The Association of Learned & Professional Society Publishers

(wileyonlinelibrary.com) doi: 10.1002/leap.1110 Received: 23 November 2016 | Accepted: 14 June 2017 | Published online in Wiley Online Library: 27 July 2017

Mapping the scientific research on open data: A bibliometric review

Yun Zhang,[1,2]* Weina Hua,[2] and Shunbo Yuan[1]

[1]School of Business, Jiaxing University, 56 South Yuexiu Road, Jiaxing 314001, Zhejiang Province, P. R. China

[2]School of Information Management, Nanjing University, 163 Xianlin Avenue, Nanjing 210023, Jiangsu Province, P. R. China

ORCID:
Y. Zhang: 0000-0002-0662-5090
W. Hua: 0000-0001-5850-4461
S. Yuan: 0000-0002-0558-4258

***Corresponding author:** Yun Zhang
E-mail: 842638105@qq.com

Abstract

This paper presents a review of open data research based on bibliometric analysis of publications in Web of Science from 1998 to 2016. It shows that research on open data has grown rapidly since 2009 with the development of various open data initiatives. We identify the different themes using science mapping and performance analysis. The most important themes are semantic web, open government, and crowdsourcing. The basic and transversal themes are data sharing and public sector information. As for the emerging themes, these are Big Data and open government data. In addition, data journalism, monitoring, and recommender systems are specific themes that deserve special attention. The UK and the USA are the leading publishing countries, both in theoretical and practical research on open data. In China, most researchers focus on practical research, and there have been efforts to promote the development of open data. Papers introducing large-scale projects receive more attention and citation quickly. Recently, researchers have been publishing more on objective topics, including possible issues and dilemmas in the era of Big Data and many problems such as budgets, ownership, licensing, culture, and sustainable development.

INTRODUCTION

Open data is the idea that some data should be freely available to everyone to use and republish as they wish, without restrictions from copyright, patents, or other mechanisms of control (Auer *et al*., 2007). The philosophy behind open data has been long-established, but the term 'open data' itself has appeared recently, gaining popularity with the development of the Internet.

There have been a large number of studies on open data, but most of them focus on the introduction of the projects in certain areas. Ahmed *et al*. (2016) proposed that drug repurposing was an innovative approach where drug candidates with the desired usage were identified by a process of re-profiling using an open-source database or knowledge of known or failed drugs already in existence. Pope, Rees, Fox, and Fleming (2014) introduced the main types and sources of remotely sensed data that were freely available and had cryospheric applications. Poldrack and Gorgolewski (2014) outlined the state of data sharing for task-based functional magnetic resonance imaging (MRI) data with a focus on various forms of data and their relative utilities for subsequent analyses in neuroimaging. They suggested that researchers and funding agencies should work together to identify and implement technical solutions that allowed the most effective data sharing without greatly increasing the burden on researchers.

Hossain, Dwivedi, and Rana (2016) used content analysis and conducted a review of 96 papers published before 2014 to address the current state of research on open data. To provide a different perspective on the growing research on open data, we conduct a retrospective bibliometrics analysis of publications on open data from 1998 to 2016 and show a recent trend on open data research. Our goals are: (1) to summarize global research trends, which may serve as a potential guide for further research; (2) to identify the

Key points

- There has been dramatic growth in articles about open data since 2009.
- Open data research focus varies between countries, with the USA taking the lead in article output.
- Articles about specific projects receive the highest citation attention.
- Emerging themes are Big Data and open government data.
- Increasingly, articles about open data discuss problems such as sustainability, ownership, and licensing.

different research emphases in some major countries; (3) to inspect the important papers with significant impacts; and (4) to identify reference bursts in the research progress.

METHODOLOGY

Data source and search strategy

This paper delineates the current state of open data research based on the documentary analysis. The structural and dynamic aspects of the scientific research on open data are revealed in two main ways, science mapping and performance analysis.

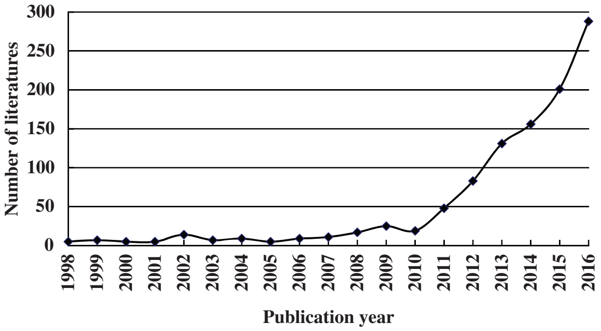

FIGURE 1 World trend of open data publications from 1998 to 2016.

Science Citation Index Expanded, Social Sciences Citation Index, and Arts & Humanities Citation Index from the Web of Science™ Core Collection, with their broad thematic coverage and renowned scientific impact, were selected as the main data sources. The query terms were set up as 'open near/1 data'. The fields were limited to title/abstract/keywords. The time spanned from 1998 to the end of 2016. The document types that we focused on were journal papers, conference proceedings, and book chapters. The data were retrieved on 19 January 2017.

This study has chosen 1998 as the starting point because the open source movement was adopted, formalized, and spearheaded by the Open Source Initiative (OSI) in 1998. Antic (2015) stated that the meanings and distinctions of 'open', 'free', and 'public' had been highly contested even during some major movements, especially the ongoing controversial arguments between Richard Stallman's Free Software Foundation and the OSI since 1998.

An overwhelming number of records resulted from this search. To assess whether the resulted set conformed to our research objectives, we performed a screening process by skimming the bibliographic information. For instance, many papers where 'open data', 'open access data', 'open government data', or 'open linked data' appeared in the topic (TS) fields may be deemed as relevant. Many papers that had 'eyes-open data', 'data on opening hours', 'data opens opportunities', or 'our data open up new venues' in the TS fields were considered irrelevant and were thus discarded. Using this screening process, our search resulted in 1,045 relevant records for the proposed analysis.

Data analysis

Retrieved records were imported to the Thomson Data Analyzer for quantitative processing. Some basic research variables were normalized, such as countries, keywords, and cited references. The normalized documents were analysed from the following aspects: (1) yearly research outputs, (2) theme evolution, (3) the research emphases in the major countries, (4) most frequently cited articles, and (5) references with top bursts.

TABLE 1 Comparison of the year-on-year growth rate of the data sets in WoS and open data (2009–2016).

Publication year	Data set in the WoS database		Data set on open data	
	Number of documents	Year-on-year growth rate (%)	Number of documents	Year-on-year growth rate (%)
2016	1,448,568	−2.13	288	43.28
2015	1,480,124	3.00	201	28.85
2014	1,437,046	2.73	156	19.08
2013	1,398,854	5.54	131	57.83
2012	1,325,455	4.82	83	72.92
2011	1,264,466	6.35	48	152.63
2010	1,188,915	4.23	19	−24
2009	1,140,633	3.98	25	47.06

WoS, Web of Science.

