A reference ontology for digital scientific journals applied to systematic literature review processes

Uma ontologia de referência para periódicos científicos digitais aplicada em processos de revisão sistemática da literatura

Fernando Benedet GHISI
Gleisy Regina Bóries FACHIN
Marcos Henrique dos SANTOS
Denilson SELL
Gregório Jean Varvakis RADOS

Abstract

This paper presents possible approaches to the use of a reference ontology for digital journals, supporting bibliographic processes in systematic literature reviews. The benefits are highlighted of using specialized services through "batch" or "on-the-fly" processing of information from different repositories, such as recognized and indexed databases and portals or specific websites. It is concluded that the use of reference ontology enables the creation of services that ensure greater interoperability between different repositories, allowing a more inclusive and accurate retrieval of information, as it standardizes the concepts related to the access points of scientific journals.

Keywords: Information retrieval. Journal. Reference ontology. Systematic review of the literature.

Resumo

O presente artigo apresenta possíveis abordagens para o uso de uma ontologia de referência para periódicos científicos digitais, apoiando os processos de levantamento bibliográfico em revisão sistemática da literatura. Destacam-se os benefícios da utilização de serviços especializados por meio de processamentos em "batch" ou "on-the-fly" de informações provenientes de diferentes repositórios, como bases de dados reconhecidas e indexadas e portais ou sites específicos. Conclui-se que a utilização de ontologia de referência possibilita a criação de serviços que asseguram maior interoperabilidade entre diferentes repositórios, permitindo a recuperação da informação de forma mais abrangente e precisa, através da padronização dos conceitos relacionados aos pontos de acesso de periódicos científicos.


1 Mestrandos, Universidade Federal de Santa Catarina, Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento. Campus Universitário Reitor João David Ferreira Lima, Trindade, 88040-970, Florianópolis, SC, Brasil. Correspondência para/Correspondence to: F.B. GHISI. E-mail: fernandoghisiegc.ufsc.br.
2 Professora Doutora, Universidade Federal de Santa Catarina, Departamento de Ciência da Informação. Florianópolis, SC, Brasil.

Received on 19/12/2011, resubmitted on 4/5/2012 and approved on 23/5/2012.
Introduction

Since the earliest inscriptions and cave drawings made by man, the need to locate, understand and disseminate information has been present in the evolution of human civilization. During the course of history, information has been represented, stored and recovered in some way. Likewise, the organization of documents and information has accompanied the evolution of society, and has been centered on the field of Librarianship, Information Science and Computing, beginning with the first libraries and advancing to modern databases, directories and repositories (Fachin, 2011).

In a digital context, the knowledge areas are responsible for researching, developing, applying and validating resources as a way to ensure that they are self-sustaining. For as long as they have existed, the knowledge areas have been recording their creations and actions, especially in scientific publications. Considering that the evolution, resources and requirements of the digital world of scientific information are constantly changing, new tools, standards, systems and movements - such as the Open Access and Open Archives Initiative movements, discussed by many authors, such as Guédon (2010) - have been created. Therefore, irrespective of the field, theme, group, profession, language or country, everyone is faced with a proliferation of technologies related to facilitating communication and availability, where one goes from intensive reading to extensive, intertextual and hypertextual reading (Miranda, 2010).

Nonetheless, undoubtedly there is constant and increasingly faster proliferation of information and technological resources on the Web every day. Moreover, every field of knowledge invests in its own sources of referential information, in order to meet users’ needs, adopting, customizing or developing their own applications, without concern about interaction or interoperability with other systems (databases, digital libraries, repositories, portals or specific sites). The research groups, entities and governments are responsible for developing actions in order to maintain the consolidation of quality, certification and legitimation of information, as well as the efficiency of technological resources for broad and accurate recovery of such information. This is especially important in the scientific context, where the "efforts of the organizations to consolidate rules, regulations and standards of scientific publishing occur worldwide, with the aim of ensuring its standardization and intercommunicability", according to Miranda (2010, p.10).

