# Towards an Ontology of ElPub/SciX: A Proposal

*Sely M S Costa [1]; Claudio Gottschalg-Duque [2]*

[1] University of Brasília, Department of Information Science
Campus Universitário Darcy Ribeiro, Brasília, DF, Brazil
e-mail: selmar@unb.br
[2] University of Brasília, Department of Information Science
Campus Universitário Darcy Ribeiro, Brasília, DF, Brazil
e-mail: klauss@unb.br

## Abstract

A proposal is presented for a standard ontology language defined as ElPub/SciX Ontology, based on the content of a web digital library of conference proceedings. This content, i.e., ElPub/SciX documents, aims to provide access to papers presented at the total editions of the International Conference in Electronic Publishing (ElPub). After completing its 10th years in 2006, ElPub/SciX is now a comprehensive repository with over 400 papers. Previous work has been used as a basis to build up the ontology described here. It has been presented at Elpub2004 and it dealt with an Information Retrieval System using Computational Linguistics (SiRILiCo). ElPub/SciX ontology constitutes a lightweight ontology (classes and just some instances) and is the result of two basic procedures. The first one is a syntactic analysis carried out through the Syntactic Parser-VISL. This free tool, based on lingsoft's ENGCG parser, is made available through the Visual Interactive Syntactic Learning, a research and development project at the University of Southern Denmark, Institute of Language and Communication (ISK). The second one, carried out after that, is a semantic analysis (concept extraction) conducted through GeraOnto, an acronym that stands for "generating an ontology", which extracts the concepts needed in order to build up the ontology. The program has been developed by Gottschalg-Duque, in 2005, in Brazil. The ensuing ontology is then edited via Protégé, a free, open source ontology editor. The motivation to carry out the work reported here came from problems faced during the preparation of a paper to Elpub2006, which aimed to present data about a number of aspects regarding the ElPub/SciX collection. While searching the collection, problems with the lack of standardization of authors and institutions names and the non-existence of any control of keywords had been identified. Such problems seem to be related to an apparent absence of "paper preparation" before entering into the SciX database. Lack of preparation, in turn, has brought about the desire of finding a solution, which is expected to support the work of those interested in searching the collection to retrieve information. ElPub/SciX ontology, therefore, is seen as that helping solution to support ElPub information retrieval.

**Keywords:** ontology; Elpub conferences; information retrieval

## 1 Introduction

Electronic publishing constitutes a hot topic of discussion within the academic environment, particularly in the study of scholarly communication. Such interest is due to the opportunities provided by the web and the Internet for a document to be available world wide in electronic format. As a free, democratic environment, the Internet provides a huge amount of information, which, on the other hand, presents a challenge for those who seek relevant hits. The digital content available today actually represents a great chaos to those interested in finding relevant information for research or in scrutinising a document collection for the same purpose.

There have been a variety of approaches to help with this matter, such as those that made possible to develop thesauri, ontologies, taxonomies, topic maps and other resources. They have been developed with the aim of facilitating the intellectual work. Therefore, a well-organised collection should be supported by one of them. In this context, an ontology can be considered one of the richest resources used for the automatic treatment of electronic documents, since it constitutes a set of definitions of a *formal* vocabulary.

In this paper, we present a proposal for a standard Ontology language defined as SciX ontology, based on the content of SciX, a digital library of conference proceedings. SciX can be viewed as a response to the need of organising and making available a collection of papers presented in an annual international conference. It is

actually a web digital library that provides access to papers presented at the International Conference in Electronic Publishing (ElPub). The conference completed its 10th years in 2006, and SciX is now a comprehensive repository with over 400 papers. The collection comprises papers presented at the two traditional tracks of sessions, namely, general and technical, as well as abstracts of keynote speeches, workshops, round tables and special sessions presentations as well as other kinds of contribution.