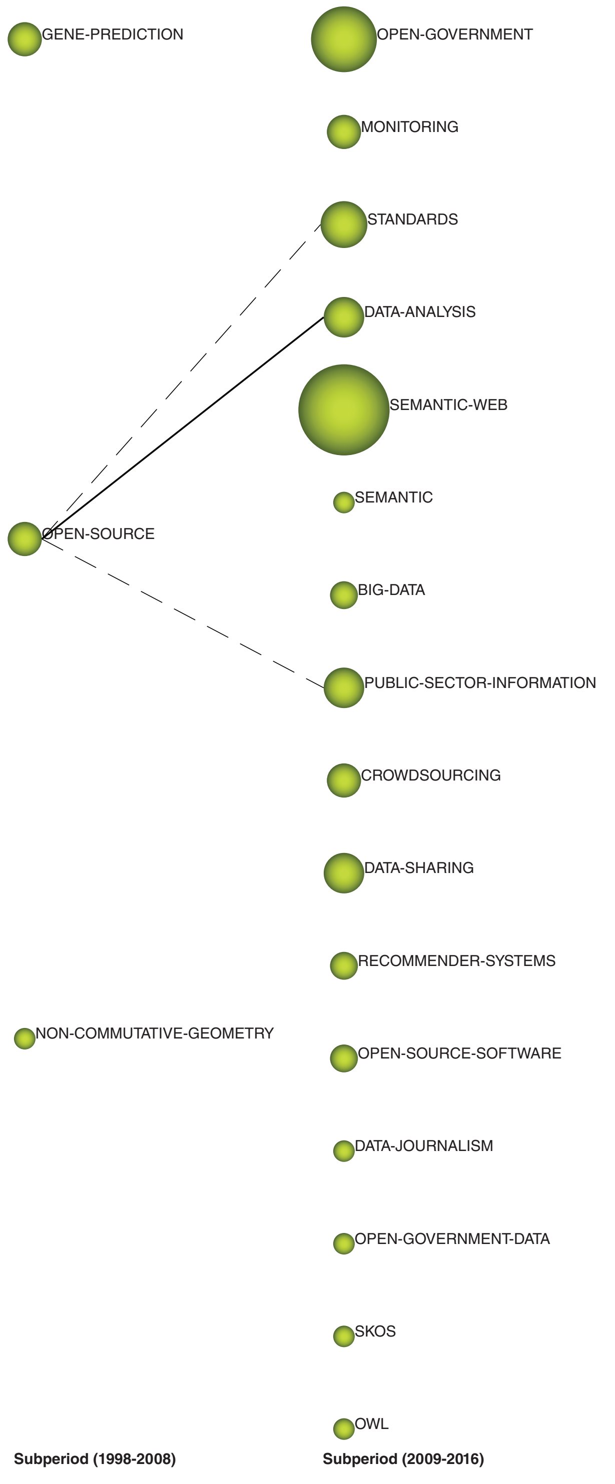

FIGURE 2 Evolution map based on *h*-index (the left column is for the period 1998–2008, and the right column is for the period 2009–2016).

Science Mapping Analysis Tool (SciMAT) was used to visualize the theme evolution. SciMAT is an open-source science mapping software tool that incorporates methods, algorithms, and measures for all the steps in the general science mapping workflow. It allows the users to carry out studies based on several bibliometric networks (co-word, co-citation, bibliographic coupling etc.). In the visualization module, three representations (strategic diagrams, cluster networks, and evolution areas) are jointly used, which enable the users to understand better the results (Cobo, López-Herrera, Herrera-Viedma, & Herrera, 2012).

UCINET, developed by Borgatti, Everett, and Freeman (2002), is a social network analysis programme. It works in tandem with a freeware programme called NetDRAW for visualizing networks. NetDRAW is a free programme written by Steve Borgatti for visualizing both 1-mode and 2-mode social network data. It can handle multiple relations at the same time, and apply node attributes to set colours, shapes, and sizes of nodes (Borgatti, 2002). In this paper, UCINET and NetDRAW were utilized to visualize the different research focuses in some major countries.

The Science of Science (Sci^2) tool is a modular toolset, specifically designed for science studies. It supports the temporal, geospatial, topical, network analysis, and visualization of data sets at the micro (individual), meso (local), and macro (global) levels (Börner, Guo, & Weingart, 2011). Sci^2 was used to identify and visualize the sudden increases or 'bursts' in the usage frequency of references over time. Burst detection can be considered as a means of identifying emerging topics, terms, or concepts.

RESULTS

Yearly research outputs

During the period 1998–2016, 1,045 relevant articles were identified from the Web of Science (WoS). Results of publication output are shown in Fig. 1. In 1998, five articles were published, while in 2016, the amount of papers rose to 288. The number of publications in the most recent years has increased significantly, from 94 during the period 1998–2008 (9%) to 951 during 2009–2016 (91%). Table 1 shows the year-on-year growth rate of the data set in the WoS database and open data and demonstrates that there is a continuous increase in the number of publications on open data.

Theme evolution

We used SciMAT to create the evolution map in order to observe the evolutions of themes of open data in two periods (1998–2008, 2009–2016) as well. A science mapping analysis wizard in SciMAT can help the user to choose the methods and algorithms to build the maps (Cobo, López-Herrera, Herrera-Viedma, & Herrera, 2011b). Several parameters were set, in which the keyword group was selected as the unit of analysis, while the minimum keyword frequency thresholds were set to 2 and 3, and the minimum edge value thresholds between keywords were set to 1 for the co-occurrence network in two different time periods. We also utilized Jaccard's index as the similarity measure. SciMAT can also enrich the science maps with bibliometric measures such as the sum, maximum, average citations,