Among the initiatives to increase the formality and credibility of researches supported by bibliographical reviews, there are "systematic literature reviewing methods" (Centre for Reviews and Dissemination, 2009; Higgins; Green, 2011). In one of the steps, a systematic search for relevant works is made, according to the study subjects on a pre-selected list of publications. This step is directly related to the mechanisms of information retrieval - and the range, quality and precision of these can have a great impact on the entire process. In this sense, one of the main problems or challenges of the Information Retrieval Systems - and more recently, of the mechanisms of intelligent information retrieval, which use intelligent agents - is only to retrieve the relevant documents looked for by the user.

To achieve proper information retrieval, the information must be processed and organized. Classifications, thesaurus and taxonomies are among the most used techniques that can be used in the treatment and organization of information. Zeng and Qin (2008) and Campos (2009) have stated that the structuring of digital documents may be done through the use of metadata. In this case, it would allow the search to be conducted in different ways: term in the title, term in the body, type of media, type of archive, period of publication, place of publication, language, among others. As a basic function, metadata provides information about the digital information and empowers an accurate retrieval, by the exchange of data among the information systems. More recently, with the emergence of the Semantic Web, ontologies have been increasingly applied to support the retrieval of significant information, according to a specific domain (Gruber, 1996; Guarino, 1998).

According to Bräscher and Carlan (2010, p.153), knowledge organization systems handle representations of “knowledge domains that delimit the meaning of terms in the context of areas, establish conceptual relationships that help position a concept in the conceptual system and are used as tools for organizing and retrieving information”. The adoption of standardized
concepts in identifying the resources available on the Web (semantic markup) aims to enable the information collection agents operating in the network (Web crawling, Web spider, Web robot) to become capable of intelligently understanding and processing the contents described and represented by the metadata. Campos (2009) collaborated in this discussion by highlighting the importance of terminological studies, multidisciplinary research and systematic studies in order to improve the standards used by the search agents.

In this context, from the problem related to the recovery of scientific information, the aim of this article was to present possible approaches to better use the vast amount of scientific material available on the Web, contributing to systematic processes for the review of literature. In addition to elucidating the use of Semantic Web resources and techniques (such as the use of reference ontology), some steps of the systematic review process that could be further automated by the use of more integrated and intelligent services, will be identified. Thus, according to Creswell (2010), this study is characterized as a qualitative, exploratory and applied research, because it seeks to model a segment of knowledge in an interdisciplinary way, using literature, operationalizing a reference ontology to standardize the essential elements of scientific journals, and demonstrating its applicability in scenarios of information retrieval that are present in systematic review of the literature.

Systematic review of literature

In the information and knowledge era, with the Web representing a huge repository of documents and information, researchers and scientists are facing an informational context that includes a miscellany of works without scientific quality or relevance. Databases - such as Scopus, Web of Science, Science Direct and EBSCO - try to mitigate part of this problem, including documents according to some quality of criteria and providing specialized search environments. However, when researchers propose a survey question, and investigate related previous publications on which to base their studies, or to support their hypotheses, they may not have had access to all the studies. They may even tendentiously leave out some studies with differing viewpoints on purpose. To attenuate this problem and increase the formality and credibility of bibliographical reviews, as well as increase their application, the methods of systematic review of literature have emerged.

The systematic review is characterized as an investigation of the evidence of a clearly formulated question using systematic and explicit methods to identify, select and critically evaluate previous relevant researches, extracting and analyzing data from studies that are included in the review (Centre for Reviews and Dissemination, 2009). For Sampaio and Mancini (2007, p.84), systematic review is a form of research that uses the literature on a particular theme as a data source, providing a “summary of evidence related to a specific intervention strategy, by applying systematic and explicit methods of searching, critical appraisal and synthesis of information selected”. In addition, several authors argue that this process, when documented, can be repeated and followed by other researchers.

Berwanger et al. (2005), conducted a study on the importance of systematic reviews, presenting a history in which they pointed out that since the seventies, psychologists have drawn attention to the need for systematization of the steps necessary to minimize systematic errors (biases) and random errors in scientific research reviews, in particular in the health area. As reported by the authors, in 1987, Yusuf and Peto pointed out the need for systematic reviews of randomized trials as an approach to obtain clinically significant responses. Furthermore, in 1987, Mulrow addressed the problem of the poor quality of clinical study reviews. In 1988, Oxman and Guyatt published guidelines to assist readers in critical evaluations (systematic reviews), concerning the validity, applicability and results. According to Freitas and Vieira (2010), the systematic review is a subject extensively discussed and used in the health area, appearing in the area of Computer Science in 2004, more specifically in the area of Software Engineering, with the work Procedures for Performing Systematic Reviews, by Kitchenham (2004).

