Submitted on: 04.10.2019

Satellite Meeting: *Knowledge Management: Artificial Intelligence and its Impact on Libraries*
Date: *22 August 2019*
Location: *Ionian University, Corfu, Greece*

Between two worlds: harmonizing automated and manual term labelling

Michalis Sfakakis
Department of Archives, Library Science and Museology, Ionian University, Corfu, Greece.
E-mail address: sfakakis@ionio.gr

Kyriaki Zoutsou
Department of Archives, Library Science and Museology, Ionian University, Corfu, Greece.
E-mail address: k.zoutsou@gmail.com

Leonidas Papachristopoulos
Hellenic Open University Distance Library and Information Center, Patras, Greece.
E-mail address: lpapachristopoulos@eap.gr

Giannis Tsakonas
Library and Information Center, University of Patras, Patras, Greece.
E-mail address: gtsak@upatras.gr

Christos Papatheodorou
Department of Archives, Library Science and Museology, Ionian University, Corfu, Greece.
E-mail address: papatheodor@ionio.gr

Abstract:

In the era of enormous information production human capabilities have reached their limits. The need for automatic information processing which would not be incommensurate to human sophistication seems to be more than imperative. Information scientists have focused on the development of techniques and processes that would assist human contribution while improve, or at least guarantee, information quality. Automatic indexing techniques may lay on various approaches offering different results in information retrieval. In this paper, we introduce an automated methodology for subject analysis, including both the determination of the aboutness of the documents and the translation of the related concepts to system terms. Focusing on a corpus consisting of articles related to the Digital Library Evaluation domain, topic modeling algorithms are utilized for the aboutness of the documents, while the context of the words in topics, as captured by Word Embeddings, are used for the translation of the extracted topics to EuroVoc concepts.

Keywords: Subject Indexing, Similarity Measures, Text Classification, Machine Learning, Word Embedding

1. Introduction

Indexing is considered to be one of the most tedious and debatable processes within the Library and Information Science field. Indexer's difficulties are summarized in the determination (a) of a set of indexing terms for the documents and (b) the degree of a term representativeness (Chu & Ajiferuke, 1989). According to Coates "the cognitive skills required for good term selection include reading comprehension, the ability to conceptualize, and the ability to articulate concepts in a concise and intelligible manner (2002, p.14)." Therefore, researchers and institutions attempted to identify and formalize the indexing process, aiming at its optimization (Hjørland, 2001). For example, ISO 5963-1985 indicates the necessary steps for indexing documents, mentioning that indexers should firstly examine the document and then establish the content. Secondly, they should identify the main concepts and express these concepts according to controlled vocabulary (International Organization for Standardization, 1985).

Indexing quality affects documents' retrieval. In an overactive document producing environment, text mining techniques enhance information retrieval and ameliorate indexers' workload. The implementation of topic modelling on a corpus provide us both its subject overview and a categorization of its documents to specific topics (Papachristopoulos et al., 2016, Papachristopoulos et al., 2015). These topics are not actual indexing terms belonging to a controlled vocabulary but bag-of-words lacking of contextualization. Contextualization means the identification of a term from a controlled vocabulary indicating a specific knowledge area that these words would have significance.

This paper attempts to introduce an approach for document subject indexing based both on topic modelling and automated topic alignment processes, aiming to the improvement of the performance of the indexing and the quality of the indexing terms assigned to a document. Firstly we identify "basic-level terms" via Topic Modelling and subsequently we attempt to upgrade them to a "significance level" according to Thellefsen, Brier and Thellefsen by translating them to specific controlled vocabularies (2003, p.214). Topic modeling was applied on a data set consisted of Digital Library (DL) Evaluation papers published in the proceedings of the most significant conferences of the domain, namely JCDL, ICADL and ECDL/TPDL, during the period 2001-2013. The topic modelling process produced 13 sets of words (topics) needing interpretation and labels. These labels extracted from the EuroVoc thesaurus by translating the generated topics to EuroVoc concepts and then assigned to each topic. Actually, the proposed labelling process aims at the identification of the underlying semantic and context relationships of the terms (words) describing the topics, as produced by topic modelling, with the sets of words labelling and describing the EuroVoc concepts by utilizing the Word Embedding vector of each word. In particular, the similarity of the Word Embedding of the terms resulted by the topic modelling algorithm and the Word Embedding of the concepts of EuroVoc along with their corresponding Scope notes and Definitions will be computed. Then, the most similar concept form EuroVoc will be considered as the candidate subject term for a topic resulted by the topic modelling.

