Qualitative and Quantitative Methods in Libraries (QQML) 1:331 –341, 2014

Faceted taxonomy for accessing digital libraries

Maculan, Benildes C M S.[1] and Lima, Gercina A. B. O.[2]

[1]Universidade Federal de Minas Gerais, UFMG
[2]Universidade Federal de Minas Gerais, UFMG

Abstract: We present the development of a faceted taxonomy of scientific works and the use of this taxonomy in a system for searching a digital library of theses and dissertations (DLTD). We also propose an algorithm for indexing these documents. Our algorithm eases the retrieval of information contained in this kind of document, and offers structured information in which results have already been refined for the user. Our methodology is based on: a) domain analysis theory; b) facet analysis theory; and c) content analysis methods, including the categorical thematic analysis technique of Bardin (2009) and the methodology proposed by Moraes (1999). The faceted taxonomy we developed is the simplified representation of the thematic content of the documents.

Keywords: Faceted taxonomy. Facet analysis theory. Domain analysis theory. Content analysis. Categorical thematic analysis technique. Scientific communication. Information retrieval.

[1] This paper presentation at the event Roma, Italy, QQML 2013, had Fundação de Amparo à Pesquisa do Estado de Minas Gerais, FAPEMIG, as a principal sponsor.

1. Introduction

One of the main purposes of information retrieval systems (IRSs) is to promote and ease sharing and transferring of information. Some IRSs are developed to offer specialized services to a specific community of users. Among these services, we focus on the digital libraries of theses and dissertations (DLTD), repositories which make possible scientific communication.

According to Targino (2000, p. 54), “scientific communication [...] provides the products [...] and the producers [...] with the necessary visibility and possible credibility in the social environment to which product and producers belong”. We

Received: 28.5.2013 / Accepted: 29.3.2014 ISSN 2241-1925

understand that the scientific communication, as it indicates the credibility or refutation of a piece of knowledge developed in any given discursive community, promotes discussion among peers and strengthens the domain. As a result, it gives visibility to the domain.

In this context, the DLTDs have an important role in scientific communication, because they make available the results and advances of research studies already concluded, which makes possible the progress of the research area. For instance, it becomes possible to retrieve valuable and detailed information about theories and applied methods.

However, the DLTDs contain specific problems that make it difficult to access to the detailed results of research studies. Some of these problems are linked to the production of the document by the author, such as (a) titles that are not representative of the document's content, (b) insufficient keywords listed by the author, and (c) incomplete summaries. In addition, there are problems connected to the data input into the IRS, caused by inadequate indexing of this sort of document. Thus, it becomes difficult to retrieve the informational content of the documents. Regarding the problems associated with the production of the document by the author, the solution involves the academic community, by raising the requirements for publishing academic works. The problems related to the information treatment in the data input in the IRS must be resolved by DLTD managers.

Within an IRS, the indexing subsystem is responsible for the process of converting information content (of the library's documents) into meaningful representations for data output. These representations will be retrieved by the user in the search for information. When theses and dissertations are available in DLTDs, it is necessary to apply an appropriate informational treatment in order to make it easier for the users to get answers for their queries.

We proposed a faceted browsing taxonomy as search interface for a DLTD (Maculan, 2011). The goal was to offer structured information with results already refined to the final user.

A faceted taxonomy allows for the presentation of the segmented domain in facets. In each facet there is a hierarchy of elements that share similar characteristics. The different possible relationships among the facets of the taxonomic structure indicate the multi-dimension al aspect of the knowledge in this domain. The navigational attribute of the faceted taxonomy enables users to find information by means of browsing its structure. This is possible because each facet and subfacet has its respective information contents attached to it.

All these features ease the search and retrieval of information by the user. This happens because the users can continue refining their search from the options which become available in the process. In addition, we eliminate the possibility of "empty results", because only the facets in which there are information contents are presented to the user.

We will present the methods and procedures we adopted, as well as the results and conclusions we obtained in this research. First, we applied domain analysis theory to identify the domain and the user of the specialized service offered by a DLTD. We used this same theory to determine in what types of information this user was interested. Then, we applied facet analysis theory, with the principles developed by Ranganathan, in order to form the set of basic thematic classes of the faceted taxonomy. We validated this set by using the content analysis method based on Bardin (2009) and applying the categorical thematic analysis technique proposed by Moraes (1999).

