Researching Digital Society: Using Data-Mining to Identify Relevant Themes from an Open Access Journal

Judith Schossböck, Noella Edelmann, Nina Rizun

Researching Digital Society: Using Data-Mining to Identify Relevant Themes from an Open Access Journal

Judith Schoßböck[0000-0002-5473-7361] [1] Noella Edelmann[0000-0001-8386-9585] [1*]

Nina Rizun[0000-0002-4343-9713] [2]

[1] Danube University Krems, Austria
[Firstname.surname]@donau-uni.ac.at
[2] Gdansk University of Technology, Poland
nina.rizun@pg.edu.pl
*corresponding author

Abstract. Open Access scholarly literature is scientific output free from economic barriers and copyright restrictions. Using a case study approach, data mining methods and qualitative analysis, the scholarly output and the meta-data of the Open Access eJournal of e-Democracy and Open Government during the time interval 2009-2020 was analysed. Our study was able to identify the most prominent research topics (defined as thematic clusters) of the journal, their evolution over time and how these were influenced by journal management factors. This kind of analysis helps editors to develop an editorial strategy, decide on the thematic development of the journal and address the expectations of future authors of the journal. It further can provide insights about research themes and trends within a scholarly community and their development over time.

Keywords: scholarly-led publishing, open access, data analysis, journals, data mining, research topics

1 Introduction

Open access (OA) publications play an important role for different stakeholders, such as academia, policy-makers, and practitioners [1]. Apart from the objective of OA to free scientific output from economic barriers and copyright restrictions, ensuring accessibility to high-quality research output in the form of open publication, the access to openly available data and metadata allows an analysis of research topical trends, methods, or "scientific turns" [2], by editors, industry and practitioners [3]. Trends are often noticeable across disciplines, and demonstrate the interaction of research and sociological, historical or political developments [4]. Publication foci are created in accordance with societal issues, specific research trends, methodological turns, or dominant theoretical frameworks, but also policy, "politics" of academia and publishing cultures [5]. For instance, "Plan S" [6], a dedicated radical open access initiative, is often referenced to both as a turning point towards more open science and as an important policy that explains shifts in the nature of research publications [7]. Similarly, topical trends or

thematic turns in research can indicate, reflect, but also influence socio-political change or action [8].

There are several ways to analyse the development of a field, research agendas, and their relation to societal and technological developments, such as literature reviews, meta-analysis, or data-based approaches. For this contribution, our work demonstrates how the computational and analytics-driven methods can be used in scholarly-led research to adjust its thematic scope, and to guide journal development. Hence, the objective of this study concerns the strategic choices of journal editors and managers, as well as the reserach community. First, it aims to provide an approach that can be used by journal editors to identify topical trends in scholarly-led publishing and to derive important clues regarding thematic clusters within a research community (RQ1), and second, we investigate what factors might impact the trends, research topics and research cooperation evolution over time (RQ2). The rest of the paper is organized as follows: Section 2 provides a review of topical trends within research communities and scholarly-led publishing from an editorial perspective. Section 3 describes the research methods in detail. Our results are provided in Section 4. Section 5 contains a discussion on consistent themes and the utility of our method. Finally, the limitations and recommendations for future research are addressed in Section 6.

2 Literature Review

The review covers exemplary research themes investigated in a scholarly community and reflects on the journal management and scholarly-led perspective relevant to the case study.

2.1 Topical trends and agendas within research communities

The “open society” has been framed as a watchword of liberal democracy and the market system in the modern globalised world, and “openness” stands for both bottom-up empowerment and top-down transparency [9]. The emergence of research areas such as “open data” [10], “open science”, “open access” [3], “open government” [11], and “open access” [12] reflects this trend across many disciplines and locations. On the policy level, the *Open Government Directive* in the US is often cited as important influence for the trend towards more “openness” in e-government research [13]. Another example is the development of the field of e-participation research, fostered by technology innovations and research programs, and was later characterised and recognised as “emerging research area” [14, p.415]. Understanding the field was, particularly in the beginning, complex, as there was no generally agreed upon definition and no clear overview of disciplines or methods to draw upon. As Sæbø et al. [14] note, their initital analysis provided a starting point for the development of a modelused to define e-participation activities such as e-voting, online discourse, online decision-making, e-activism, e-consultation, e-campaigning, and e-petitioning. Such work helps to better understand emerging fields, the role of related disciplines, and provides the basis for developing research agendas in the future [14]. Another study [8] analysed a total of

150 publications during a timeframe of 10 years related to this research field and a conference series, showed how the agenda has changed over a decade of research. This data-based approach, in contrast to literature reviews, offers insights on the development of a field over time.