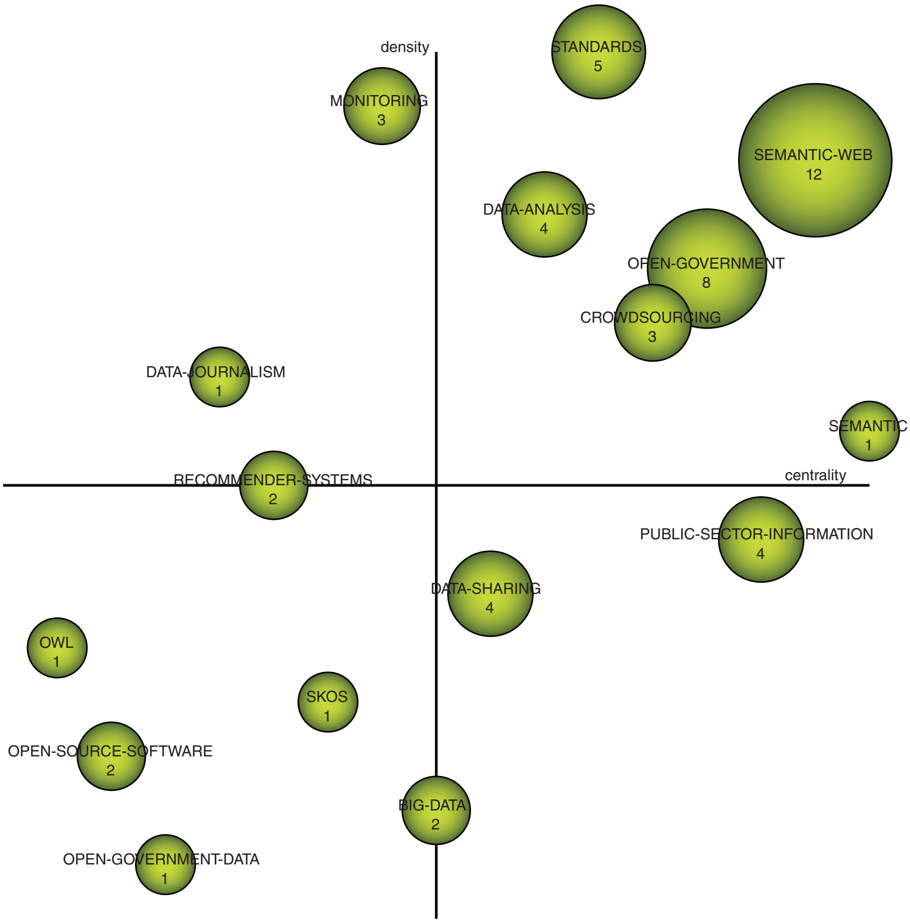

FIGURE 3 Strategic diagram in the sub-period (2009–2016) based on the *h*-index.

TABLE 2 Performance measurements for the themes in the sub-period (2009–2016).

Theme name	Number of documents	Number of citations	Average of citations	*h*-Index
Open-government	49	316	6.45	8
Monitoring	7	290	41.43	3
Standards	11	68	6.18	5
Data-analysis	7	65	9.29	4
Semantic-web	85	469	5.52	12
Semantic	6	12	2.00	1
Big-Data	13	18	1.38	2
Public-sector-information	10	43	4.30	4
Crowdsourcing	10	36	3.60	3
Data-sharing	12	46	3.83	4
Recommender-systems	7	48	6.86	2
Open-source-software	4	5	1.25	2
Data-journalism	2	3	1.50	1
Open-government-data	3	40	13.33	1
SKOS	2	3	1.50	1
OWL	2	2	1.00	1

OWL, Web Ontology Language; SKOS, Simple Knowledge Organization System.

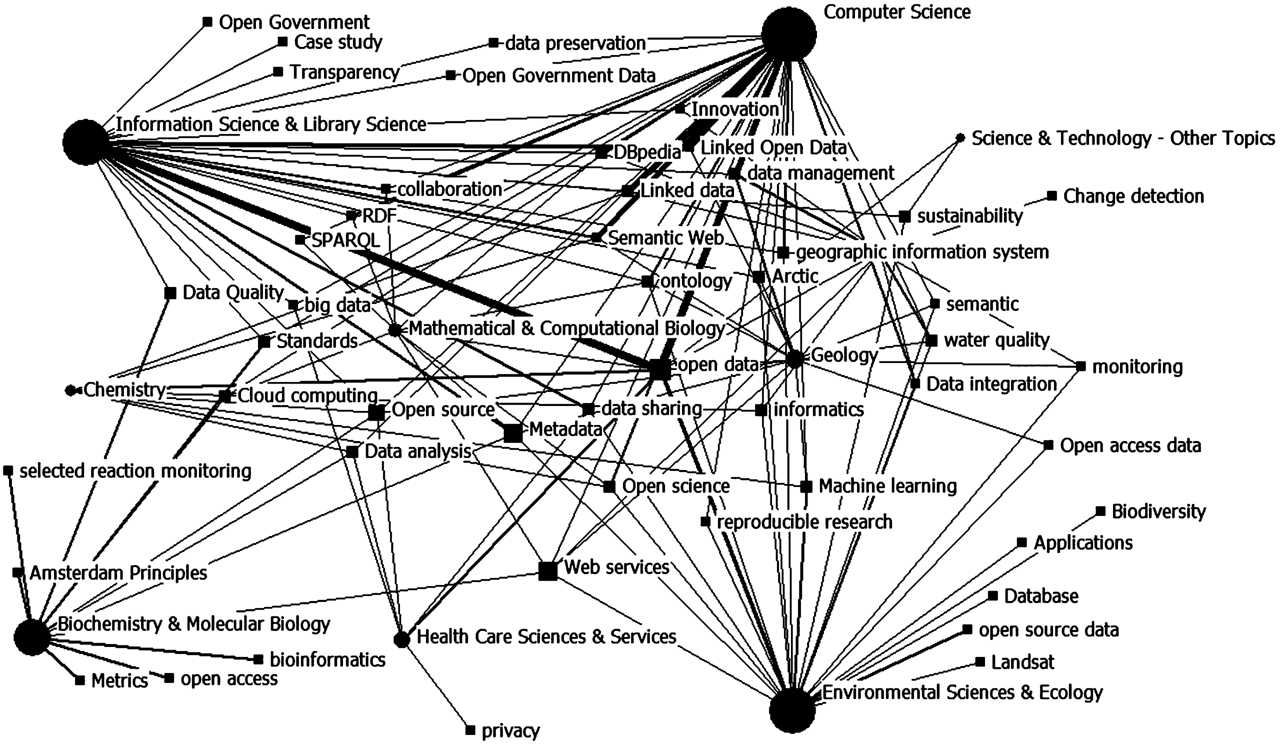

FIGURE 4 Research focus in USA.

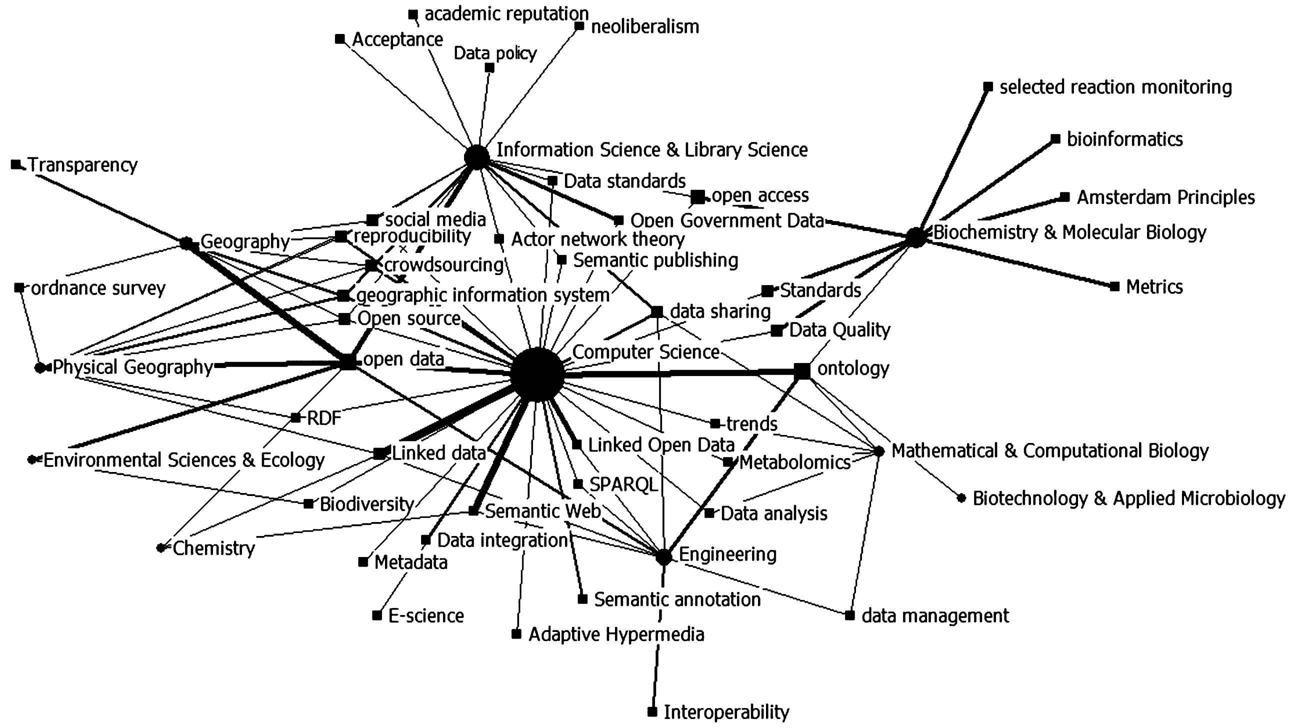

FIGURE 5 Research focus in UK.