## 2 Motivation and Expectations

During ElPub2006, the 10th version of the conference, a short paper was published with observations on a few quantitative data [1]. The analysis of papers from the 10 years conferences showed that SciX content does needs a standardisation process of its data in order to improve search and retrieval for research purposes.

Since the work of Paul Otlet and Henri La Fontaine, regarding documentation, retrieval of relevant content of a document is deemed the key factor of success in any information service/product. In this regard, ontologies have the capability of significantly improving retrieval needed in information services. The proposed ontology will certainly help the exploration of SciX content in both quantitative and qualitative approaches, in the extent that ontologies constitute a set of *classes* (for example, author, title, key-words), *individuals* (for example, Leslie Chan, University of Toronto) and *properties* (http://www.utsc.utoronto.ca) that allows a sounder work on the data available.

It is interesting to note that a vocabulary ontology expressed in a formal specification, such as the Web Ontology Language (OWL), makes possible machine processing of information (in a very basic level), rather than simple data, adding expediency to web content search and retrieval. Based on this understanding, the authors of this paper have decided to develop an ontology to help the work of researchers or practitioners interested in using SciX data for research. The leading objective is to provide an "information resource", allowing the generation of a richer domain-specific knowledge, which is a formal specification of a controlled vocabulary.

The work carried out on the ElPub/SciX collection in 2006 has actually allowed the standardisation of authors and institutions names. The output, however, has not been aggregated to that collection until now. Taking into account that the digital collection so far reproduces the information provided by authors themselves, the work on keywords standardization requires a controlled vocabulary, gradually built up while processing SciX collection. Nevertheless, despite more than 10 years, it seems possible to create this control and help future authors to rely on the output for better stating keywords pertaining to their papers.

The ontology makes it possible to define nodes of semantic relationships and make inferences concerning the topics covered by authors. In addition, a number of relationships between concepts are possible to identify, which, in turn, can respond to the need of standardising them.

This paper, therefore, reports the experience of developing a semi-automated process of extracting concepts from an electronic document collection, in order to create an ontology. It is semi-automatic because of the non-automated procedures concerning part of the data extracted from the database. The idea is to develop an optimised, interesting output of a scholarly papers collection, with the aim of making it easier to be handled by researchers. Through its semantic net, ElPub/SciX ontology is intended to provide better conditions for the user to find the information needed more efficiently. These standard metadata, besides other possibilities, can contribute to define a new environment based on ElPub/SciX Ontology concepts.

It is noteworthy to call attention to the fact that authors have always used different ways of informing both their own names and their institutions names wherever, and whenever they publish. Moreover, different authors define the same topic differently. Concerning Elpub, another aspect that deserves attention is related to the conference sessions' title, which do not always represents a 'core topic' to which the content of those session papers converge. This, in turn, makes an accurate content analysis difficult to carry out.

Such ambiguities and inconsistencies are probably related to the way Elpub has been conducted. It is observable that, as the conference progressed, a number of procedures started to be implemented to make it more organised. At least three of these improvements are clearly identified. Firstly, preoccupation with well-formatted/presented papers based on a well-planned template, along with guidelines about what is expected from authors, has helped improve the content entered into SciX. Secondly, requirements of an abstract, keywords and other data, not present in some of the first editions of the conference, seem to have had an impact on such improvement. Thirdly, the definition of the conference sessions' titles, over the last editions (and 1997's!!), well depicts the content of papers presented in those sessions. Nevertheless, for the enhancement of the access and use of this

content, a standardisation is urgently needed. ElPub/SciX ontology, with no doubt, can definitely contribute to that. In spite of still being in an incipient stage as yet, the ontology has the potential of accomplishing the aforementioned purpose.