Thus, it is critically important to study the intellectual core and the dynamics of knowledge development over time [15]. Many scholars have paid attention to evolution within different research fields, using several research methods: exploratory content analysis, to investigate the progress and development of e-government research [15], Natural Language Processing (NLP) techniques, in order to extract key terms or their co-occurrence, to discover the underlying knowledge structure and evolution of the Journal of Knowledge Management [16], topic modeling (LDA) or Structural Topic Modelling (STM) to profile the research hotspots, exploring annual topic proportion trends and topic correlations, potential future research directions in information management [17] and transportation engineering [18] research.

2.2 Scholarly-led publishing: managing a diamond OA journal

Journal editors face various issues and choices in order to support the development of the journals they lead and manage. One issue are the strategic choices that are to enhance the visibility, accessibility and impact of research published in their journal. Another issue is how to provide potential authors of the journal with useful information that will support them with their publication choice [19]. This is important as scholarly-led publishing is increasingly connected to academic metric systems such as journal indexes, e.g. Scopus or Web of Science, as well as authors' citation scores and academic evaluation such as Publish-or-Perish, h-index, or Google Scholar.

In order to define their scope and strategy, editors of scholarly-led (or academic-lead) journals [20] have to know the scientific trends, common themes and research methods within their publishing communities. Journal establishment and maintenance might be motivated by offering an independent arena for specific topics and the research community, but also by strategic or political reasons, such as offering open access publication venues. Editors must offer continuity and stability to authors when establishing their brand and choosing their thematic scope throughout the years. Further, research trends can lead to the development and coining of new disciplines, such as gender studies, once considered a niche discipline (e.g. the OA journal Aspasia). Journal editors must react to research trends within their field, and need to have knowledge of current research agendas [21]. They will also have to consider the different action domains across disciplines that are typical for inter- and transdisciplinary science and learning processes [22]. This is reflected in the journal's name, strategy, and the submission guidelines: while catering to different disciplines, editors need to be aware of common research cultures and long-standing research areas, methodological consensus, or emerging themes within their chosen scope [8]. While informal processes undoubtedly occur, there is only little research that reflects on such strategies for editors. Hence, it is important to investigate the emerging topics trends and thematic clusters within research communities and to understand how they might change and

evolve over time. Our study offers an opportunity (i) to ground such reflections in empirical data and to relate the strategy to the output dimension, and (ii) to demonstrate that deep insights can be extracted from open access articles' content and metadata to improve our comprehension and to derive influential factors and drivers for journal development. Based on the above observations, our research questions for this study are: RQ1: What are the research topics of papers and their evolution over time? RQ2: What factors influence the research cooperations and research topics from a journal management and editorial perspective?

3 Methods

A threefold approach to study the topic was applied: a case study of a scholarly community represented by the platinum OA scholarly-led eJournal of e-Democracy and Open Government (JeDEM, www.jedem.org), data mining algorithms applied to the content and metadata of all articles published in JeDEM since 2009, and qualitative analysis.

3.1 Case study

The case study approach allows the investigation of a phenomenon within its real-life context and the use of multiple sources of evidence [23], such as publication data and the reflexive perspective of different editors. The case further defines the scholarly field the results are applicable to: an inter- and transdisciplinary field of technology and social studies.

The investigated journal, JeDEM, is published by the Department for E-Governance and Administration (Danube University Krems) and has an acceptance rate of 50% on average. Publication frequency is twice a year, and the journal is indexed in Scopus, EBSCO, DOAJ, Google scholar, and the Public knowledge project (PKP) metadata harvester. Articles are referenced with a digital object identifier (DOIs).

