Information Services & Use 34 (2014) 241–248
DOI 10.3233/ISU-140744
IOS Press

Researchers' information needs in the bibliographic database: A literature review

Morgana Carneiro de Andrade [a,*] and Ana Alice Baptista [b]

[a] *Doctoral Program in Technology and Information Systems, University of Minho, Guimarães, Portugal*
[b] *Algoritmi Research Center / Information Systems Department, University of Minho, Guimarães, Portugal*

Abstract. This article presents a literature review whose aim was to identify the reported information needs of researchers when they consult bibliographic databases. Initially, 192 articles were retrieved using Scopus, Web of Science and Google Scholar databases. After applying the criteria for exclusion, the number of articles was reduced to 16, which is already an indicator of the small number of studies on this specific topic. The results show that it is hard to identify the information needs of researchers. They also show that the researchers have been requiring information with a higher degree of granularity. We conclude that although the available studies provide important information about the researchers' information needs and hints on how to address them, there is a need for more in-depth studies. The results of these deeper studies may be useful to serve as an indication for the creation of new procedures and tools, including those based on new metadata elements drawn to improve search results on Linked Open Data tools.

Keywords: Information needs, bibliographic databases, researchers

1. Introduction

The theme of information needs is present since the first studies in the field of Library and Documentation Science, and subsequently in Information Science. With the advent of the Internet there was an increase of studies on this topic, especially with focus on information services, such as digital libraries and bibliographic databases.

According to Kuruppu and Gruber [17] "understanding information needs, information-seeking behavior and information use of researchers is challenging" and it gets more complicated as they play several roles (researcher, teacher, administrator, ...), their needs and interests change over time and they are "continuously" affected by technological advances.

The huge amount of information present on the Internet and the diversity of services have paradoxically contributed to hinder the identification of the most relevant papers. In order to find relevant information in less time, it is required from the user that he knows "[...] what to get, from where to get, how to get it" [1]. These questions are related to information needs [1] that in the scope of this literature review are defined as "a state or process started when one perceives that there is a gap between the information and knowledge available to solve a problem and the actual solution of the problem" [23].

*Corresponding author: Morgana Carneiro de Andrade, Information Systems Department, University of Minho, 4810-058 Guimarães, Portugal. Email: morganaandrade@hotmail.com.

One of the forerunners in this area was Taylor [30], who in the article "The process of asking question" brought insights on information needs that are relevant till today. The author proposes four levels of information needs:

- First level – the conscious and unconscious needs, which when identified refer to the "perfect question".
- Second level – the conscious needs that are poorly defined which will be made understandable from interactions with other people.
- Third level – the conscious needs that are well defined, but that may not be properly "translated" into the information system.
- Fourth level – the conscious needs that are well defined and may be "translated" into the language of the information system in a way they can be processed.

From the 3rd and 4th approaches, Taylor lists some aspects that affect the man–machine relation: (a) system organization, which includes input and output characteristics; (b) types, complexities and characteristics of the subject related to the question; and (c) the researcher's competence.

The "internal organization", part of the system organization and its input characteristics, as the author understands it, corresponds to the access points, which in his view are related to the degree of sophistication in the use of terms, the depth of analysis and indexing and the level of specificity. These access points could assume a "multidimensional space", ranging from empirical data to theoretical concepts, by way of descriptive data, experimental evidence, historical material, results analysis, interpretation of descriptive categories of information. For Taylor, the way the information service can be exploited has implications on the way the researcher formulates his questions and on number of relevant answers he gets from the system.

In Taylor's second to forth level it is present the role of the librarian as an interpreter and a translator of user needs to the system. Due to technological advances and the familiarity of researchers with the technology, this role of the librarian is becoming less present. Because users do not resort so often to the presence of the librarian, information services need to offer functionalities that meet their needs by providing access points that allow them to retrieve relevant documents. Therefore, for this literature review, we are especially focused on one of the items identified by Taylor: the way the services are internally organized, classified, and indexed, and their access points. For this purpose we consider that the researchers are aware of what they want, but for their information needs to be met they depend on the "internal organization" of information services.

In this sense, the identification of access points that meet the user's needs may be useful to serve as an indication for the creation of new procedures and tools, including those based on metadata elements drawn to improve search results on Linked Open Data tools. The tendency to increase granularity at the description level allows that initiatives such as the W3C's "The RDF data cube vocabulary", "Data Catalog Vocabulary", linked data can be more easily adopted [12,13]. Cyganiak, Reynolds and Tennison [12], by publishing guidelines at the W3C initiatives to the multidimensional publication of data, confirm the thought of Taylor [30], aired 52 years ago as Taylor also stressed the need for a multidimensional space to the empirical data and theoretical concepts. Also according to Ismail and Kareem [16], the Web does not provide support for the novice researcher and thus one of the solutions might be that information services become semantically interoperable. One suggestion to resolve some of these issues was the use of semantic web technologies.