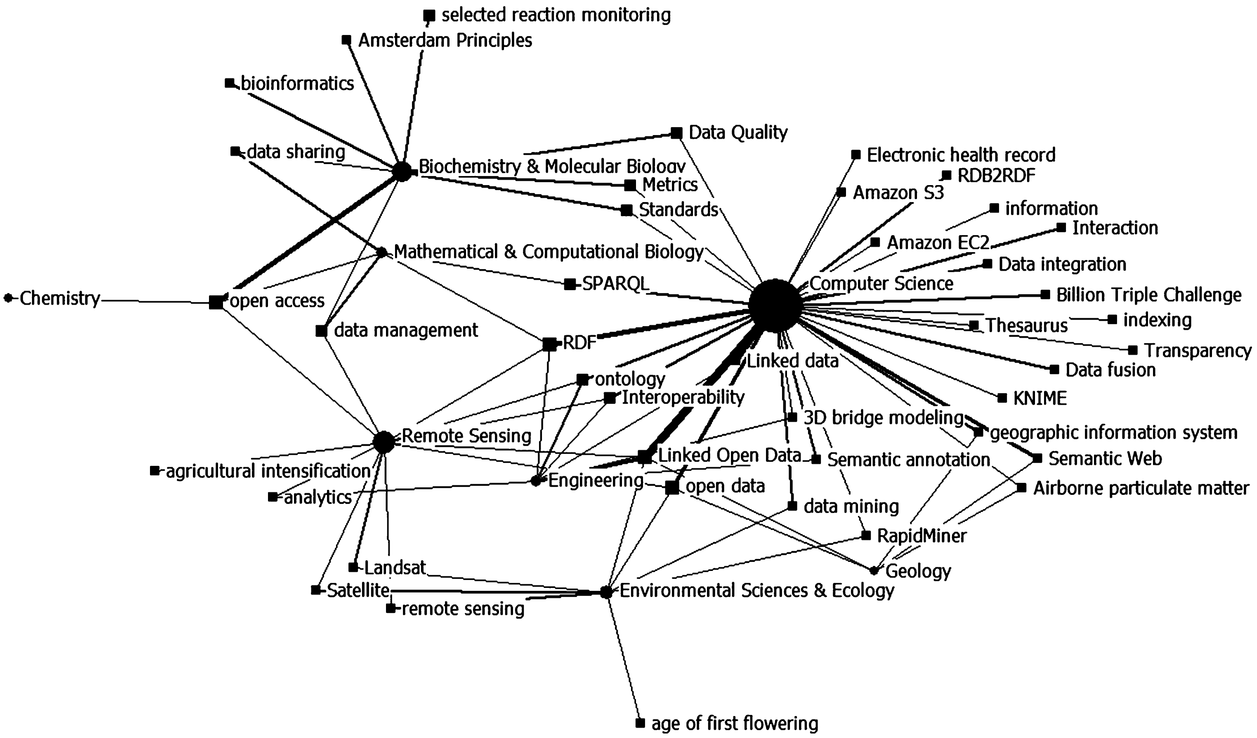

FIGURE 6 Research focus in Germany.

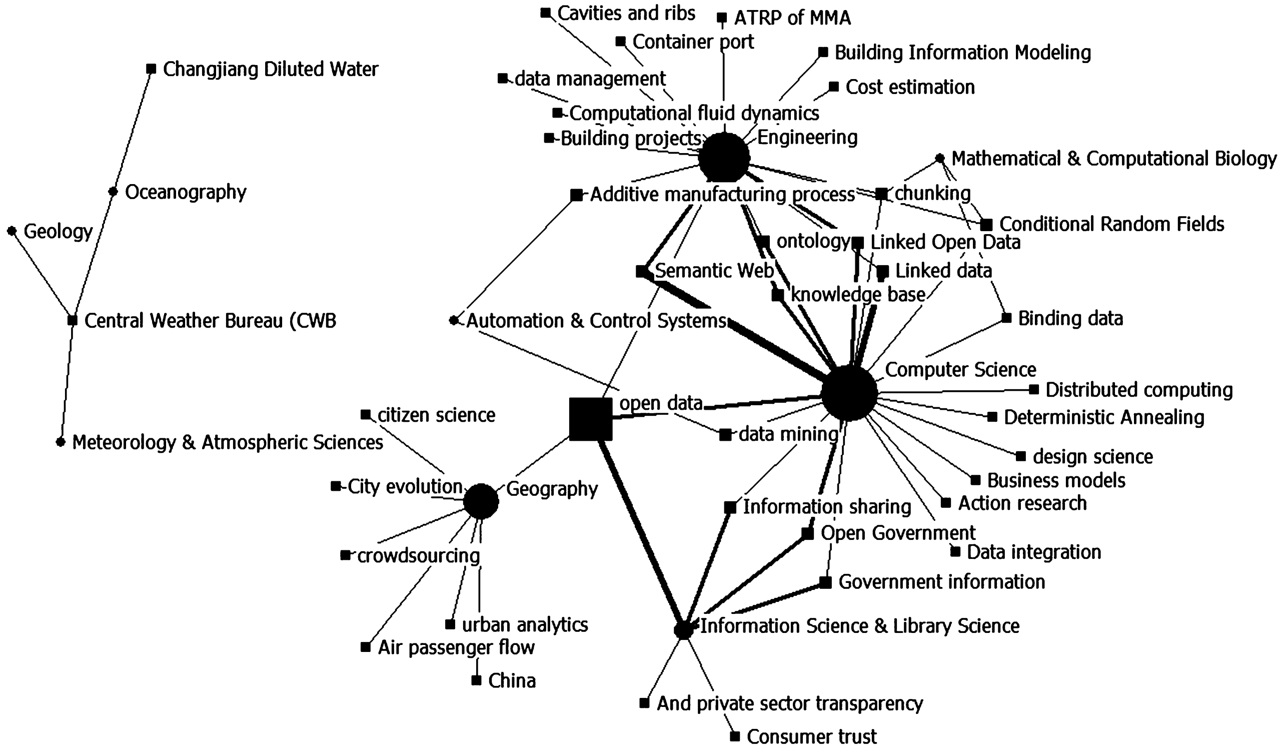

FIGURE 7 Research focus in China.

and *h*-index to provide the performance information of the nodes in the network (Cobo *et al.*, 2011b). In Fig. 2, *h*-index was chosen to show the performances of different themes. There are three themes, open source, gene prediction, and non-commutative geometry, shown in the first sub-period (1998–2008). The values of *h*-index for the three themes are 3, 3, and 1, respectively, whereas

TABLE 3 Most frequently cited papers in the WoS database.