This article is organized as follows: Section 2 demonstrates the recent research on automatic indexing, as Topic Modelling and Word embedding implementations; Section 3 presents the methodological steps of our experimental approach. Section 4 analyses the outcomes of the research, while Section 5 debates the results of the study and signalizes points that future research should focus.

2. Literature Review

The amount of vast digital content productivity puts pressure on the research community for analogous outcomes on digital content management. This is reflected to the steady increase to related scientific literature activity (Pulgarín & Gil-Leiva, 2004). According to Brown & Barrière (2006, p.603) "automatic indexing aims at the automatic creation of a list of index terms associated with a document often in the purpose of text retrieval. Two main approaches are Natural Language indexing and Controlled Vocabulary indexing, respectively extracting words from a text and assigning them from an external lexical resource". Our approach aims at the automatic controlled vocabulary indexing where the terms of a controlled vocabulary (thesaurus, subject headings etc.) would be mapped to a document without human contribution.

Early attempts on the field of automatic indexing were applied on the diagnostic summaries of pathology reports which were attempted to automatically encoded into the Systematized Nomenclature of Pathology via a morphosyntactic approach (Dunham, Pacak, & Pratt, 1978). The authors admitted that their approach is a possible solution, although presented an important drawback the lack of semantic context.

Outcomes' success of automatic subject indexing lies either on the mapping technique or the complexity of subject headings. Névéol et al. (2009) applied various automatic indexing methods ("Jigsaw puzzle" methods, Rule-based methods, statistical) as a recommendation tool in order to assign MeSH terms to MEDLINE documents. Indexers evaluated the aforementioned outcomes as inconsistent and inadequate as they were missing and erroneous recommendations. The whole experimentation highlighted the fact that complex subject headings need more sophisticated approaches.

On the other hand, the emergence of Topic Modelling induced the need for the reduction of human involvement in topic interpretation (Blei, Ng, & Jordan, 2003). While Lau et al. (2010) attempted to label the emerged topics by picking the most appropriate word from the word set of each topic, other researchers tried to develop mechanisms in order to flag these bags-of-words by dragging terms from external resources. For example, ALOT algorithm was applied on topics generated from Topic Modelling in order to map them to a topic Hierarchy obtained from Google Directory Service and the OpenOffice English Thesaurus (Magatti et al., 2009). Another approach for automatic labelling was based on the exploitation of best term of each topic in order to be mapped with Wikipedia terms (Lau et al., 2011).

Lately, Mikolov et al. (2013) managed to attach contextual, morphological, hierarchical and semantical information to each word in a document via a new method called "word embeddings". The aforementioned method will contribute to the materialization of automatic alignment and labelling of our topics derived from topic modelling with the labels used for the concepts in EuroVoc.

The massive document production of EU's bodies, led the EU Publication Office to develop a thesaurus for its better management (Publications Office of the European Union, 2015). EuroVoc is a multilingual tool (23 EU languages) aiming at the terminological standardization within various fields (finance, law, international relations etc.), designed to cover the general needs of EU's publications and not national specific needs. EuroVoc thesaurus is divided in 21 domains (two-digit identification) and 127 microthesauri (four-digit identification). EuroVoc's concepts relationships may be either hierarchical (Broader, BT, or Narrower, NT, term), or associative (Related Term, RT). Its current edition (4.4. edition) includes 6,883 concepts, 4,904 reciprocal hierarchical relationships and 6,922 reciprocal associative relationships.

Our study attempts to take advantage of the specific artificial intelligence progress in order to align the set of words of each topic generated from Topic Modelling with terms from a controlled vocabulary. This alignment will accomplish effectively the first step of topic extraction by assigning to them (topics) established terms reflecting the analogous semantics.