# 3 Methodology (Theoretical Approach and Methodological Procedures)

In order to carry out the study, a number of procedures have been performed, in accordance to what has been stated in theories that underlie content analysis and retrieval in a web environment. As it has been asserted, *the use of ontology as a formal explicit specification of a shared conceptualisation, can help to solve the problem of inefficiency, overloaded "fake information", ambiguity and chaos* [2]. These authors draw attention to the fact that the use of automatic semantic analysers (despite its earlier own approaches with some incipient, but encourage results) makes possible extract the conceptual structure, describe phrases and use semantic relationships between words and concepts to establish connections between them. This structure, which constitutes a meta-level description, is a representation that brings order to the collection of documents, so it can be understood as an ontology, in the sense defined by Gruber [3] and others. That is, in computer science, the term 'ontology' expresses explicit formal specifications of terms in a domain and the relationships among them.
Notwithstanding the improvement afforded by an ontology, two problems remain, ambiguity and inconsistencies. Names ambiguity have been approached as one of the major problems in retrieving information from a database. As observed by Han et al [4], "*because of name variations, identical names, name misspellings or pseudonyms, two types of ambiguities in research papers and bibliographies can be observed*". They are authoring multiple name labels and multiple authors sharing the same name label. The same authors point out that "*it may affect the quality of scientific data gathering, can decrease the performance of information retrieval and web search, and even may cause incorrect identification of and credit attribution to authors*".

The solution, that is, name disambiguation, has been approached in a variety of ways and has always been related to the creation of authority files. Auld, cited by French et al [5], stressed that this sort of strategy has been called 'authority work' and have mostly benefited from computational procedures.

It is interesting to emphasise that name ambiguity can be related to a number of entities, such as authors, institutions, journal or conference titles and so forth. Ambiguities in institutions names have been approached by French et al, who looked at techniques to aid in detecting variant forms of strings in bibliographic databases. They highlight Taylor's approach [6], whose first principle of authority control is concerned with all variants of a name being "brought together under a single form so that once users find that form, they will be confident that they have located everything relating to the name". This 'single form' has been defined as 'canonical name' and, in the work of French and his colleagues, consisted of deciding on a set of canonical affiliation strings and then, assigning each affiliation string in the database to one of these canonical strings.

As can be inferred, disambiguation of names is crucial to the work proposed here, as the simple creation of the ontology itself could not solve this sort of problem by itself. It has been partially and preliminary performed by Costa et al [7] in order to carry their analysis out. Nevertheless, it has been a non-automated process in the sense that no computational procedure was used.

The result, however, corresponds to Taylor's first principle of authority control and is used in this second work upon ElPub/SciX collection. That is why it is still a 'semi-automatic' extraction process. Further work will develop an automated procedure for the creation of canonical names of both authors and institutions.

As regard the ontology and the procedures developed in order to create it, the work was based on the previous model developed by Gottschalg-Duque [8], adapted for this application (Figure 1), as it does not yet include the indexing module. As can be observed, the whole process consists of the stages *file conversion, natural language processing and ontology creation and editing*. The implementation of the modules and sub-modules (figure 2 shows a detailed view of the natural language processing module, comprised of two sub-modules, syntactic and semantic) is done by means of three programs, which are Syntactic Parser, GeraOnto and Protegé.

The stages involved in the analyses and in the ontology creation are succinctly described further, and show how each of them is performed, along with the indication of the software used. It is important to highlight that GeraOnto, because of patent problems, does not allow giving any detail.