3.2 Data mining

Methods, techniques and algorithms for data-mining have already been successfully applied to such scholarly studies [20]. Text mining scholarly research refers to the extraction of implicit, potentially valuable information and patterns from natural language texts by identifying promising topics that characterize the document collections, document classifications based on predetermined keywords set, or the detection of a group of semantically close documents [24]. In our study, we apply the set of data mining and text mining algorithms to realize the following methodological steps. *First*, we preprocessed the scientific papers (title, abstract and content), including word normalization, stemming, removal of stop words, punctuations, and numbers. *Second*, each subset of the paper collections was pre-ordered over time (years) and divided into a disjoint set, then the top 30 keywords (and bigrams) were extracted. We consider these key-

words sets as a thematic context parameter, characterizing a particular epoch of research field development. *Third*, to derive the insights regarding thematic clusters within the analyzed paper collection, unsupervised k-means clustering algorithm was applied. Cosine similarity was adapted as a measure of thematic semantic closeness for clusters building [25]. The resulting thematic cluster (TC) were manually labelled based on their keywords set. *Fourth*, to deeper investigate the contextual structure of each TC, the LDA [26] topic model was used. To determine the optimal number of sub-topics for each cluster, the perplexity of a held-out set of papers has been calculated. As a result, 3 sub-topics were identified for cluster 1, and 4 each for clusters 2 and 3. The topics labels were determined based on a set of keywords describing them according to the LDA algorithm results.

3.3 Qualitative Analysis

Qualitative methodologies seek to portray a world in which reality is socially constructed, complex and ever changing [27]. The qualitative interpretation was conducted from a journal management and editorial perspective and represents the reflexivity inherent to auto-ethnographic methods [28]. An interpretative approach focuses on the analysis of meanings ascribed to data and the perceptions of a phenomenon, whist reflexivity enabled us to relate the trends in the data to experience and knowledge about cooperations, events, or advertising strategies that impact strategic decisions and publication.

4 Results

We focus on the main insights from the data on publications, namely: (1) the prominent research topics (and sub-topics) and their evolution over time and (2) the reflection on potentially influential factors from a journal management and editorial perspective.

4.1 Research topics, thematic clusters and sub-topics

Research topics and evolution over time
We identified 3 clusters of prominent thematic trends emerging in the collection of scientific papers. Table 1 presents the thematic clusters' (TC) labels, cluster proportion (CP, %) over articles collection, cluster description, and word clouds for the research topic clusters.

Table 1. Research topics clusters

Cluster 1	**CP, %**	**Cluster 2**	**CP, %**	**Cluster 3**	**CP, %**
Citizen engagement	24.49	Disruptive technology	47.96	Smart governance	27.55
Cluster Description					

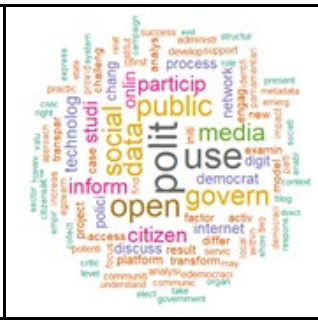
Electronic petitions, e-participation initiative & citizen participation	Open data, social media & social networking, disruptive technologies & digital transformation	Open data-driven government, smart city & co-creation
Cluster Wordcloud		
		

Figure 1 below shows the thematic evolution the research publihsed, where the identified clusters are highlighted by colours and their contextual evolution is represented by a set of major keywords (bigrams). The results enable us to define some "scientific turns" such as "disruptive technology" and "smart governance": (1) while the thematic citizen engagement (cluster 1) was present during the start-up phase of the journal (2009-2010), there is a trend towards disruptive technologies (cluster 2) as from 2011. Cluster 2 concerns many different innovations and disruptions, e.g. social media, parliamentary informatics and open data. From 2014, we also notice the emergence of the theme of smart governance (cluster 3), which became a more established cluster from 2016 onwards. Drawing on this kind of data can help determine the emergence of certain research foci and its representation in publications (such as 2014 for "smart" and 2011 for "open"). For example, following the Open Government Directive in 2009 [29] the investigated research community was quick to publish on this subject within a timeframe of two years.