3. Methodology

Corpus formation

Conference proceedings are recognized as a credible channel for the dissemination of a state-of-the-art research. In the Digital Libraries (DL) domain, JCDL, ECDL/TPDL and ICADL conferences constitute historical venues, where anyone can follow the DL's evolution from its incunabulum to the latest advanced state. Accordingly, DL evaluation domain is a crucial domain due to users' expectations for high quality services, content and performance (Fuhr et al., 2007), as challenging due to its interdisciplinary founts which raise the level of complexity (Papachristopoulos et al., 2016). Our experimental study is based on a corpus of papers selected by the proceedings of JCDL (123 papers), ECDL/TPDL (147 papers) and ICADL (125 papers) for the period 2001-2013. A Bayesian classifier was trained to select the DL evaluation oriented papers and three domain experts who worked independently validated its results (Afiontzi et al., 2013; Papachristopoulos et al., 2016).

Topic Modeling implementation

The selected papers were pre-processed so as to generate a "bag of words". We reduced the size of this bag of words removing the most frequent and rare words (above 2,100 or under 5 appearances) and stopwords included in Fox's list (Fox, 1989). The outcome of the pre-processing was used as input to a web-based implementation of the Latent Dirichlet Allocation algorithm, called jsLDA (Mimno, 2018), which generated a pre-defined number of topics. A topic is considered a set of words; the LDA algorithm computes the probability of a word to belong in a topic. Moreover, it estimates the probability of a topic to be included in each document of the corpus. We run the algorithm to produce various numbers of topics (30, 25, 20, 15, 14, 13, 12, 11, 10) and we concluded that 13 was the most well interpretable number of topics that could be generated. The set of terms that describes each topic that derived from topic modelling, were limited to those having probability greater than 0.004.

Word embedding alignment

Word embeddings are word representations extensively used in natural language processing (Li et al., 2015; Mikolov et al., 2013). Each word is converted to a numerical multidimensional vector which contains syntactic and semantic information (Mikolov et al., 2013a; Mikolov et al., 2013b) in order to human language can be understood by computers. In this paper we used word representations from the pre-trained Word Vectors set,[1] trained using fastText open-source library, [2] according to (Mikolov et al., 2018). Pre-trained word vectors set consists of

[1] https://dl.fbaipublicfiles.com/fasttext/vectors-english/wiki-news-300d-1M.vec.zip

[2] https://fasttext.cc/

one million word vectors trained on Wikipedia 2017, UMBC web-based corpus and statmt.org news dataset, while each word representation is a word vector of 300 features.

Thereafter, each one of the words of each of the thirteen topics generated by the LDA algorithm was assigned a word embedding i.e. a word representation by a vector of 300 features. Then for each topic, we calculated the weighted average of the word embedding of the words of the topic, to produce the word embedding of the topic, which is also a vector of 300 features and forms the topic's word embedding.

Respectively, for each concept of the EuroVoc Thesaurus a word vector extracted using words from the concepts' descriptive labels. It is worth mentioning that the SKOS version of EuroVoc was used for our experiments. Concretely, two use cases were evaluated using different sources to select words. For the first case, the words were selected from the skos:prefLabel only, while for the second case the words were selected from a wider set of labels, the skos:prefLabel, skos:altLabel, skos:hiddenLabel, rdfs:label, skos:scopeNote, and skos:definition. For each EuroVoc concept a word vector of 300 features was calculated by averaging the respective word vectors for the words selected from the aforementioned sources. Thus for each use case the word embeddings of the 7,247 EuroVoc's concepts were produced.

Topics to concepts matching

In order to find the most suitable EuroVoc concept for each topic, we calculated the cosine similarity between the word embedding vector of each topic and the word embedding vectors of the EuroVoc concepts. When a topic and a concept are very similar, their word vectors' cosine similarity will be close to 1, while if they are dissimilar, their word embedding vectors' cosine similarity will take a smaller value. The concept with the highest cosine similarity to a topic was considered as the most suitable concept to represent the topic.

As mentioned in the first use case we attempted to align the word embeddings of the topics with the word embeddings of the words in the skos:prefLabel property of the EuroVoc's concepts. In the second use case, we extended the set of the selected words by adding words from the properties skos:prefLabel, skos:altLabel, skos:hiddenLabel, rdfs:label, skos:scopeNote and skos:definition of the EuroVoc concepts and hence we produce more descriptive word embeddings for each concept.