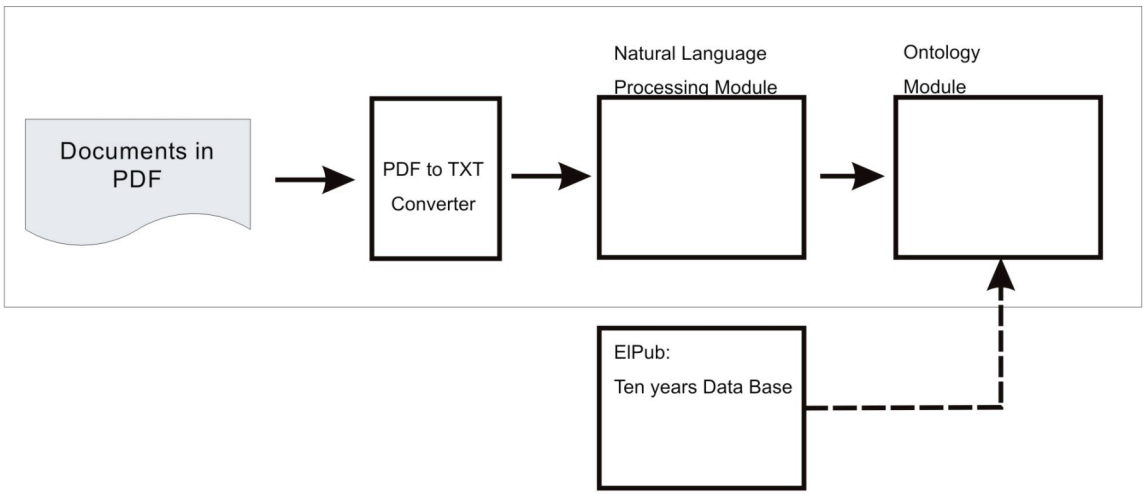


**Figure 1: The process of creating the ElPub/SciX Ontology**

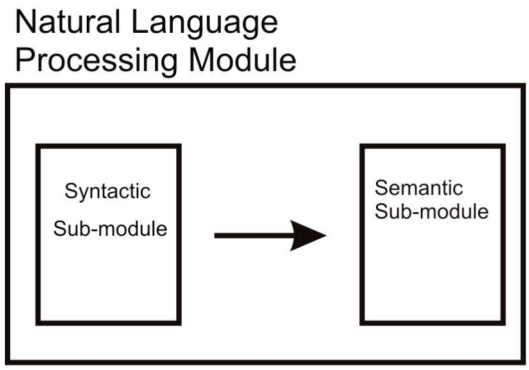


**Figure 2: Detail of the two sub-modules of the natural language processing**

The ontology construction policy adopted pointed to the definition of what constitutes the relevant concepts that should compose the ontology structure (Figure 3).

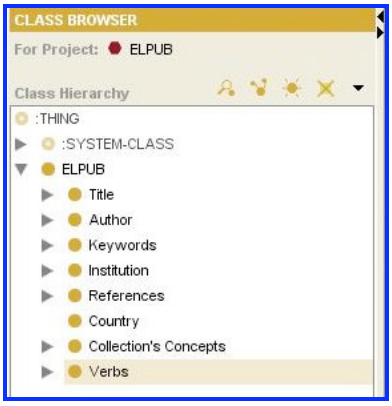


**Figure 3: The ontology structure**

# 4 Results

The procedures carried out comprised the following steps and produced results as exemplified in figures 4, 5 and 6:

- Visit SciX site and collect the entire collection of ElPub papers;
- Transfer the collection into a native database;
- Manually extract titles, author's and institution's names, as well as keywords;
- Replace authors and institution names in the native database by the canonical names created by Costa et al [11]. It is interesting to point out that, for institution names, canonical affiliation strings have been created by applying the rule of putting names given by authors in a standard order. That is: university, faculty/school/institute, department and programme/project, whatever appears. For authors names, the rule was to adopt the most complete form of a name;
- Convert all pdf files into txt files;
- Send the texts (from the introduction to just before the references) to a syntactic analyser (**Syntactic Parser - VISL**), which automatically performs the analysis and generates a syntactic tree with all syntactic tags (example in figure 4);