Another contribution to the understanding and construction of systematic reviews is Cochrane Handbook for Systematic Reviews of Interventions, which focuses the health area; however, these instructions could be used in other areas, serving as guide for building more concise
reviews. According to this manual, a systematic review brings together all the empirical evidence, according to predefined criteria, to answer a question, promoting more reliable results. Among others, these features of systematic reviews are highlighted: a set of goals with pre-defined eligibility criteria; an explicit and reproducible methodology; a survey of all studies that meet the criteria adopted, within the investigated issue; the constant evaluation of the risks of biases; and the presentation synthesizing the results of the studies - included in a systematic way (Higgins; Green, 2011).

For Kitchenham (2004) and Centre for Reviews and Dissemination (2009), a systematic review can be conducted in three major steps: planning, conducting and publication of the results. In the planning step, the protocol of the review is defined, i.e., the purpose of the review and the procedures to be adopted for the research are described. In the conducting step, the search for relevant works is made (bibliographic review) - considering the object of study, the selection of publications and the extraction and registration of data from each of the selected sources. And finally, in the publishing step, the considerations about the selected sources and conclusions are described.

Freitas and Vieira (2010) have stated that even if one adopts a systematic review, it is not possible and nor is it sure that all the works about the subject searched will be obtained. According these authors, some works may not have been found due to the number of selected sources, limitations of the repositories related to the quality of its search system resources, difficulties in defining the search syntax, and invariably, the researcher bias when choosing articles. Thus, it appears necessary to improve the process of systematic literature search, improving the conducting of the information retrieval process as a whole.

**Information retrieval**

Information retrieval is an essential tool for research and the evolution of science in every area of knowledge. One of its main issues is to retrieve documents that are relevant to the users. Viewed from a global perspective, information retrieval is about enabling and realizing the meeting of a question with the stored information and providing positive return to the requesting user.

The book *Modern Information Retrieval*, by Baeza-Yates and Ribeiro-Neto (1999), is one of the preeminent works on information retrieval, providing a broad discussion about this subject. For the authors, information retrieval deals with the representation, storage, organization and access to information. They also make a distinction between data retrieval and information retrieval, stating that to retrieve data consists mainly of identifying documents of a particular collection that have as keywords the same terms as those used by the user in a query. Thus, according to those authors, the recovery process does not satisfy the information needs of users, it only retrieves data. For these authors, information retrieval consists of obtaining relevant information about a particular subject, and not simply acquiring documents that statically answer a query requested.

Grossman and Frieder (2004), presented two categories of documents that corresponded to any given query: the relevant documents and retrieved documents (Figure 1). In a perfect system, the two sets would be equivalent (returning only relevant documents), according to the authors; however, in reality, the systems also return many irrelevant documents. Thus, two metrics are used to evaluate systems, precision and recall. The first is the ratio between the number of relevant documents returned and the total number of documents returned. The second is the number of relevant documents returned in the total number of documents believed to be relevant.

![Figure 1](image-url) Categories of documents that correspond to any issued query.

Source: Based on Grossman and Frieder (2004).
For Belkin (1996), the process developed by an information retrieval system is complex and should handle: the representation (users stating the problem); the comparison (the representation of user-informed problem with the content); the interaction (user with someone - mediator - or machine - human-computer interaction); the judgment (user analysis and relevance); and the modification (change of the search depending on the results and/or reformulation of the question).

Hoeschl (2006) emphasizes the historical aspects of search engines. The first generation based on directories such as Yahoo; the second based on robots and automated technologies such as Altavista; the third generation brought the metasearchers; the fourth led to refinement in the organization of the results (All the Web); the fifth marked by Google, in its way, combined sophistication and a wide coverage, brought its Page Rank algorithm to the Web; the sixth, under development, is likely to be the merging of the various types of media into a single search - text, images, videos and sound; the seventh generation, according to Hoeschl (2006), is marked by quality in the selection of information, through intelligent content analysis, based on ontologies and artificial intelligence techniques. The author reports that many advanced technologies are at present seeking more qualified information retrieval, such as the use of Structured Contextual Search, Dynamically Contextualized Knowledge Representation, Text Mining and Case Based Reasoning (RBC).