2. The domain analysis theory

Birger Hjorland was the pioneer in the formulation of domain analysis theory. But in the early 70s, Jesse H. Shera had already thought about information as originated from and dependent on the cultural and social context. The domain analysis is characterized as an approach that combines theory and practice. It presents a functional method to understand the implicit and explicit functions of information in the context of communication within a discursive community. As defined by Jacob and Shaw (1998), a discursive community is a domain or social group in which its subjects are synchronized in thought, language, and knowledge, where the individuals function as a social construction, not as autonomous entities.

According to Hjorland and Albrechtsen (1995), the most effective way to perceive information is establishing the knowledge domain one wants to know, and this must be done from the domain analysis. This perspective sees the knowledge process as a product that is historically, culturally, and socially developed, and, for that reason, focused on the knowledge domains (HJORLAND, 2002a). Hjorland (2002b) states that individuals do not process information according to a system of individualistic rules, but according to criteria grounded on the social and cultural environment of this subject. Thus, concepts are generated within discourse communities, within each domain or social group.

We understand that to create an information service, through a digital library of theses and dissertations, it is necessary to identify the domain (discursive community), the user, and the user's information needs. Determining these elements requires answers to questions related to "who produced", "in what social context was it produced ", "what is the purpose of the production", and "who uses the product".

In order to answer these questions, we applied domain analysis theory in four phases in a qualitative approach. The procedures we used were adapted from the methodology of the domain analysis which refers to the software engineering area:

1. Identification of the domain: we analyzed the domain using as guide the following questions: what is the domain? is the domain known? what kind of information does this domain produce?
2. Data analysis: we have identified the main characteristics of the domain, establishing relationships and functions among the pieces of information, based on the following questions: who produces the information identified in the previous phase? what is this information produced for? who is interested in the use of the information produced?
3. Domain modeling: we selected the types of information that are considered important to be available in the information service to be created, according to the criteria proposed in the second phase (data analysis).

4. Collection and selection of data and validation: the types of information selected in the third phase were validated and dated through mapping of the literature and informal interviews with specialists of the area under study.

The final result of the analysis of the data collected in the literature, based on the domain analysis theory, enabled us to define the domain as the "scientific community". We found that this domain consists of researchers and scientists who investigate different themes within the same theoretical-methodological framework. These scientists develop and produce theses and dissertations, and they are the most interested in this kind of document.

This result is supported by Kuhn (2006), who characterizes the members of a scientific community as having similar scientific interests, similar bibliography, a particular vocabulary (even it is divergent at some points), and similar models and methodological approaches. Also, they usually start their research studies from results of research studies already published. Kuhn states that a community is strengthened as the knowledge produced can be transferred and used in new investigations within it, intensifying the use of its own theories, methods, and techniques.

We surveyed the literature of the area to identify what sort of information would serve the needs of this social group when performing their professional activities. The decision of analyzing the literature instead of applying a users' study was based on Hjorland (2002b). He states that, in general, users are unable to express their needs and relevant criteria when "browsing". According to Stevenson and Byerly's (1995), scientists explore a specific object in order to understand and amplify their scientific knowledge about this object or to improve the existing theories. We believe that the primary activity of researchers is to produce and to use information within their own social group (their peers). They play the role of producers and communicators, sharing information relative to, mainly, works which have been concluded already.

Scientists need distinct types of information in each different phase of their research efforts. According to Alvarado and Oliveira (2008), during an investigation, the researcher needs three basic types of information: 1) themes already researched; 2) methods and techniques already used in research studies about the theme; 3) results obtained in research studies already concluded.

Guided by these three types of information and by the facet analysis theory, we created the set of basic thematic classes that composed the structure of the faceted browsing taxonomy.

3. Composition of the set of basic thematic classes

To implement the specialized service of a DLTD, it is necessary to have criteria for indexing and for information retrieval, both created from paradigms and from methodological and epistemological norms. The scientific knowledge of any domain is complex, and it doesn't have any rigidly set boundaries. Thus, we decided to use the faceted taxonomy, which is a chain of relations in which each facet, within any given basic thematic class, may be linked to other facets of different basic thematic class.

This aspect gives the meaning of multidimensional relationships chain, which is present among the principles of the facet analysis theory. These principles emerge from two studies: 1) Ranganathan's, whose theory has 46 canons, 13 postulates and 22 principles, and is present in five works: Five Laws of Library Science (1931), Colon Classification (1933), Prolegomena to Library Classification (1937), Philosophy of Book Classification (1951), and Elements of Library Classification (1962); and 2) the studies of the members of the Classification Research Group (CRG), created in the 50s, in the United Kingdom, which included: D. J. Campell, E. J. Coates, Derek Austin, J. E. L. Farradane, D. J. Foskett, G. Jones, J. Mills, T. S. Morgan, B. I. Palmer, O. W. Pendleton, L. G. M. Roberts, B. C. Vickery, Robert Fairthorne, Barbara Kyle, A.J. Walford, K. E. Watkins, Derek Langridge, and Jack Wells.