Authors	Title	Source	Times cited
O'Boyle, NM; Banck, M; James, CA; Morley, C; Vandermeersch, T; Hutchison, GR	Open Babel: An open chemical toolbox	*Journal of Cheminformatics*. 2011; 3	785
Gaulton, A; Bellis, LJ; Bento, AP; Chambers, J; Davies, M; Hersey, A; *et al.*	ChEMBL: Large-scale bioactivity database for drug discovery	*Nucleic Acids Research*. 2012; 40(D1): D1100–D1107	772
Kleyer, M; Bekker, RM; Knevel, IC; Bakker, JP; Thompson, K; *et al.*	The LEDA Traitbase: A database of life-history traits of the Northwest European flora	*Journal of Ecology*. 2008; 96(6): 1266–1274	452
Pluskal, T; Castillo, S; Villar-Briones, A; Oresic, M	MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data	*BMC Bioinformatics*. 2010; 11: 395	397
Kerrien, S; Aranda, B; Breuza, L; Bridge, A; Broackes-Carter, F; Chen, C; *et al.*	The IntAct molecular interaction database in 2012	*Nucleic Acids Research*. 2012; 40(D1): D841–D846	387
Aranda, B; Achuthan, P; Alam-Faruque, Y; Armean, I; Bridge, A; Derow, C; *et al.*	The IntAct molecular interaction database in 2010	*Nucleic Acids Research*. 2010; 38: D525–D531	381
Croft, D; Mundo, AF; Haw, R; Milacic, M; Weiser, J; Wu, GM; *et al.*	The reactome pathway knowledgebase	*Nucleic Acids Research*. 2014; 42(D1): D472–D477	344
Avants, BB; Tustison, NJ; Song, G; Cook, PA; Klein, A; Gee, JC	A reproducible evaluation of ANTs similarity metric performance in brain image registration	*Neuroimage*. 2011; 54(3): 2033–2044	342
Guha, R; Howard, MT; Hutchison, GR; Murray-Rust, P; Rzepa, H; Steinbeck, C; *et al.*	Blue Obelisk—Interoperability in chemical informatics	*Journal of Chemical Information and Modeling*. 2006; 46(3): 991–998	290
Groom, CR; Bruno, IJ; Lightfoot, MP; Ward, SC	The Cambridge structural database	*ACTA Crystallographica Section B: Structural Science, Crystal Engineering and Materials*. 2016; 72: 171–179	253

16 themes are identified in the second sub-period (2009–2016), including semantic web (*h*-index = 12), open government (*h*-index = 8), standards (*h*-index = 5), and the rest (see Fig. 2 for full list of themes and their corresponding *h*-index values).

A strategic diagram (Fig. 3) was drawn to investigate the different impacts of themes on open data since 2009. By default, SciMAT adds Callon's density and centrality as network measures to each detected cluster in the selected period. Callon's centrality measures the degree of interaction of a network with other networks and is considered the external cohesion of the network. Callon's density measures the internal strength of the network and is considered the internal cohesion of the network. These measures are useful to categorize the detected clusters of a given period in a strategic diagram (Cobo, López-Herrera, Herrera-Viedma, & Herrera, 2011a; Cobo *et al.*, 2012). Themes in the upper-right quadrant are both well developed and important for structuring a research field. Themes in the upper-left quadrant are regarded as marginal in the field. Themes in the lower-left quadrant are either emerging or disappearing themes. Themes in the lower-right quadrant are transversal and general, basic themes (Cobo *et al.*, 2011a). In Fig. 3, the size of the sphere was proportional to the *h*-index of each theme. We also present the quantitative measures of each theme in Table 2. Our results (Fig. 3 and Table 2) show that the most important themes are semantic web, open government, standards, data analysis, and crowdsourcing. When considering the basic and transversal themes only, data sharing and public sector information stand out. As for the emerging themes, there are Big Data and open government data. In addition, monitoring, data journalism, and recommender systems are specific themes.

Different research emphases in major countries

The country-based analysis on open data researches can help us understand the different focuses in various countries. Among the 1,045 selected articles published from 1998 to 2016, the authors' affiliated addresses are from 74 countries/territories. The top 10 countries are USA, UK, Germany, Spain, Italy, Netherlands, China, Canada, France, and Australia in the order of number of publications.

USA, UK, Germany, and China were chosen as four major countries for more detailed analysis. The top 10 research areas and the top 50 keywords in the four countries were selected to create the category–keyword matrices from the different subsets of data. The matrices were then imported into UCINET so that the data could be used for NetDRAW. Figures 4–7 were generated by NetDRAW.

The whole network in the USA (Fig. 4) is closely connected and the subjects cover a wide range of topics. The four most important research areas are computer science, environment