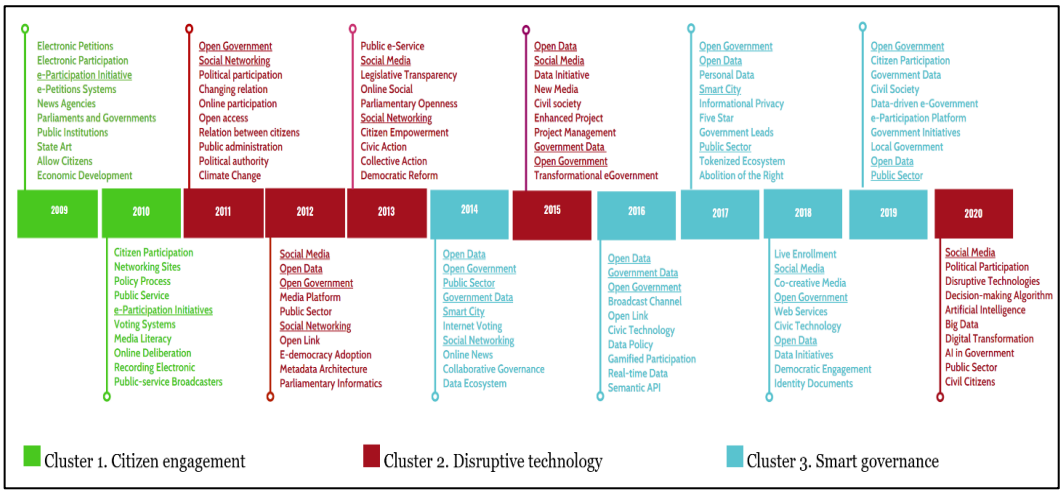


Fig. 1. Thematic clusters throughout the years

Regarding the main factors influencing the thematic development of the journal, our study noted: (1) Specific research topics addressing the call for papers, demonstrating a strong correlation relationship with identified thematic clusters in the corresponding year. Thus, in our case-study, Cramer's V statistic, calculated based on the data about correspondence between thematic clusters associated with the period and special issue topics, is equal to 8.75 and is statistically significant (p-value=0.0377); (2) Special offline events (mainly conferences) and the related materials, which are annually published in one of the journal issues, contribute to the continual evolution of e-democracy and e-government as a research domain in JeDEM.

The following main factors influencing article quantity were defined: (1) Offline events such as conference series, connected to the publishing community, could be an additional initiator of an increase in article quantity; (2) changing the journal title, as well as the title of the related community conference. In our case study, when the E-Democracy Conference (EDEM) changed its title to Conference for E-Democracy (CeDEM) in May 2011, this led to the disappearance of citizen engagement topics and, at the same time, led to the emergence of a new large thematic cluster named disruptive technology with “open government” as one of the most significant contextual bigrams. Research topics in calls for papers impact the regional dimension of the articles’ authors, if the title of the calls mentions the regional focus. In our case-study, from the years 2013 and 2015 onwards, the journal had special issues addressing Asian experiences, which led to a natural influx of articles from authors based in Asia.

Sub-topics and word co-occurrence in papers over time

For a deeper understanding of the nature of thematic trends, the Topic Modelling approach was used to identify the sets of subtopics for each of the clusters (Fig. 2). The

following insights were identified: (1) the subtopic of data openness undergoes a transformation stage from the Open Data concept in cluster 2 to Open Linked Data (or linked open data) in cluster 3; (2) the same phenomena can be noticed for the Open Government subtopic in cluster 2, which in cluster 3 evolved into an Open Data-driven Government concept; (3) there is a large predominance of subtopic such as Online Deliberation, E-Petitions, Open Linked Data in the context of Smart Cities, and Decision-making and Social Media.