To accommodate all the knowledge of a domain, Ranganathan conceived five fundamental categories, which are called PMEST: Personality, Material, Energy, Space and Time. On the other hand, CRG suggested the following categories: Type of final product, Parts, Materials, Property, Process, Operation, Agent, Space, Time, and Presentation format (LIMA, 2004b). CHART 1 shows a qualitative analysis of the two conceptions:

CHART 1 – Comparison between Ranganathan categories and CRG

RANGANATHAN	**CRG**
Personality	Types of final product
Material	Parts Materials Properties
Energy	Processes Operations Agents
Space	Space
Time	Time

Source: Maculan (2011)

The categories of CRG can be mapped to those of Ranganathan. It is important to point out that Ranganathan never affirmed that a domain cannot be decomposed into more than these five fundamental categories. However, based on his own experience in libraries, he suggested that this set of five fundamental categories is enough to categorize any domain.

We used Ranganathan's principles of facet analysis theory, applying the PMEST to analyze the domain and determine the set of basic thematic classes that shaped the faceted taxonomy structure. We applied this method in three phases: 1) selection of documents to be analyzed; 2) reading of documents to identify their constituent parts, through dissection, which is the segmentation of the *corpora* in their constitutive parts, represented by terms of the same level; 3) selection of the most relevant constitutive parts to make up the set of basic thematic classes, followed by the determination of their arrangement. These procedures aimed to search for the essential and common elements in the production of this kind of document. CHART 2 shows the results we found:

CHART 1 – The set of basic thematic classes

RANGANATHAN	DOMAIN: Basic Classes
Personality	Theme Historical/contextual foundation Theoretical foundation Product (final result)
Material	Data collection Type of research Object
Energy	Methods
Space	Environment
Time	Not a class (Year of publication)

Source: Maculan (2011)

Having reached this result, we started to apply the categorical thematic analysis technique to validate the nine basic thematic classes.

3.1 Validation of the set of basic thematic classes

The set of basic thematic classes was validated through the content analysis method, with the technique of the categorical thematic analysis, which is based on Bardin (2009) and was applied with the methodological procedures proposed by Moraes (1999): 1) preparation of information, with the analysis of documents; 2) unitarization, which is the transformation of the content into record units (RU) and content units (CU), with the codification and acquisition of CUs and RUs from the textual structure of the documents, according to the literature or methodology of research and its specific rules; 3) categorization that is classification of the units into categories; 4) description, with the definition of each category, and continuous validation of the category set; 5) interpretation, with an analysis of the achieved result.

At the end of the validation, we found that all scientific research is based on theoretical fundamentals of the investigated theme, and the researchers need to know the "state of the art" of the subject, as well as the methods and techniques that were already applied to the research problem. Moreover, we found that the set of basic thematic classes is the "simplified representation" of all the structural and textual content of thesis and dissertation documents. We also found that the logical sequence of the ideas presented in the documents was respected, following the sequence of its chapters and descriptive data.

Once we finished this procedure, it was necessary to develop an algorithm in which the criteria for indexing documents of theses and dissertations were determined.

4. Criteria for indexing documents

The algorithm we created for indexing documents is a sequence of tasks, with clear and well defined instructions. It was developed as an indexing matrix in order to guide the extraction of concepts from documents. These concepts will feed the set of basic thematic classes, which make up the faceted taxonomy structure.

The matrix has three columns: 1) the first column is the set of basic thematic classes; 2) the second column contains the questions, which were elaborated according to the NBR 12676:1992 guidelines and the principles of indexing system analysis PRECIS (FUJITA; RUBI, 2006); 3) the third column contains the parts of the structural analysis of theses and dissertations, in which we can find the answers to the questions of column 2. Part of the matrix is shown in CHART 3.

CHART 3 – Matrix for indexing theses and dissertations

BTC	**QUESTIONINGS (RULE 12.676) and PRECIS**	**PART OF THE TEXTUAL STRUCTURE**
C1. THEME	What subject is this document about?	SUMMARY/ INTRODUCTIO
C2. EMPIRICAL OBJECT	What is the empirical object of the study in question? What object was used and/or analyzed in the research?	SUMMARY/ INTRODUCTIO N

Source: Maculan (2011).

Following the completion of the Matrix we applied it to index the documents.