With new information retrieval techniques emerging every day, one can envision solutions for recovering documents and scientific information more effectively, as well as optimization in conducting systematic processes of literature review (specifically at the systematic search stage). With the use of ontological resources - such as reference ontologies - supporting and connecting the information retrieval services, promoting interoperability between them and facilitating the processing, retrieval and dissemination of information - it is possible to use more intelligent search agents for conducting systematic and replicable surveys, optimizing time and avoiding rework, as well reducing research biases, promoting a greater credibility of the results.

In this interim period, an example is presented of reference ontology for digital scientific journals and its application is demonstrated in supporting systematic literature review processes based on smart information retrieval techniques.

**Reference ontology**

Many resources have been applied by numerous public and private institutions since the ease of access to information and its vast availability on the Web, creating a profusion of resource types, standards and guidelines, hindering the efficient retrieval of relevant information.

The credibility assigned to pages, sites, repositories, portals, and even databases is questionable, pushing people, research groups, and especially institutions and research centers to study ways to provide effective recovery on the Web, as well as to unify and integrate these resources - enabling a greater interoperability - in order to enable a wide and accurate recovery from the databases of the different knowledge areas. In this sense, ontologies have recently been broadly applied.

Fonseca (2007) mentions that ontology can be defined as a tool that helps to describe a specific world and that information systems will only be as good as the associated ontologies. Guarino (2008) collaborates in this context, when he states that reference ontology addresses the aspects of scientific theory, as it seeks to improve the adequacy of the representation of the object under study. According to the author, reference ontology can be renamed as foundational ontology, referring to the ontological theories whose focus is to define or clarify the intended meanings of terms used in very specific areas. Corroborating this context, in the present study, the choice was to use an ontology that “describes a specific world”, i.e., describing the conceptual structure of digital scientific journals, standardizing and consolidating its metadata, to support its validation as a vehicle for disseminating knowledge, promoting information exchange and dissemination of knowledge.

Thus, the ontology of digital scientific journals to be used, proposed by Fachin (2011), is characterized as a reference ontology, as it consolidates a set of metadata that standardizes the basic elements related to digital scientific journals, providing understanding and sharing in this area - where there is interaction between people and systems. As described in Fachin et al. (2010) and Fachin...
(2011), a wide and deep research has been conducted, consolidating these metadata and developing the ontology based on Fernández-López et al. (1997) and Noy and McGuiness (2005) methodologies. The tools used were the ontoKEM - Ontology for Knowledge Engineering and Management (http://ontokem.egc.ufsc.br) - and the Protégé. The Figure 2, created with the Jambalaya plugin for Protégé, shows some classes of this ontology, with its relations of hierarchy. Each concept used and their relationships, as well as other concepts that have been suppressed in this figure, are presented in detail by Fachin et al. (2010) and Fachin (2011).

In this ontology, the classes and subclasses were created following a hierarchical structure, starting from the broader knowledge - Digital Scientific Journal - to the more specific, trying to group the classes semantically, and thus, showing a structure of all the necessary information for the creation, maintenance and management of digital journals. The following scenarios, shown below, demonstrate the applicability of this ontology in information retrieval contexts present in systematic processes of literature review.

**Ontology applicability**

Considering the reference ontology of digital scientific journals included in scientific work information retrieval contexts, and supporting an improvement in processes of systematic review, there is the possibility of two main approaches to the operationalization of this scenario: on one hand with a “batch information processing”, and, on the other hand, with an ‘on-the-fly information processing’. In the first approach, (Figure 3), the searches would be performed in a centralized database of information previously loaded, following the concepts of the reference ontology. In this context, there are two possibilities for the database loading: providing a service for journals and/or indexers to submit their information (passive mode) or using a service to perform collection, integration and storage of this information (active mode).

In the first case, the journals and indexers would have to adapt themselves to the submission service available (which would follow the concepts of the reference ontology), creating other services for automatic
upload information - in the case of larger repositories - or even making manual entries - in the case of smaller repositories. In the second case, a service would be developed to support the collection, integration and storage of information from different repositories. As journals and indexers do not follow a standard for the publication of information, there is a need to create other services specialized in collecting and processing the data from each repository - dealing with the issues of semantic and syntactic differences. The two possibilities for loading the centralized database (passive mode and active mode) - used in the approach with batch processing of information (Figure 3).