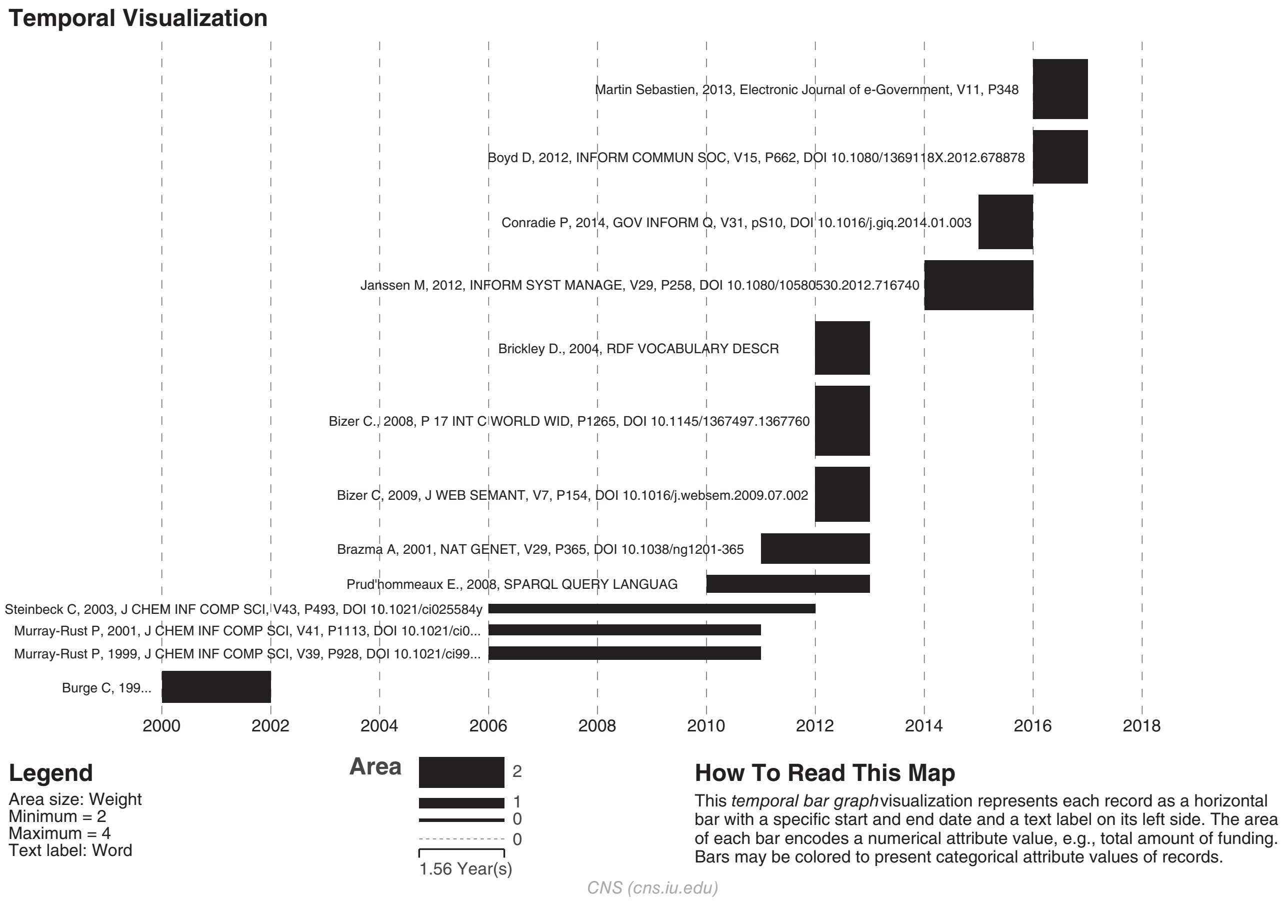

FIGURE 8 Top reference bursts in the 'open data' data set (Gamma = 0.75).

science and ecology, information science and library science, and biochemistry and molecular biology. There are very close connections among the different research areas.

In the UK (Fig. 5), it can be seen that computer science, information science and library science, and geology are closely connected and share keywords like social media, crowdsourcing, and reproducibility. In addition, computer science and engineering share keywords related to semantic web technologies such as linked data, semantic web, and SPARQL [SPARQL Protocol and Resource Description Framework (RDF) Query Language].

In Germany (Fig. 6), it can be seen that computer science is the most important research area, with emphasis on information indexing, data quality, data mining, semantic web, and related technologies. It also demonstrates that semantic web and related technologies are widely used in environment science and ecology, engineering, and geology.

In China (Fig. 7), the subject scope is relatively narrow because the number of published research papers is relatively low. Computer science is the most important research area and the semantic web technologies have been primarily used in engineering. The special research area is geography, which is closely related to the research on citizen science, crowdsourcing, and urban analytics.

Most frequently cited articles

The citation count was obtained from the WoS database. Table 3 lists 10 most frequently cited papers during the period from 1998 to 2016. These 10 papers were published in seven different journals. Among them, five papers are of biochemistry and molecular biology and three papers are of chemistry. O'Boyle *et al.* (2011) is the most frequent cited paper and has been cited 785 times since its publication.

Some projects were introduced in these papers. For example, ChEMBL is an open data database containing binding, functional, and ADMET (absorption, distribution, metabolism, excretion, and toxicity) information for a large number of drug-like bioactive compounds and is accessible via https://www.ebi.ac.uk/chembldb (Gaulton *et al.*, 2012). The Cambridge Structural Database (CSD) includes a comprehensive and fully retrospective historical archive of small-molecule crystallography, which is a very important source for research in structural chemistry, materials science, and the life sciences (Groom, Bruno, Lightfoot, & Ward, 2016).

Top reference bursts

Attribute values of scholarly entities and their diverse aggregations increase and decrease at different rates and respond with

TABLE 4 Keyword distributions in different themes.

Theme name	Keywords (the number of documents)
Open-government	Open data (171); open government (34); transparency (21); e-government (9); business models (5); public administration (4); accountability (3); adoption (3); governance (3); information reuse (3); information sharing (3); local government (3)
Monitoring	Data management (15); remote sensing (6); citizen science (5); monitoring (5); biodiversity (4); landsat (4); mapping (4); policy (4); web services (4); observation (3); satellite (3); water data (3)
Standards	Metdata (18); open access (15); data quality (8); metrics (7); standards (7); research data (4); vocabularies (4); Amsterdam principles (3); bioinformatics (3); data preservation (3); epidemiology (3); selected reaction monitoring (3)
Data-analysis	Open source (15); data analysis (8); data format (4); metabolomics (4); social networks (4); world wide web (4); industry foundation classes (3); infrastructure (3); mass-spectrometry (3); open-architecture (3); volunteered geographic information (3); XML (3)
Semantic-web	Linked open data (98); linked data (63); semantic web (61); ontology (39); RDF (34); knowledge management (6); human computer interaction (4); knowledge representation (4); libraries (4); control vocabulary (3); instance matching (3); usability (3)
Semantic	Data integration (12); SPARQL (10); data mining (8); thesaurus (8); semantic (7); agriculture (5); repositories (4); wikipedia (4); dynamic (3); e-science (3); open data model (3); sensors (3)
Big-Data	Big data (24); smart city (10); visualization (9); APPS (5); analytics (4); data fusion (4); KNIME (4); modelling (4); algorithms (3); map (3); meta analysis (3); social science (3)
Public-sector-information	Database (15); public sector information (10); data (9); innovation (7); evaluation (6); ICT (6); open access data (5); semantic interoperability (5); decision support system (4); information (4); spatial data infrastructure (3); standardization (3)
Crowdsourcing	Geographic information system (11); crowdsourcing (10); social media (9); cloud computing (5); design (5); ranking (4); urban analytics (4); web 2.0 (4); China (3); geospatial data (3)
Data-sharing	Data sharing (24); interoperability (15); open science (11); privacy (8); data standards (4); data citation (3); data publication (3); replication (3); reproducible research (3); research methods (3)
Recommender-systems	DBpedia (10); recommender systems (7); social tagging (4); personalization (3); semantic similarity (3)
Open-source-software	Open data kit (6); open source software (6); sensor network (4); decision tree (3); participation (3)
Data-journalism	Data viualization (6); statistics (4); civil society (3); data journalism (3)
Open-government-data	Open government data (21); datasets (5); taxonomy (4); freedom of information (3)
SKOS	SKOS (5); classification (3); environment (3)
OWL	Knowledge base (4); OWL (4); LEMON (3)

APPS, Applications; ICT, Information and Communication Technology; KNIME, Konstanz Information Miner; LEMON, Lexicon Model for Ontologies; OWL, Web Ontology Language; RDF, Resource Description Framework; SPARQL, SPARQL Protocol and Resource Description Framework (RDF) Query Language; SKOS, Simple Knowledge Organization System; XML, Extensible Markup Language.

different latency rates to internal and external events. In Sci^2, Kleinberg's burst detection algorithm (Kleinberg, 2003) can be used to detect bursts related to author names, references, Institute for Scientific Information keywords, or terms used in the title and/or abstract of a paper. The algorithm generates a list of the bursts in the document stream, ranked by the burst weight, along with the intervals of time in which these bursts occur. The algorithm has four important parameters: the gamma value, first ratio, general ratio, and the number of bursting states. The gamma parameter controls the ease with which the automaton can change states. The higher the gamma value is, the smaller the list of bursts generated (Börner *et al*., 2011). In Fig. 8, 13 references were identified when gamma was set to 0.75 and the total area of the bar corresponded to the value of the burst weight.