This article has the following structure: Section 2, Research design, which contains the description of the strategy for conducting the search. Section 3, Results, which are presented and discussed in or-

der to clarify what has been developed so far to meet the information needs of researchers. Section 4, Conclusion, which brings the final considerations.

2. Research design

We started by identifying studies addressing the needs of researchers when consulting bibliographic databases and what has been developed to meet this need. The databases used for the search were: Scopus, Web of Science, Networked Digital Library of Theses and Dissertations (NDLTD), Library & Information Science, and Technology Abstracts (LISTA) and Google scholar. The keywords used were: information needs × bibliographic database; information seeking behavior × scientific communication; user profile × scientific information; scientific articles × user's needs; retrieval × bibliographic database × user's needs; user studies × bibliographic database; information needs × scientific communication.

The above terms were used in the fields of keywords/topics in databases, and in Google scholar the title field was used. When searching the databases, we have made the following choices: temporal limits: none; language: English, Portuguese and Spanish. The same procedure was used for Google scholar, but in this case the results' analysis was limited to the first 100 most relevant articles.

The papers resulting from this search were selected based on the titles and abstracts (step 1). The articles cited in these studies were also selected based on titles and abstracts (step 2). We adopted the same procedure to select the articles that cite the ones retrieved in the first step (step 3). This way, we sought an outcome with greater coverage but without losing relevance to the theme.

In order to maintain consistency of the proposed study we excluded articles that we considered to be out of scope, such as the ones on: software applications, information behavior, relevance analysis, technical analysis of information retrieval systems. In addition, we identified a few studies whose approach was more general, and which analyzed aspects like kind of used sources, language or subject [8,9]. In this case, we chose not to include these articles in the results, for not bringing relevant information as to achieve the purpose of this literature review. Inclusion was restricted to scientific articles. Based on these criteria, the number of articles that were actually aligned with the objective of the present study was reduced from 192 to 16.

3. Results

The analysis of the 16 articles results in a panorama of what has been researched on the researchers' information needs when seeking in bibliographic databases as presented in Table 1.

The four main aspects that arise from this analysis are: components of articles, indexing and metadata, domain, and user profile.

- *Use of components of scientific articles as a way to enhance search results*
  In the context of this study, components of articles represent physical or logical structures of a document [5,6,27]. Tables and figures are examples of physical components, whereas the data resulting of some experiment represent logical structures or narratives [5].
  Bishop [5] conducted a study that examined how the components of a scientific paper are identified, stored and utilized by users in digital libraries. In her study she used DeLIver which is aimed to allow researchers at the University of Illinois to search for components of documents. Bishop

Table 1
Subjects and aspects approached in articles

Authors	Aspects				
	Subject	Components	Domain	Indexing/Metadata	User profile
Amato and Straccia, 1999 [1]	IN				×
Bates, Wilde and Siegfried, 1993 [3]	ISB		×		
Bates, 1996 [2]	ISB		×		
Bishop, 1998 [5]	ISB	×	×	×	
Bishop, 1999 [6]	IN	×		×	
Borgman, 1986 [7]	IN			×	
Courtright, 2007 [10]	ISB				×
Crowston and Kwasnik, 2003 [11]	IN			×	
Dogan et al., 2009 [14]	IN		×		
Hjørland, Nielsen and Williams, 2001 [15]	ISB	×		×	
Ismail and Kareem, 2011 [16]	IN			×	
Lee and Downie, 2004 [18]	IN			×	
Markey, 2007 [20]	ISB		×		
Markey, 2007 [21]	ISB		×		
Rowlands, 2007 [26]	ISB		×		
Sandusky and Tenopir, 2008 [27]	ISB	×		×	×

Note: IN – information needs; ISB – Information Seeking Behavior.

[6] reported that researchers appreciate the use of specific components in specific situations. The researchers also demonstrated the importance of using the components to decide which articles resulting from the search process should be read. Bishop's research, as she discusses, is aligned with the principles advocated by Paul Otlet [5,6].

According to Otlet, as chemistry researchers moved from the analysis of molecules to atoms, efforts should converge to think of ways to allow science to have access to specific parts of the content of the publications. In a visionary way, in the early decades of 1900, Otlet said: "methods will be found to index works quickly and completely in order to permit the retrieval, instantly and without trouble or difficulty, of the substance of what each publication contributes to knowledge" [24]. Rayward [25] adds that these statements referred to the atoms of information that could be reconfigured in order to meet the information interests and needs of users.