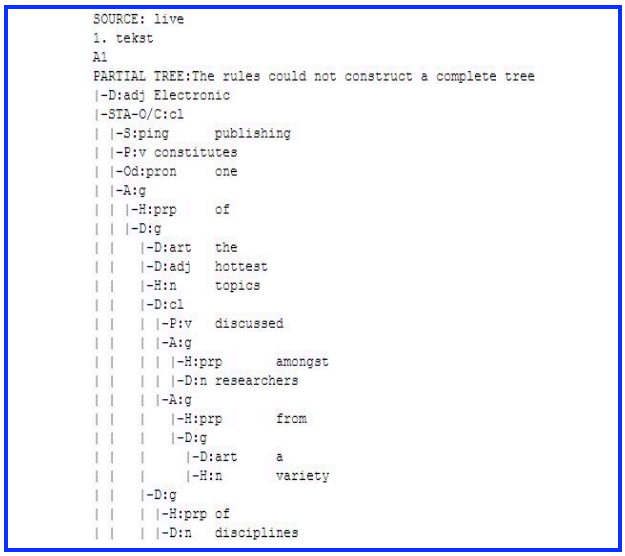


**Figure 4: Syntactic Parser output**

- Send the syntactic tree to *GeraOnto*, which extracts the semantic elements (noun phrases and verbs) of interest for the construction of the ontology. These are concepts that can or cannot be composed of more than one term or concept;

- Insert these concepts into *Protegé*, which edits the ElPub/SciX Ontology, using SciX's record identifiers as its slots (examples of the output are shown in figures 5, 6 and 7).