4.1. Indexing theses and dissertations

Indexing documents is done in two basic phases: subject analysis and translation. In the first phase, there is the identification and selection of concepts that represent the thematic content of documents. In the second phase, the translation is carried out either through an indexing language or a controlled vocabulary. The use of a tool of controlled vocabulary for the translation phase is important in order to avoid inconsistencies and ambiguities, such as the use of more than one indexing term to represent a similar concept.

We suggest that this procedure be applied in the data input in the digital library of theses and dissertations, aiming to improve the information retrieval. The first indexing phase must be guided by the Matrix (CHART 3), and the translation must be done by a specific controlled vocabulary for the domain of documents.

After indexing the documents, all the indexing terms extracted will feed the set of basic thematic classes. The set of terms will constitute all the taxonomy structure: facets and subfacets. Each of these elements can be used during the faceted browsing.

5. The faceted browsing

In a digital environment, an IRS is considered faceted when it contains a minimal number of operations and has a set of elements that the user can browse. The faceted browsing "enables the user to make up a question progressively, and to keep on observing the effect that his choice of facets has on other facets" (TUNKELANG, 2009, p.23). The faceted browsing has already been used in some projects of digital libraries, as reviewed by Maculan (2011). In Brazil, the FLAMENCO system stands out as one of the first and most cited examples of faceted browsing.

Tunkelang (2009) states that the faceted browsing presents advantages such as (1) user guidance, (2) progressive formulation of search questions, and (3) exploration and retrieval of information through the faceted structure. He also states that the faceted browsing approach is indicated for semi-structured texts, such as the contents of theses and dissertations. A semi-structured text contains structured elements such as author, date, title of work, and title of sections, and elements without structure such as summary and content.

The browsing and access mechanism – the browsing faceted taxonomy (TAFNAVEGA) – presented in this article and proposed by Maculan (2011), worked with semi-structured documents like theses and dissertations.

In this proposal, the final product enabled the organization of the information available in the digital library, thus improving it. By offering the informational resources in a structured and orderly way, our prototype was able to make it easier for the user to search and retrieve information.

In this project, the search and faceted browsing made it possible to organize large data collections, helping the user not to feel displaced while navigating the search system. In addition, the faceted browsing can be combined with free search, which further improves the user experience. This tool was applied in a study case, whose results can be found in Maculan (2011). Below we present a synthesis of the application.

5.1 The case study

The research environment was the digital library of theses and dissertations (DLTD) of the Federal University of Minas Gerais (UFMG). The investigation's universe was the set of documents analyzed in the Post Graduation Program in Information Science (PPGCI), School of Information Science (ECI). The empirical object were the theses and dissertations which came from the PPGCIs Organization and Use of Information line of research, between 1998 and 2009, and were made available in the DLTD database in July 2010. We obtained a total of 290 papers: 62 theses and 228 dissertations.

Of these 290 documents, we focused on the analysis of the *corpus* formed by theses and dissertations presented in the PPGCI/ECI/UFMG, whose line of research was the Organization and Use of Information (OUI). The selected works amounted to 41 documents – 12 theses and 29 dissertations.

6. Conclusions

Strategies for faceted searching and browsing have been employed in the area of information retrieval for at least two decades. Many of these studies were based on the principles proposed by Ranganathan. Two works stand out as pioneers in this kind of methodology: (1) the work of Ahlberg and Shneiderman (1994), from the University of Maryland; (2) the work of Pollitt et al. (1996), from University of Huddersfield. Ahlberg and Shneiderman (1994) created the *FilmFinder* prototype, which allows the exploration of a database of movies by facets. These facets are determined from a set of pre-established parameters called "parametric search". The deficiency of this product is that the user may get an empty response to the question formulated, because the system doesn't limit responses to the existing informational resource in the database. Pollitt et al. (1996) created the prototype HIBROWSE (HIgh resolution interface for BROWsing and SEarching) which is a faceted navigation project using a "view-based searching system". This mechanism was created for a bibliographic database of medical studies, the EMBASE, and for the digital library of the European Parliament, the EPOQUE. These systems were structured using a faceted thesaurus in the medical field as knowledge base.