With this approach, involving the use of a centralized repository of previously loaded and integrated information, faster information retrieval is enabled, as well as a greater level of control over the availability and integrity of this information. However, the main disadvantage is the loss of dynamism in adding new content or even in changes or updates of previously published content.

The alternative to this scenario is the "on-the-fly" approach. In this second approach, (Figure 4), the search would be performed in real time and results would be processed dynamically with the aim of integrating and determining the relevance of information, following the concepts of the reference ontology. In this context, two possibilities for operationalization of this service are envisioned: the standardization of services of the different repositories - so that queries would be performed in a uniform way (following the concepts of the reference ontology); or the use of non-standard services - so that specialized operations for querying and processing the information on each service would be required.

In the first case, the repositories that have interest in providing their information would have to create a standardized service to perform queries, according to the system protocols and based on the reference ontology, and the system would simply orchestrate the queries among the services, as well as the information retrieval and integration. In the second case, specialized services would be developed for collecting and transforming the information available in the (non-standard) services of the different repositories - tailoring this information to the concepts available in the reference ontology.

After this stage of "pre-processing", the information could then be integrated and made available by the system. The two possibilities for using the services from the different repositories (using standard or non-standard services) - in the approach with "on-the-fly information processing".

As previously mentioned, by means of this approach, without involving the use of a centralized repository of information (previously loaded and integrated), the searches would be performed in real time and results would be processed dynamically, for integration and relevance determination of the information, following the concepts of the reference ontology - thus obtaining the most up-to-date information available from the different providers. However, there would be a loss in the information recovery speed, as well as less control over the availability and integrity.

Moreover, by consolidating a set of metadata related to digital scientific journals, the reference ontology could provide a standardized representation, availability and access to information in the field of scientific publications, enabling greater interoperability between the different systems and repositories and the development of more accurate and inclusive information retrieval services.

Consequently, improving the mechanisms for systematic search and directly assisting in the process of systematic review of literature (both at the conducting and publishing stages), researchers would not need to deal with different search techniques and tools available in the databases.

In addition, it would be possible to create even more intelligent search agents that would recover the most relevant studies according to different inclusion and exclusion criteria (pre-defined by the researchers), following the ontology concepts and making inferences. These agents could incorporate explicit and formal methods of search, extraction, grouping and summarization of the studies.

Supporting the researchers, even at the stage of critical evaluation of the studies, it would also be possible to generate a "trail" of all procedures performed - making it possible to replicate and update (running over more
Figure 3. Approach for realization of meta search using a centralized information repository - “batch information processing”.
Source: By authors.
Figure 4. Approach for performing the metasearch "on-the-fly", without a centralized information repository.

Source: By authors.
recent works) the process automatically, as well as validate the methods performed by the agent. Thus, benefits would also be obtained at the stage of publication of procedures used in the systematic review, without taking into account a possible increase in the credibility of bibliographic reviews - because the automation of some steps might reduce the researcher bias, which may occur in the selection and evaluation of literature, and with the manual use of different tools.

Conclusion

This article presented possible approaches to the use of a reference ontology for digital scientific journals to enhance information retrieval, supporting the development of systematic search in processes of systematic literature review.

The ontology intends to support the standardization of concepts related to scientific journals, which would ensure greater interoperability between different data repositories and a more accurate and comprehensive information retrieval, in addition to smart inferences in the systematic review steps.

The proposed ontology could also support intelligent search agents, which would recover the most relevant studies according to different inclusion and exclusion criteria, based on the systematic review goals. These agents could also generate a “trail” of all procedures performed - enabling automatic replication and update of the process, as well as the validation of the process conducted by the agent. The aim is also to benefit the stage of publication of the systematic review processes, and make it possible to increase the credibility of bibliographic reviews - due to the automation of some steps that might reduce researcher bias, which may occur in the selection and evaluation of literature, and in the application of different tools.

The approach discussed in this study is now under evaluation to support scenarios of information integration and inference as well, with regard to the evaluation of courses and journals, based on the Brazilian Graduate Program Evaluation - Coleta Capes.

References