**Figure 5: Concepts automatically extracted from a text**

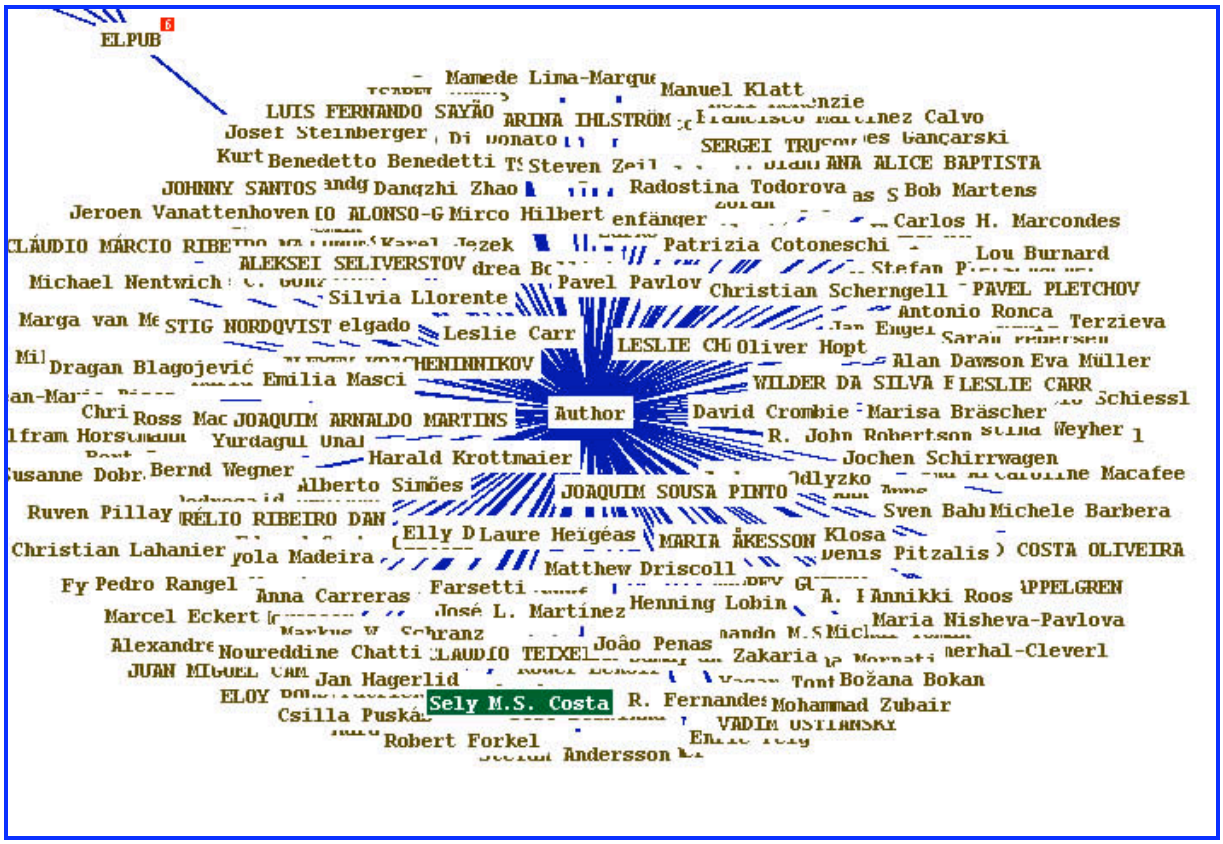


**Figure 6: Graphic presentation of the 'author' super class and its sub-classes**

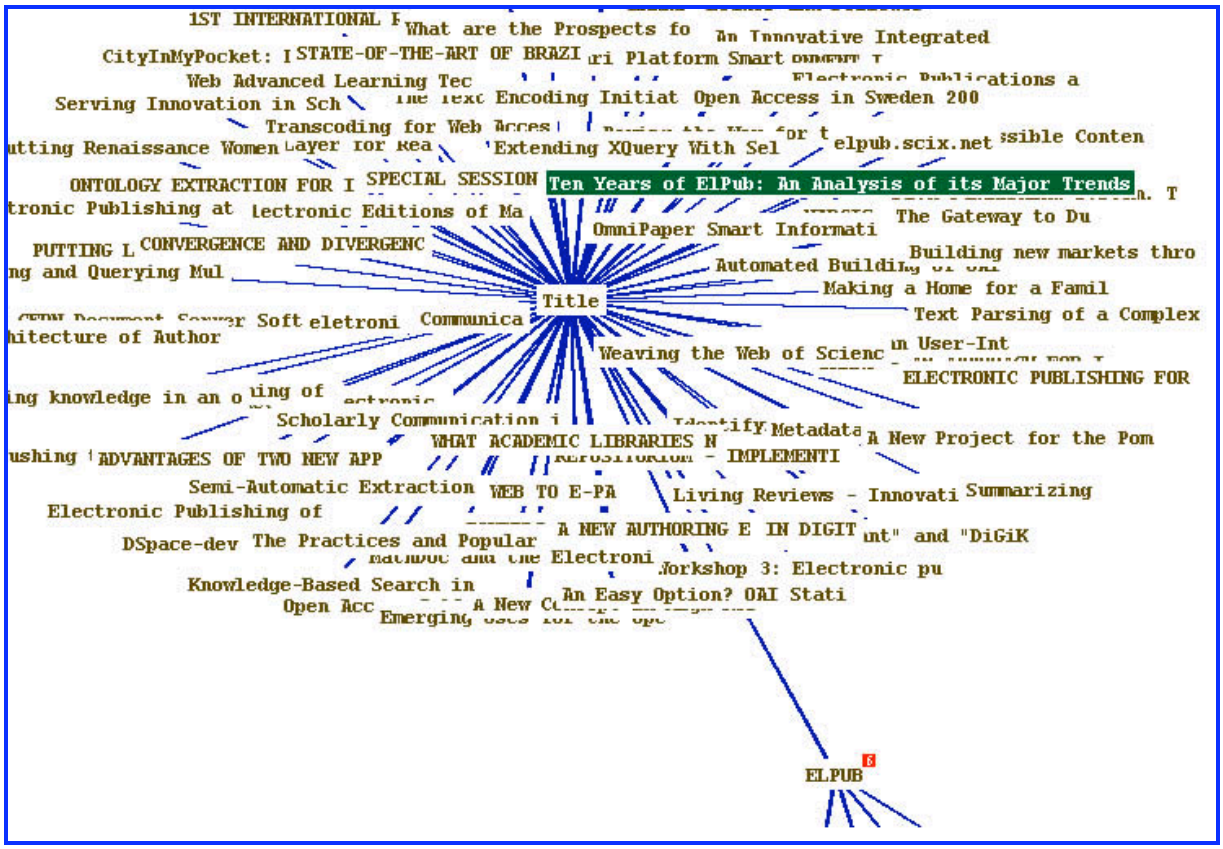


**Figure 7: Graphic presentation of the 'title' super class, with its sub-classes**

Results obtained so far have extracted more than 4,000 concepts. Some of them, especially those related to keywords, need human interference in order to be refined and standardised. Nevertheless, as a work in progress, a number of improvements are still taking place. One of them is to automatically generate the authority names file by creating authors and institutions names as super classes of the ontology, and all their variant names as their sub-classes.

# 5 Conclusions and Recommendations

This proposal minimises, rather than completely eliminates noises identified in information retrieval from ElPub/SciX collection, by creating an ontology. Besides the creation of authority names control, it does reduce ambiguities by means of the syntactic analysis. Instances (SciX record identifiers, such as ELPUB2004_11elpub2004.content.pdf) identified help to prevent ambiguity of explicit repetition of terms. The combined use of Syntactic Parser-VISL, GeraOnto and Protegé has proved to be very helpful and useful for this work.

The resultant ontology will be available at SciX, as well as a tutorial that is intended to be developed and made available for those interested in it. The major learning, however, has been the solutions and strategies identified so far in order to deal with the problem studied. It appeared very clear that the use of such ontology indeed enhances the information retrieval from a collection like ElPub/SciX, particularly because it reduces ambiguities. Regarding keywords, the creation of a controlled vocabulary is highly recommended, based on what is already available on ElPub/SciX collection, in order to guide prospective authors in defining them.