Hjørland, Nielsen and Williams [15] cite Bishop's results to highlight the "discussion of the need to replace the traditional linear structures in documents with a free combination of 'info-bricks"', which are "the extraction of individual facts and ideas as separate units" [5]. The authors also mention the studies of Al-Hawamdeh et al., Al-Hawamdeh and Willett and Lalmas and Ruthven, whose studies support the usage of components of texts as subject access point (SAP).

- *Relationship with the domain*

  The domain is considered by several authors to be a factor that interferes with the researchers' information needs and with the way the users proceed when searching. For example, it could be expected that faceted search conjugated with Boolean operators in online databases would return in better results independently of the domain of the researcher. However, according to Bates [2] research in the humanities, end users find it difficult to perform those kinds of searches, as well as finding optimal results.

This idea is reinforced by Markey [20,21], which considers that experts in a particular field seek high degree of accuracy, limiting the search field. Their strategies are based on identifying clues contained in any word or phrase in the title, the name of an author, a variable, a test or a particular research center, reducing the number of items retrieved, but getting high degree of relevance.
The domain is also related to the modification of the researchers' information needs through, for example, the arising of new more specific research fields [26]. An example is given by the study of Dogan et al. [14], who found that the majority of searches in PubMed in the subject field are performed by gene, protein and/or disease. This is particularly interesting if we take into account that just a few decades ago instead of gene or protein, the searches used terms related to some kind of treatment or diagnosis.

- *Process of indexing and access points*
The users' information needs are often articulated ambiguously not only in what concerns the terms, but also in what concerns the usage of the structure of the system being searched. The process of conducting a search involves issues such as syntax, semantics, structure and purpose of the search; how the access points are used to reduce and expand the results; search alternatives, and if the fault in the search derives from a personal or a system failure [7].
Sandusky and Tenopir [27] claim that there is a great difficulty for researchers to identify relevant articles, as these are still indexed in a way that does not comprise detailed information about the document. Reference is made to ProQuest CSA, which developed a prototype system that provides detailed information by indexing individual components of the articles. This model allows the realization of Boolean searches using author, title, statistics, geographic and taxonomic terms. Searches can be refined by maps, figures, photographs, type of article or predictive models. This procedure is related to the increased level of granularity used in the design of databases. Bishop [5] complements Sandusky and Tenopir claims by stating that the indexing of articles is highly standardized, including the identification of the author, title, abstract and keywords. Bishop further argues that existing items in the articles that are not explored in the search fields may promote the retrieval of the most relevant documents.
Bates, Wilde and Siegfried [3] present the results of the analysis of online bibliographic databases usage. These results contributed to improvements of (1) facets of the Styles and Periods Art And Architecture Thesaurus, with the inclusion of some terms that were previously neglected in the thesaurus; (2) database structure, with the inclusion of names of artists, based on the variation of names and terms that identify the academic disciplines in the database. For these authors, only a good indexing enables the obtention of high precision and relevance search results.
Like Bishop [5,6], Crowston and Kwasnik [11] also mention the issue of indexing and underline its relationship with a critical factor: context. Crowston and Kwasnik identify the difficulty to meet the users' needs with relevant results due to issues such as inaccurate or incomplete information in the databases which are related to indexing.
Studies in which the indexing process is approached are still incipient in what concerns the possibilities of expanding the level of analysis and granularity. This does not go along with some initiatives in which the use of descriptive metadata ceases to be limited to the identification of title, author and subject. Items such as outcome treatment, risk factors and conclusion as topic begin to be explored either manually or automatically, especially with the use of Semantic Web tools [28,29].
Hjørland, Nielsen and Williams [15] add that the different structures that exist in texts have consequences in the search. The search strategy in scientific databases can vary according to specific

access points such as methodological issues or findings, which are considered topics of greatest interest.

As regards to description and recovery, there are metadata or access points, which result from the activity of indexing documents, i.e., access points determine objective possibilities of document retrieval by users through algorithmic or automated procedures [15]. In this sense, studies have shown that software agents can process articles' information using semantic treatment as a new form of content exploitation [19].

Lee and Downie [18] developed a study in relation to music information retrieval (MIR) in the field of music digital libraries (MDL). One of the quests was "What types of metadata or access points should be provided to users." Regarding this quest, they identified the need for new types of metadata as access points that include information about music or music objects and information that contextualize searches of users' the real-world.