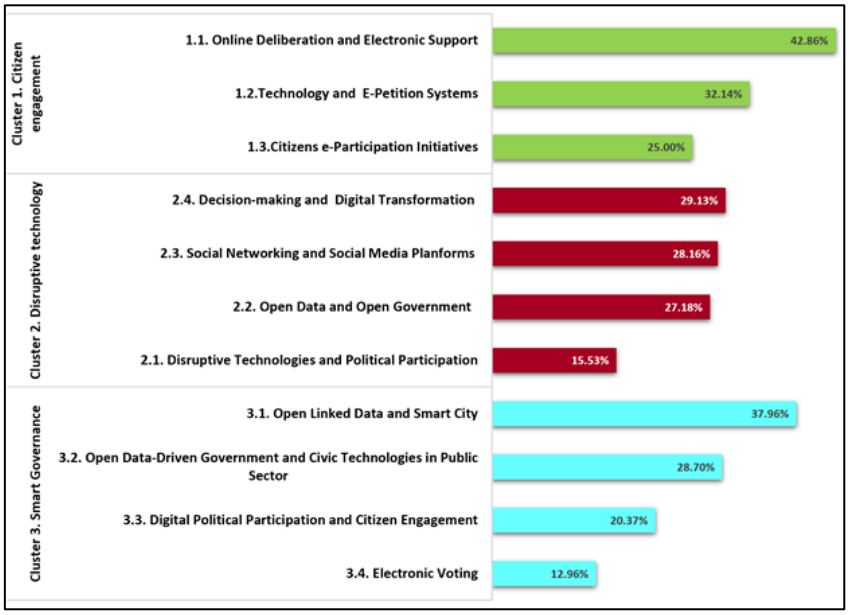


Fig. 2. Journal papers subtopics

To define a contextual profile of the research development of the scholarly community over the past 11 years, the analysis of the co-occurrence of words (Fig. 3) hinted at the anchoring of research themes within specific contexts, such as (1) governance and the public sector, as well as open, data and technology; but also (2) the context of participation and politics was more visible than the context of democracy or collaboration.

4.2 Influential factors of journal management

Based on our above analysis, we are able to derive and summarise four influential factors of journal development, related to its thematic focus, research cooperations and other challenges addressed by the journal on the visibility, accessibility and impact of the journal. Table 2 gives an overview on the most important factors, their description and the area of influence as detected in the analysis and evaluated by managing editors. Topical influence refers to the development of research areas and thematic clusters over time; regional influence to spatial ties of research cooperations; quantitative to the number of publications; and qualitative to the quality of submissions.

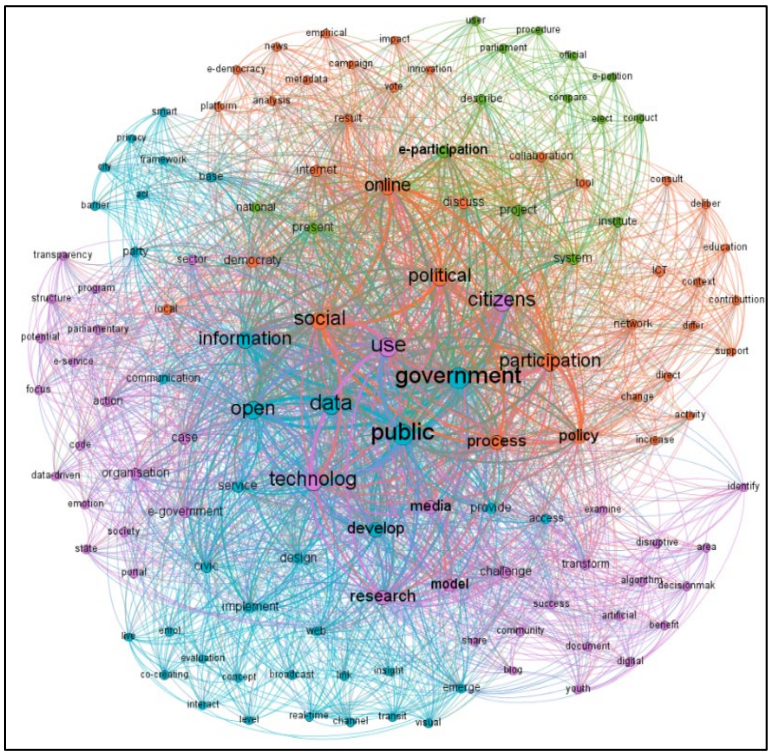


Fig. 3. Contextual profile of the scholarly community

Table 2. Influential factors of journal development

Influential factors	**Description**	**Area/Nature of Influence**
Research cooperations: Guest editors	Guest editors work on special issues and call for papers addressing specific research topics	Topical, regional
Research cooperations: Offline events	Mainly international scientific conferences	Topical, quantitative
Editorial strategy: Title	The title and naming of the journal, including strategic decisions regarding name change	Topical
Editorial strategy: Indexing*	Listing of the journal in indexing systems such as Scopus, EBSCO, DOAJ, Google scholar, and the Public knowledge project (PKP) metadata harvester.	Quantitative, qualitative

*Indexing systems were strongly perceived as one of the biggest factors enhancing the number of publications (e.g. with Scopus, in 2019) [30]