In addition to these works, strategies for faceted searching and browsing have also been employed in the creation of other prototypes, such as: (1) the *Relation Browser* project or *RAVE*, started in 1998-1999, used for mapping statistical information available in the database, providing a mechanism for searching and browsing; (2) the *FACET* Project , which used faceted thesauri in order to improve the retrieval of information from semantic approximation measures, through existing relationships in the thesaurus semantic structure; (3) the *MuseumFinland* project based on ontologies and on Dublin Core standards; (4) the *Suomi.fi* which was created in categories using an ontology as knowledge base; this prototype was developed with the "Yahoo! Approach", Semantic Web principles, Resource Description Framework (RDF) standards and *Web Ontology Language* (OWL); besides the faceted navigation, the site offers a search by keywords, allowing the combination of both; (5) the *TerveSuomi.fi* has a multifaceted search interface, in which the system gives the user an overview of the portal contents, and the user can combine one or more values (facets) in the navigation process; (6) the *HealthFinland* has a faceted search interface, and currently has 21 categories of content; (7) the *POSEDU* (Educational Semantic Portal), a search and faceted navigation system that has structural facets built by formal rules from an ontological knowledge base; (8) the *Explorator*, a mechanism that allows navigation and search, built on *Resource Description Framework* (RDF) language base.

Unlike previous works, we present a faceted taxonomy as search interface to aid information retrieval. We propose a specific conceptual model for digital libraries of theses and dissertations, which follows the structure of production of these types of documents. The faceted taxonomy, called TAFNAVEGA, gives options for the user to retrieve information by elements, often implicit, such as theories, methods, data collection instruments, and results of research already completed. Usually, this kind of information is not described by metadata in digital libraries of theses and dissertations, hindering their recovery. With TAFNAVEGA, the researchers expand their ability to retrieve this valuable information in different ways, such as: providing broad vision of the entire contents of the repository, guiding the user in formulating searches, and never showing an empty search result. Thus, the user can view, explore, and get the information they need for starting a new research effort or making progress on an ongoing one.

References

Alvarado, Rubén U.; Oliveira, Marlene (2008). A comunidade científica da Biblioteconomia e Ciência da Informação brasileira. *Pesquisa Brasileira em Ciência da Informação e Biblioteconomia,* v. 3, n. 2.

Associação Brasileira de Normas Técnicas (1992). *NBR 12676. Métodos para análise de documentos*: determinação de seus assuntos e seleção de termos de indexação. Rio de Janeiro: ABNT.

Bardin, Laurence (2009). *Análise de conteúdo*. 4. ed. rev. e atual. Lisboa: Edições 70.

Fujita, Mariângela S. L.; Rubi, Milena P. (2006). Um modelo de leitura documentária para a indexação de artigos científicos: princípios de elaboração e uso para a formação de indexadores. *DataGramaZero – Rev Ci. Inf.,* Rio de Janeiro, v. 7, n. 3, jun.

Hjorland, B. (2002a). Domain analysis in information science: eleven approaches: traditional as well as innovative. *Journal of Documentation,* v. 58, n. 4, p. 422-462.

Hjorland, B. (2002b). Epistemology and the socio-cognitive perspective in information science. J*ournal of the American Society for Information Science and Technology - JASIST*, v. 53, n. 4, p. 257-270.

Hjorland, B.; Albrechtsen, H. (1995). Toward a new horizon in information science: domain analysis. *Journal of the American Society for Information Science - JASIS*, v. 46, n. 6, p. 400-425.

Jacob, Elin K.; Shaw, Debora (1998). Sociocognitive perspectives on representation. *Annual Review of Information Science & Technology*, v.33, p.131-185.

Kuhn, Thomas S. (2006). *A tensão essencial.* 9. ed. Lisboa: Edições 70.

Lima, Gercina Ângela B. de O. (2004a). *Mapa hipertextual (MHTX) um modelo para a organização hipertextual de documentos.* 2004. 199f. Tese (Doutorado em Ciência da Informação) – Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte.

Lima, Gercina Ângela B. de O. (2004b). O modelo simplificado para análise facetada de Spiteri a partir de Ranganathan e do Classification Research Group (CRG). *Inf. Cult. Soc.*, Ciudad Autónoma de Buenos Aires, n. 11, jul./dic.

Maculan, Benildes C. M. dos S. (2011). *Taxonomia facetada navegacional*: construção a partir de uma matriz categorial para trabalhos acadêmicos. 2011. 185f. Mestrado em Ciência da Informação, UFMG, Escola de Ciência da Informação, Belo Horizonte.

Moraes, Roque (1999). Análise de Conteúdo. *Revista Educação*, Porto Alegre, v. 22, n. 37, p. 7-32.

Stevenson, Leslie; Byerly, Henry (1995). *The many faces of science*: and introduction to scientists, values, and society. Boulder: Westview.

Targino, M. G. (2000). Comunicação científica: uma revisão de seus elementos básicos. *Revista Informação & Sociedade: Estudos*, João Pessoa, v. 10, n. 2.

Tunkelang, Daniel (2009). *Faceted search*. North Carolina: Morgan e Claypool.