Hjørland, Nielsen and Williams [15] stress that the SAP are critical to the retrieval of documents. Thus, if these access points are not supplied by the information services, users may not be able to retrieve the documents they need. Some studies also show that increasing the granularity of the information in the information services contributes to the improvement of the search, since it provides a wider range of access points [2,5,7,15,18,27]. By restraining these claims, we argue that a higher degree of granularity of access points is expected to promote high precision and relevant search results that are more aligned with users' needs. This is related to Otlet's claims about information atomization. As Menzel [22] points out, "the expressed and conscious wants of individuals in any area are constrained by their perception of what is feasible". As what is feasible currently in most information services does not include high granular searches and access points, researchers' needs may not be properly externalized, forwarding to Taylor's 3rd level of information needs. Perhaps this is what better explains why there are not enough studies on researcher's information needs regarding bibliographic searches. Hence, it is relevant to explore the perspective of the researcher, what he needs and what access points would meet his needs.

- *User profile*

  The concern with the user is externalized by Courtright [10], when she observed that there was a change in the direction of the studies in what concerns information needs. The studies used to focus on a system-centered model and were redirected to a user-centric model, where research focuses on the information of the participating actors.

  Amato and Straccia [1] and Courtright [10] consider that the information needs may vary depending on the type of user and on the context in which these needs are analyzed or required. Thus, one of the concerns in the development of information needs studies is to establish which user profile is to be analyzed.

  Additionally, researchers are users whose experience in developing searches presupposes information with the highest level of specificity and relevance. Sandusky and Tenopir [27] found that, for this type of user, the identification of relevant items in a short time reflects on certain types of behavior as, for instance, the time allocated for reading articles. These results are corroborated by the studies of Berghel et al. [4], Shotton [29], Tenopir et al. [31,32].

4. Conclusion

This literature review presented some interesting and promising aspects related to information needs of researchers and which are worth exploring: components of articles, indexing and metadata, domain, and

user profile. It is clear that Otlet was indeed beyond his time in what concerns his suggestions regarding the need to access to specific parts of documents quickly and completely. However, most authors reported difficulties in identifying the information needs of researchers. Perhaps these difficulties are related to the difficulty that researchers themselves have in expressing their real needs when using the information services, forwarding them to Taylor's level 3 of information needs.

Another aspect that was identified with this literature review concerns Taylor's level 4 of information needs, i.e., the conscious needs that are well defined and may be "translated" into the language of the information system in a way they can be processed. There is a tendency to use a high degree of information granularity that can be optimized by the use of semantic Web and linked data services that are able to provide more relevant search results in less time and with less effort of the researchers.

Regarding the development of this study, it is worth noting the difficulties and limitations we found. Some of the difficulties we had are related to some problems identified by authors cited in this literature review. In particular, we had problems that are related to the indexing process of the databases (the inclusion of specific keywords), which caused a limitation in the number of retrieved articles. This implied the need to adjust the search strategy in order to identify other relevant articles for this literature review.

As future work, we suggest that studies should focus on the information needs that could be met by the use of semantic Web technologies. For example, How do researchers enjoy the potential offered by these technologies? How are these technologies being used by the information services? What kind of new services could be created using these technologies to better meet the researchers' needs?

Acknowledgements

We thank for Espírito Santo Federal University, Brazil; CAPES Foundation, Ministry of Education of Brazil for financial support to our research activities.

Part of this work has been supported by FCT – Fundacão para a Ciência e Tecnologia within the Project Scope: PEst-OE/EEI/UI0319/2014".

References

[1] G. Amato and U. Straccia, User profile modeling and applications to digital libraries, in: *Research and Advanced Technology for Digital Libraries*, 1999, Springer, Berlin, pp. 184–197.

[2] M.J. Bates, The getty end-user online searching project in the humanities: Report No. 6: Overview and conclusions, *College and Research Libraries* **57**(6) (1996), 514–523.

[3] M.J. Bates, D.N. Wilde and S. Siegfried, An analysis of search terminology used by humanities scholars: The Getty Online Searching Project Report No. 1, *The Library Quarterly* (1993), 1–39.

[4] H. Berghel, D. Berleant, T. Foy and M. McGuire, Cyberbrowsing: Information customization on the Web, *Journal of the American Society for Information Science* **50**(6) (1999), 505–513

[5] A.P. Bishop, Digital libraries and knowledge disaggregation: the use of journal article components, in: *Proceedings of the Third ACM Conference on Digital Libraries*, 1998, pp. 29–39.

[6] A.P. Bishop, Document structure and digital libraries: how researchers mobilize information in journal articles, *Information Processing and Management* **35**(3) (1999), 255–279.