The 13 references (in Fig. 8) reflect the different focuses of the researchers. For example, Murray-Rust and Rzepa (1999, 2001) introduced chemical markup language (CML), and showed how a document object model (DOM) for chemistry can be constructed based on the development of CML, which can be used as a primary means of representing chemical information in computers. Bizer, Heath, Idehen, and Berners-Lee (2008) pointed out that the web was increasingly understood as a global information space consisting not just of linked documents but also of linked data. Janssen, Charalabidis, and Zuiderwijk (2012) proposed that in order to ensure that the open system is further adopted, researchers should pay more attention to the following research: how to deal with obstacles, how to gain insight from the perspective of users, and how to establish the incentive system of stimulating and using open data.

DISCUSSION

Our result (Fig. 1) shows that the research on open data has grown rapidly since 2009. Several governments (such as in the USA, UK, Canada, and New Zealand) have continued to announce initiatives with regard to opening up their public information. For examples, USA's President Obama issued the Open Government Directive on 8 December 2009. The USA established its *data.gov* portal in 2009. Canada launched the open data portal in 2011. The promulgation of the policy, the development of demonstration projects, and the future prospects interest many researchers. Therefore, the theoretical and practical explorations of many areas of study have been gradually explored and expanded.

The booming trend of open data research can also be seen in Fig. 2. In the first sub-period (1998–2008), there are only three themes out of 94 published original papers. The three themes involve 15 papers, accounting for 15.96% of the total because of the reduced network and the convergence effects of important themes. In the second sub-period (2009–2016), 16 themes are identified, which involve 550 papers, accounting for 57.83% of all the 951 articles. It clearly shows that the research on open data has developed rapidly since 2009, manifested in the increasing number of themes and the enhancement of convergence effects among articles.

The evolution and impact of themes on open data are shown in Figs 2 and 3, respectively. Table 4 is dedicated to supplementing the results by showing the keyword distributions in different themes. Based on the combined analysis, we can elicit more information on the themes. In the upper-right quadrant (Fig. 3), semantic web is a theme with the greatest *h*-index value, which focus on the following aspects: the new data organization technologies, such as linked data, linked open data, ontology, and RDF; the research on knowledge management and knowledge representation; and the human–computer interaction to promote the efficiency of machine-readable resources for humans. In addition, open government is also an important theme that emphasizes the following aspects: working on the provision of open data, improving the quality of service, and reforming the mode to speed up the process of e-government and public governance. We also find that the theme of crowdsourcing has become an the important force to promote the development of open data, with the rapid development of information technology like cloud computing, the enhancing support capabilities of social medias for public participation, and the rapid development in application areas in geospatial metadata and geographic information systems.

Data sharing in the lower-right quadrant (Fig. 3), however, is not a new theme; now, it provides prominence to the repeatability and reproducibility of research with the development of open science. It also values the efforts in privacy protection and data interoperability and deals with the new challenges in data sharing practices, such as data citation and data publication.

There are some emerging themes such as Big Data and open government data in the lower-left quadrant (Fig. 3). Big Data is a very hot topic in recent years, containing many related keywords such as algorithms, modelling, visualization, and APPS. Our results show that smart city is one of the important keywords related to the keyword Big Data. We think the research of Big Data is closely dependent on the development of computer algorithms and tools, and the research results of Big Data are of great significance to the construction of smart city. From the perspective of popularity of the research, themes of SKOS (Simple Knowledge Organization System) and OWL (Web Ontology Language), which refer to the knowledge management in semantic web, are now declining. However, there have been quite a lot of research findings from them.

The digital journalism in the upper-left quadrant (Fig. 3) reflects the essential role of numerical data when reaching and distributing information in the digital era, such as helping a journalist tell a complicated story based on infographics (Gray, Chambers, & Bounegru, 2012). The research on data journalism will embrace some new issues, like its actors, practices, conditions of data access, the computer and statistical skills required, and training programmes.

Comparing the figures of the four different countries (Figs 4–7), there are some similarities and dissimilarities. In general, computer science is the support discipline for open data research in many countries by providing algorithms, standards, tools, and some other advanced technologies. Semantic web and its related technologies have been widely used, in which the linked open data sets provide abundant and open repositories to be incorporated in many application programmes. Furthermore, RDF, SPARQL, and OWL are the technologies used to support the development processes of the researchers.

USA ranks first with 216 papers covering a wide range of themes. It is worth noting that the research in computer science and information science and library science form the basis for other fields, and share many common topics. We also find that the papers on environment science and ecology and biochemistry and molecular biology often introduce some practical projects, such as large-scale databases, open data format, and open source platforms.

The other three countries have different emphases. UK research emphasizes the use of open data but there is relatively little fundamental research compared to the USA. German research emphasizes technology issues, and information science and library science is not among the top 10 research areas. China mainly concentrates on applied research. The connections between different disciplines are weak and some increases have been made with the development of citizen science.

Additionally, the Open Data Barometer report in 2016 (World Wide Web Foundation, 2016) shows that the UK and the USA, as the long-standing leaders of open data research, remain top in the global rankings, while Germany ranks 11th and China ranks 55th. It can be surmised that the UK and the USA have strong capabilities in theoretical and applied research. China still faces many challenges to catch up in the development of open data.

The top 10 high-impact articles in Table 3 were published from 2006 to 2016, among which eight articles were published after 2009 and received significant attention. It is worth noting that the top 10 articles are all related to the introduction of large-scale projects, mainly in the areas of biochemistry and molecular biology, chemistry, plant sciences and radiology, and nuclear medicine and medical imaging.

These projects tend to have some common features. First, the projects contain a large number of multi-source and heterogeneous data; they increase the possibilities of new discoveries and knowledge creation. Second, the projects are maintained by specific entities, in which the standard exchange formats and standard recognition mechanisms are designed to ensure that the data are widely discoverable and readily reusable. Third, the projects often provide third-party application software, application-programming interface, or other toolkits to facilitate the ease of access or the integration with other applications. Finally, the projects are often posted on the Internet to provide a dedicated website and many free network services. We think that these papers provide important directions for the researchers; for instance, some researchers have come up with more in-depth research findings, whereas others have proposed improvements or new research designs based on the shortcomings of existing designs. As a trend, there could be more papers on the introduction of the large-scale open data projects and further research based on them.

The 13 references (in Fig. 8) identified by burst detection reflect the changes in research focus, from gene prediction and CML in the early stages, to linked data and the semantic web, and finally to implementation difficulties and problems of sustainable development on open data and Big Data in recent years. We classify the references into three broad stages. In the early stages, the studies focus on the integration and development of heterogeneous data in specific fields, especially by using computer technologies. Then, there is the flourishing development of semantic web research. In recent years, more research has been on practical problems associated with open data, open government, and Big Data.