5 Discussion

Using data mining methods, editors can analyse whether the topics included in their call for papers match the topics the journal aims to publish or whether adaptations need to be made to better manage the expectations of future journal authors. The results gained help authors' publication choices by tracing the evolutionary stages of the study of scientific problems, and using a conscious analysis in the shaping of research agendas (f.i. what is involved, what is not). It offers the advantage to tie such activities to data gained from the publishing communities, and to reflect on scholarly developments, research trends and setting the agendas. On one hand, the research cooperations detected in this study seem to be tied to certain economic privileges of specific countries and showcase a certain hegemonic bias with expectable domination and lack of representation of some countries. One the other hand, we could identify specific research realities or interests. Journal editors can therefore use this approach to better target their managing and advertising efforts and include a spatial dimension in calls for papers, for instance by adapting the call or advertisement according to the themes and cooperations identified. We were able to confirm trends in research agendas and related connotations and timestamps, such as the use of "smart" (2014) or "open" (2011) as reflected in the publications of the investigated scholarly community. For editors, this can provide a starting point for reflecting on emerging trends, which influence journal strategies, such as the name of a journal, the scope of the calls for papers and announcements to the scholarly community. As evident in our data, consistent themes within the thematic clusters could be taken more into account in calls for papers and special issues in order to enhance the identification and reach of targeted authors.

The data mining method offers a fresh perspective and specific weighting of the editor evaluations, as, without this approach, journal editors might think that their topics are represented mostly equally throughout the call, or they might miss important topical developments or long-term emphasises. The topic modelling algorithm can be used for the formulation of specific sub-topics, f.i. for calls for papers for special issues, and confirm or contest the general topical orientation of a journal. For instance, for the editors, the strong prevalence of the topic e-petitions in publications was surprising, made visible by the data mining. This also points towards a potentially crucial development regarding the definition of research agendas, as strategic decisions can be supported by publication-based data.

6 Conclusion and Outlook

Our study was able to identify research cooperations, the most prominent research topics (defined as thematic clusters) of the studied journal, as well as their evolution over time. The most important influential factors were identified. From the methodological side, we demonstrated that deep and valuable insights can be extracted from freely available open access journal content and article metadata in order to improve the whole publishing community's comprehension of research themes, influential factors and

drivers of journal development. This approach was found to be useful for gaining insights concerning the influence of research cooperations and editorial decisions and might also be used by editors of all other types of journals to determine their strategy as well as consider the readers' demands for articles on particular research topics. Our next steps concern (1) the analysis of further data points such as regions, methods, rejection rate, or factors of openness (f.i. whether research data underlying the studies is made openly available), and (2) the validation of our results by comparison to the prominent topics emerging in the scientific papers published in other OA digital society journals. The limitations of this study lie in its focus on one case as well as the reflexive aspects of data interpretation. As this is partly a qualitative endeavour, there are always other means of interpretation and defining umbrella terms, f.i. for thematic clusters.

References

1. Edelmann, N., and Schoßböck, J.: 'Open access perceptions, strategies, and digital literacies: A case study of a scholarly-led journal', Publications, 8, (2020).
2. Alperin, J.P., Gomez, C.J., and Haustein, S.: 'Identifying diffusion patterns of research articles on Twitter: A case study of online engagement with open access articles', Public Understanding of Science, (2018).
3. Bach, T.A., and Ray-Sannerud, B.: 'Benefits of open access articles for industry', Nordic Perspectives on Open Science, 1, (2017).
4. Shehata, A., Ellis, D., and Foster, A.: 'Scholarly communication trends in the digital age', The Electronic Library, (2015).
5. Agate, N.: 'From Evaluated Outputs to Values-Embedded Practices', Septentrio Conference Series, 2020.
6. European Science Foundation: www.coalition-s.org, last accessed 01.06.2020.
7. International Science Council: https://council.science/current/blog/could-plan-s-be-a-turning-point-for-global-open-science-interview-with-robert-jan-smits/, last accessed 23.01.2019.
8. Johannessen, M.R., and Berntzen, J.: 'A Decade of eParticipation Research: An Overview of the ePart Conference 2009-2018', in: 'Book A Decade of eParticipation Research: An Overview of the ePart Conference 2009-2018' pp. 3-14 (2019).
9. Götz, N., and Marklund, C.: 'The Paradox of Openness', The Paradox of Openness, 2014
10. 10 Wessels, B., Finn, R., Sveinsdottir, T., and Wadhwa, K.: 'Open Data and the Knowledge Society', (2017).
11. Parycek, P., and Schossböck, J.: 'Adopting a new political culture: obstacles and opportunities for open government in Austria', in: 'Book Adopting a new political culture: obstacles and opportunities for open government in Austria' pp. 210-236 (2014).
12. Chesbrough, H.W.: 'The era of open access', Managing Innovation and Change, 127, pp. (2016).
13. Sachs, M., and Parycek, P.: 'Open government - information flow in Web 2.0', European Journal of ePractice, 9, pp. 1-70 (2010).
14. Sæbø, Ø., Rose, J., and Flak, L.S.: 'The shape of eParticipation: Characterizing an emerging research area', Government information quarterly, 25, (3), pp. 400-428 (2008).
15. Rodríguez Bolívar, M.P., Alcaide Muñoz, L., and López Hernández, A.M.: 'Scientometric Study of the Progress and Development of e-Government Research During the Period 2000–2012', Information Technology for Development, 22, pp. 36-74 (2016).