4. Results

The proposed approach seems to be promising as in both cases of our experiment the outcomes of the cosine similarity performance were satisfying. The derived topic alignment with words selected from the EuroVoc's skos:prefLabel label achieved only an average degree of similarity 0.84 (84%), while the alignment based on the words selected from the extended set of the concepts' descriptive properties proved to be more efficient exhibiting 0.88 (88%) average similarity.

Table 1 presents a detailed view of the match. The first and second columns refer to the topics, the third and fourth refer to concepts and last column refers to cosine similarity between the topic and the most similar concept. It is worth mentioning that all cells in the last four columns are split into two parts. The upper part refers to the results from the first case, in which the words for the concept were selected from the skos:prefLabel only, while the lower part refers to the second case, where the words for the concept were selected from the extended set of the concepts' descriptive properties. More specifically, the second column of the table presents the words of each topic with probability to belong to the topic greater than 0.04. The cells of the third column depicts the skos:prefLabel of the most similar EuroVoc concept to the topic derived by each use case. The lower part of the cells of the fourth column shows the words used for the most similar concept derived by the second case; the upper parts of this column are empty because the skod:prefLabel are depicted in the upper cells of the third column. Finally, the fifth column presents the corresponding cosine similarities for both the use cases.

Topic 6 has the best similarity scores and was matched with the concept *'Metadata'* from EuroVoc in both experiments (0.85 and 0.94 respectively). Topic 9 exhibited a significant improvement when its skos:altLabels and skos:scopeNote used in the second case. The matching EuroVoc concept in the first experiment was '*Search engines'* with a similarity 0.86, while in the second experiment the matching concept changed to the more similar one *'Internet site'* providing a similarity 0.89. Also, topic 13 improved significantly its similarity degree from 0.86 to 0.89 respectively.

The alignments for the topic 1 showed a significant amelioration between the two experiments. While it had been matched with EuroVoc's concept labeled *'Persons in work'*, in the second experiment was assigned to the *'Citizen science'* concept.

Table 1. Topics to Eurivoc Concept alignment

Topic	Topic Description	Eurovoc Concept prefLabel	Concept Description	Cosine Similarity
1	participants, study, text, book, students, books, reading, paper, page, notes, read, participant, using, audio, personal, researchers, example, physical, materials, studies, questions, electronic	persons in work		0.79
		citizen science	('prefLabel', 'citizen science'), ('definition', 'Participation of the citizens in the scientific research process in different possible ways: as observers, as funders, in identifying images or analysing data, or providing data themselve.')	0.86
2	similarity, entities, entity, name, names, set, data, features, quality, blocks, chemical, semantic, wikipedia, block, value, based, references, article, attribute, measures, distance, extraction, author, section, values, records, recognition, matching, syllabus, geographic, measure, articles	geographical information system		0.83
		financial derivative	('prefLabel', 'financial derivative'), ('altLabel', 'derivatives market -- derivative financial instrument'), ('definition', 'A financial contract whose value depends on the value of one or more underlying reference assets, rates or indices, on a measure of economic value or on factual events.')	0.88
3	students, resources, learning, design, resource, project, educational, teachers, knowledge, questions, student, education, nsdl, science, technology, research, development, projects, gportal, using, mobile, goals, content, web, course, tools, perceived, concept, teacher, study, classroom, online, teaching, activities	evaluation of resources		0.86
		new educational methods	('prefLabel', 'new educational methods'), ('altLabel', 'parallel school -- open-access school -- experimental school -- educational experiment -- educational innovation -- educational research -- pilot school')	0.90
4	data, system, server, content, service, distributed, node, network, nodes, file, architecture, web, using, files, index, processing, path, key, stored, size, collection, approach, fig, access, client	on line data service		0.84
		cloud computing	('prefLabel', 'cloud computing'), ('altLabel', 'HaaS -- platform as a service -- hardware as a service -- infrastructure as a service -- PaaS -- application service provider -- ASP -- software as a service -- cloud service -- SaaS -- IaaS'), ('definition', 'Storing, processing and use of data on remotely	0.88