[7] C.L. Borgman, Why are online catalogs hard to use? Lessons learned from information-retrieval studies, *Journal of the American Society for Information Science* **37**(6) (1986), 387–400.

[8] J.J. Calva González, Las necesidades de información del usuario en la automatización de unidades de información, *Revista Biblioteca Universitaria* **1**(1) (2009), 6.

[9] J.J. Calva-González, Las necesidades de información de los investigadores del área de humanidades y ciencias sociales, *Revista General de Información y Documentación* **13**(2) (2003), 155–180,

[10] C. Courtright, Context in information behavior research, *Annual Review of information Science and Technology* **41**(1) (2007), 273–306.
[11] K. Crowston and B.H. Kwasnik, Can document-genre metadata improve information access to large digital collections, *Library Trends* **52**(2) (2003), 345–361.
[12] R. Cyganiak, D. Reynolds and J. Tennison, The RDF data cube vocabulary, W3C candidate recommendation 2013, available at: http://www.w3.org/TR/2013/CR-vocab-data-cube-20130625.
[13] Data Catalog Vocabulary 2014, available at: http://www.w3.org/TR/vocab-dcat/.
[14] R.I. Dogan, G.C. Murray, A. Névéol and Z. Lu, Understanding PubMed® user search behavior through log analysis, *The Journal of Biological Databases and Curation* **2009** (2009).
[15] B. Hjørland, L.K. Nielsen and M.E. Williams, Subject access points in electronic retrieval, *Annual Review of Information Science and Technology* **35** (2001), 249–298.
[16] M.A. Ismail and S.A. Kareem, Identifying how novice researchers search, locate, choose and use web resources at the early stage of research, *Malaysian Journal of Library and Information Science* **16**(3) (2011).
[17] P.U. Kuruppu and A.M. Gruber, Understanding the information needs of academic scholars in agricultural and biological sciences, *The Journal of Academic Librarianship* **32**(6) (2006), 609–623.
[18] J.H. Lee and J.S. Downie, Survey of music information needs, uses, and seeking behaviors: Preliminary findings, *ISMIR* **2004** (2004), 5.
[19] C.H. Marcondes, M.A.R. Mendonça, L.R. Malheiros, L.C. Da Costa and T.C.P. Santos, Ontological and conceptual bases for a scientific knowledge model in biomedical articles, *RECIIS* **3**(1) (2009).
[20] K. Markey, Twenty-five years of end-user searching, Part 1: Research findings, *Journal of the American Society for Information Science and Technology* **58**(8) (2007), 1071–1081.
[21] K. Markey, Twenty-five years of end-user searching, Part 2: Future research directions, *Journal of the American Society for Information Science and Technology* **58**(8) (2007), 1123–1130.
[22] H. Menzel, The information needs of current scientific research, *The Library Quarterly* **34**(1) (1964), 4–19.
[23] S.V. Miranda and K. Tarapanoff, Information needs and information competencies: a case study of the off-site supervision of financial institutions in Brazil, *Information Research* **13**(2) (2008), 5.
[24] P. Otlet, The science of bibliography and documentation, in: *International Organisation and Dissemination of Knowledge: Selected Essays of Paul Otlet*, W.B. Rayward, ed., Elsevier for the International Federation of Documentation, Amsterdam, 1990.
[25] W.B. Rayward (ed.) Introduction, in: *International Organisation and Dissemination of Knowledge: Selected Essays of Paul Otlet*, 1990, Elsevier for the International Federation of Documentation, Amsterdam, pp. 7–10.
[26] I. Rowlands, Electronic journals and user behavior: A review of recent research, *Library & Information Science Research* **29**(3) (2007), 369–396.
[27] R.J. Sandusky and C. Tenopir, Finding and using journal-article components: Impacts of disaggregation on teaching and research practice, *Journal of the American Society for Information Science and Technology* **59**(6) (2008), 970–982.
[28] Scientific Data 2014, available at: http://www.nature.com/scientificdata/.
[29] D. Shotton, Semantic publishing: the coming revolution in scientific journal publishing, *Learned Publishing* **22**(2) (2009), 85–94.
[30] R.S. Taylor, The process of asking questions, *American Documentation* **13**(4) (1962), 391–396,
[31] C. Tenopir, D.W. King, S. Edwards and L. Wu, Electronic journals and changes in scholarly article seeking and reading patterns, *ASLIB Proceedings* **61** (2009), 5–32.
[32] C. Tenopir, D.W. King, J. Spencer and L. Wu, Variations in article seeking and reading patterns of academics: What makes a difference?, *Library and Information Science Research* **31**(3) (2009), 139–148.