## Acknowledgements

The first author of this work has been generously sponsored by *Finatec* (http://www.finatec.org.br). We are grateful to Rafael Odon de Alencar, who developed java implementations that allowed the development of the ontology with a collection in English, since GeraOnto was created for a Portuguese collection of texts and needed some adaptations.

## Notes and References

[1] COSTA, S. M. S.; BRÄSCHER, M; MADEIRA, F.; SCHIESSL, M. Ten years of ElPub: an analysis of its major trends. In: Martens, B.; Dobreva, M. (Eds.) *Digital spectrum: integrating technology and culture*. Proceedings of the Elpub conference. Bansko : FOI-COMMERCE, 2006. pp. 395-399.

[2] GOTTSCHALG-DUQUE, C. ; LOBIN, H. Ontology extraction for index generation.. In: COSTA, Sely M. S.; ENGELEN, Jan; MOREIRA, A. C. S. (Eds.) *Building digital bridges: linking culture, commerce and science*. Proceedings of the 8th ICCC International Conference on Electronic Publishing. Brasília, 2004. pp. 111-120.

[3] GRUBER, T. *What is an ontology*, 1996. Available at: http://www-ksl.stanford.edu/kst/what-is-an-ontology.html

[4] HAN, H.; ZHA, H.; GILES, C. L. Name disambiguation in author citations using a K-way spectral clustering method. In: *International Conference on Digital Libraries archive*. Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries. New York : ACM Press, 2005.

[5] FRENCH, J. C.; POWELL, A. L.; SCHULMAN, E. Using clustering strategies for creating authority files. *Journal of the American Society for Information Science*, 2000, vol. 51, no. 8, pp. 774-786.

[6] FRENCH, J. C.; POWELL, A. L.; SCHULMAN, E.; PFALTS, J. L. cite the work of Taylor, published in 1984. Their article is about authority files and has been published in 1997, in the proceedings of the ECDL 1997.

[7] COSTA, ref. [1]

[8] GOTTSCHALG-DUQUE, C. *SiRILiCO uma proposta para um Sistema de Recuperação de Informação åbaseado em Teorias da Lingüística Computacional e Ontologia*. Belo Horizonte, 2005.