16. Chaudhuri, R., Chavan, G., Vadalkar, S., Vrontis, D., and Pereira, V.: 'Two-decade bibliometric overview of publications in the Journal of Knowledge Management', in: 'Book Two-decade bibliometric overview of publications in the Journal of Knowledge Management' (2020).
17. Sharma, A., Rana, N.P., and Nunkoo, R.: 'Fifty years of information management research: A conceptual structure analysis using structural topic modeling', International Journal of Information Management, 58, pp. 102316 (2021).
18. Das, S., Dixon, K., Sun, X., Dutta, A., and Zupancich, M.: 'Trends in transportation research exploring content analysis in topics', in: 'Book Trends in transportation research exploring content analysis in topics', pp. 27-38 (2017).
19. Schoßböck, J., Edelmann, N., Rizun, N., and Zuiderwijk, A.: 'Scholarly Research and Publications Over Time: Identifying Trends for an Open Access Journal by Applying Data-Mining Methods', in: 'Book Scholarly Research and Publications Over Time: Identifying Trends for an Open Access Journal by Applying Data-Mining Methods' (2020) .
20. Dridi, A., Gaber, M.M., Azad, R.M.A., and Bhogal, J.: 'Scholarly data mining: A systematic review of its applications', Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, pp. 1-23 (2020).
21. Teixeira Da Silva, J.A.: 'Conflicts of Interest Arising from Simultaneous Service by Editors of Competing Journals or Publishers', Publications 2021, Vol. 9, (2021).
22. Viale Pereira, G., Estevez, E., Cardona, D., Chesñevar, C., Collazzo-Yelpo, P., Cunha, M.A., Diniz, E.H., Ferraresi, A.A., Fischer, F.M., and Cardinelle Oliveira Garcia, F.J.S.: 'South American expert roundtable: increasing adaptive governance capacity for coping with unintended side effects of digital transformation', 12, (2), pp. 718 (2020).
23. Yin, R.K.: 'Case Study Research: Design and Methods (5th Edition)' SAGE Publications, (2014).
24. Hassani, H., Beneki, C., Unger, S., Mazinani, M.T., and Yeganegi, M.R.: 'Text mining in big data analytics', Big Data and Cognitive Computing, 4, pp. 1-34 (2020).
25. Qayyum, F., and Afzal, M.T.: 'Identification of important citations by exploiting research articles' metadata and cue-terms from content', Scientometrics, 118, pp. 21-43 (2019)
26. Blei, D.M., Ng, A.Y., and Edu, J.: 'Latent Dirichlet Allocation ', in: 'Book Latent Dirichlet Allocation ' pp. 993-1022 (2003.).
27. Sloan, A., and Bowe, B.: 'Phenomenology and hermeneutic phenomenology: The philosophy, the methodologies, and using hermeneutic phenomenology to investigate lecturers' experiences of curriculum design', Quality and Quantity, 48, pp. 1291-1303 (2014).
28. Guillemin, M., and Gillam, L.: 'Ethics, Reflexivity, and "Ethically Important Moments" in Research', Qualitative Inquiry, 10, pp. 261-280 (2004).
29. Lathrop, D., and Ruma, L.: 'Open Government. Collaboration, Transparency, and Participation in Practice', (2010).
30. Pranckutė, R.: 'Web of Science (WoS) and Scopus: The Titans of Bibliographic Information in Today's Academic World', Publications, 9, (1), pp. 12 (2021).