In general, scholars have tried to increase the interconnections and interoperations of data, and have accumulated good results in some specific areas. More recently, researchers have taken more objective attitudes with the expansion in the scale of the projects, the coming of the era of Big Data, and many actual problems, such as budgets, ownership, licensing, culture, and sustainable development. Some researchers claim that data release by governments is still novel and there is little experience and knowledge thus far about its benefits, costs, and barriers. They propose that the way data is stored, the way data is obtained, and the way data is used by a department are crucial indicators for open data release (Conradie & Choenni, 2014). Some researchers point out possible dilemmas that may cause conflicts – for example, Big Data redefines the meaning of knowledge – but bigger data does not always mean better data. Big Data may lose some information when taken out of context, and limited access to Big Data may create new digital divides (Boyd & Crawford, 2012).

CONCLUSION

Motivated by the needs to explore the recent development in open data research, we conducted a review on published research and enriched the research results with visualization and bibliometrics. Our results show that open data is an extremely hot topic in recent years, where several achievements have been made and some new and special topics are emerging. Although the application prospect of open data is very promising, the academic environment is far more diverse and complex, particularly around questions of data ownership and lack of research funding in some disciplines. Our study sheds lights on further research directions for researchers, and it also provides a preliminary assessment for government funding plans, such as what areas are making great progress and what areas are emerging and needed to be supported. For future research, more in-depth studies and results from specific themes could be investigated. Open government, data journalism, data citation, and data publication are key topics. Other growing areas may include government policies and funding reports.

ACKNOWLEDGEMENTS

Research in this article is funded by National Social Science Foundation of China – Youth Project (Grants 14CTQ046). The authors are grateful for the valuable suggestions of anonymous reviewers and the Editors, which significantly improved the article.

REFERENCES

Ahmed, M. U., Bennett, D. J., Hsieh, T. C., Doonan, B. B., Ahmed, S., & Wu, J. M. (2016). Repositioning of drugs using open-access data portal DTome: A test case with probenecid [Review]. *International Journal of Molecular Medicine*, *37*(1), 3–10. https://doi.org/10.3892/ijmm.2015.2411

Antic, A. (2015, September). *John Dewey's philosophical legacy for the "Open Movements" in the digital age*. Paper presented at the epc2: European Pragmatism Conference II, Paris, France. Retrieved from https://epc2.sciencesconf.org/59307/document

Auer, S. R., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). DBpedia: A nucleus for a web of open data. In K. Aberer, K.-S. Choi, N. Noy, D. Allemang, K.-I. Lee, L. Nixon et al. (Eds.), *The semantic web* (pp. 722–735). Berlin and Heidelberg, Germany: Springer.

Bizer, C., Heath, T., Idehen, K., & Berners-Lee, T. (2008). *Linked data on the web (LDOW2008)*. Paper presented at the WWW '08 Proceedings of the 17th International Conference on World Wide Web, April 2008, Beijing, China. Retrieved from http://www2008.org/papers/pdf/p1265-bizer.pdf

Borgatti, S. P. (2002). *NetDraw: Graph visualization software*. Harvard, MA: Analytic Technologies. Retrieved from http://www.analytictech.com/Netdraw/netdraw.htm

Borgatti, S. P., Everett, M. G., & Freeman, L. C. (2002). *Ucinet for windows: Software for social network analysis*. Harvard, MA: Analytic Technologies. Retrieved from http://www.analytictech.com/ucinet/

Börner, K., Guo, H., & Weingart, S. (2011). *Science of Science (Sci^2) tool manual*. Retrieved from Cyberinfrastructure for Network Science Center, School of Library and Information Science, Indiana University website: http://sci2.wiki.cns.iu.edu/

Boyd, D., & Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society*, *15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2011a). An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the Fuzzy Sets Theory field. *Journal of Informetrics*, *5*, 146–166. https://doi.org/10.1016/j.joi.2010.10.002

Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2011b). *SciMAT version 1.0 user guide*. Retrieved from http://sci2s.ugr.es/scimat/software/v1.01/SciMAT-v1.0-userGuide.pdf

Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2012). SciMAT: A new science mapping analysis software tool. *Journal of the American Society for Information Science and Technology*, *63*(8), 1609–1630. https://doi.org/10.1002/asi.22688

Conradie, P., & Choenni, S. (2014). On the barriers for local government releasing open data. *Government Information Quarterly*, *31*, S10–S17. https://doi.org/10.1016/j.giq.2014.01.003

Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., … Overington, J. P. (2012). ChEMBL: Large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, *40*(D1), D1100–D1107. https://doi.org/10.1093/nar/gkr777

Gray, J., Chambers, L., & Bounegru, L. (2012). *Data journalism handbook 1.0 beta*. Retrieved from http://datajournalismhandbook.org/1.0/en/introduction_0.html

Groom, C. R., Bruno, I. J., Lightfoot, M. P., & Ward, S. C. (2016). The Cambridge Structural Database. *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials.*, *72*, 171–179. https://doi.org/10.1107/S2052520616003954

Hossain, M. A., Dwivedi, Y. K., & Rana, N. P. (2016). State-of-the-art in open data research: Insights from existing literature and a research agenda. *Journal of Organizational Computing and Electronic Commerce*, *26*(1–2), 14–40. https://doi.org/10.1080/10919392.2015.1124007

Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. *Information Systems Management*, *29*(4), 258–268. https://doi.org/10.1080/10580530.2012.716740

Kleinberg, J. (2003). Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, *7*, 373–397. https://doi.org/10.1023/A:1024940629314

Murray-Rust, P., & Rzepa, H. S. (1999). Chemical markup, XML, and the worldwide web. 1: Basic principles. *Journal of Chemical Information and Computer Sciences*, *39*(6), 928–942. https://doi.org/10.1021/ci990052b

Murray-Rust, P., & Rzepa, H. S. (2001). Chemical markup, XML and the world-wide web. 2: Information objects and the CMLDOM. *Journal of Chemical Information and Computer Sciences*, *41*(5), 1113–1123. https://doi.org/10.1021/ci000404a

O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., & Hutchison, G. R. (2011). Open Babel: An open chemical toolbox. *Journal of Cheminformatics*, *3*, 33. https://doi.org/10.1186/1758-2946-3-33. Retrieved from http://jcheminf.springeropen.com/articles/10.1186/ 1758-2946-3-33

World Wide Web Foundation. (2016, April). *ODB global report third edition*. Washington, DC: World Wide Web Foundation. Retrieved from http://opendatabarometer.org/doc/3rdEdition/ODB-3rdEdition-GlobalReport.pdf

Poldrack, R. A., & Gorgolewski, K. J. (2014). Making big data open: Data sharing in neuroimaging. *Nature Neuroscience*, *17*(11), 1510–1517. https://doi.org/10.1038/nn.3818

Pope, A., Rees, W. G., Fox, A. J., & Fleming, A. (2014). Open access data in polar and cryospheric remote sensing. *Remote Sensing*, *6*(7), 6183–6220. https://doi.org/10.3390/rs6076183