			located computers accessed over the internet.')	
5	user, users, paper, papers, algorithm, citation, set, recommendation, cluster, tags, clusters, clustering, using, based, algorithms, similarity, filtering, results, similar, collaborative, personalized, matrix, items, model, profile, approach, method, social, data, research, recommendations, system, experiments, citations, tag, score, weight, category, books, recommender, ranking, figure	information technology user		0.83
		browser	('prefLabel', 'browser'), ('altLabel', 'navigator -- explorer -- browser software application'), ('definition', 'A client program which allows users to look for and read hypertext documents on the world wide web and navigate between them.')	0.869
6	metadata, data, records, resources, content, services, objects, elements, language, quality, service, reference, fields, repository, field, collection, resource, subject, community, project, record, description, repositories	metadata		0.85
		metadata	('prefLabel', 'metadata'), ('altLabel', 'metadata repository -- metainformation -- data about data -- meta-information -- metadata registry')	0.94
7	video, image, images, videos, task, topics, topic, performance, surrogates, text, collection, subjects, recognition, tasks, human, surrogate, visual, content, participants, shots, awareness, activity, pmi, frame, retrieval, user, study, shown, slides, keywords, viewing	video display unit work		0.85
		hypermedia	('prefLabel', 'hypermedia'), ('scopeNote', 'The linking of multimedia to web documents; the integration of text, images, sound, graphics, animation and video through hyperlinks.')	0.87
8	text, words, word, performance, method, table, classification, data, using, results, set, test, average, feature, methods, section, training, model, assigned	on line data service		0.83
		optical character recognition	('prefLabel', 'optical character recognition'), ('altLabel', 'optical reading -- OCR'), ('scopeNote', 'The process of optical character recognition consists in the conversion, by means of a scanner or machine which recognises letters and characters through a mechanism of electronic conversion using a light source, of a numeric image into text, which can then subsequently be modified and processed by means of a word-processing programme.')	0.88
9	search, web, results, page, pages, users, relevance, google, engines, relevant, site, features, information, context, sites, engine, user, topic, content, found, table, using, top, knowledge, online, session, list, queries, links, urls, link, internet, study	search engine		0.81
		Internet site	('prefLabel', 'Internet site'), ('altLabel', 'webpage -- web page -- website -- list of websites'), ('scopeNote', 'To be used only for indexing documents which refer to the creation, development and maintenance of websites.')	0.89
10	information, users, system, evaluation, user, research, process, analysis, specific, systems, collection, provide, support, access, collections, study, particular, provided	management information system		0.90
		access to information	('prefLabel', 'access to information'), ('altLabel', 'public information -- free movement of information')	0.89
11	music, preservation, file, data, cluster, sentiment, files, cost, musical, analysis, songs, example, audio, mood, feature, mturk, format, song, genre, structure, judgments, migration, comparison, section, values	music		0.85
		hypermedia	('prefLabel', 'hypermedia'), ('scopeNote', 'The linking of multimedia to web documents; the integration of text, images, sound, graphics, animation and video through hyperlinks.')	0.85

12	query, terms, queries, term, retrieval, search, using, results, relevant, set, index, subject, precision, top, system, related, thesaurus, retrieved, language, based, relevance, concepts, knowledge, semantic, expansion, recall, phrases, result, information, phrase, title, list, approach, indexing, experiments, example, method	objections to an election result		0.77
		electronic document management	('prefLabel', 'electronic document management'), ('altLabel', 'EDMS -- electronic data management -- electronic document management system -- EDM'), ('scopeNote', 'Refers to the complete range of equipment, software and technical means used in and for storing and archiving data in numerical form.')	0.83
13	user, search, users, interface, task, system, tasks, participants, browsing, interfaces, using, view, results, map, subjects, usability, photos, list, searching, interaction, design, study, information, structure, visual, support, result, photo, figure, tools	information technology user		0.86
		information technology user	('prefLabel', 'information technology user'), ('altLabel', 'information system user -- computer system user -- data-processing system user')	0.89

The words of Topic 3 provide evidence for relevant points to educational content. The first experiment assigned this topic to the Concept *'Evaluation of resources'*, with a similarity score 0.86, while the second experiment assigned it to the Concept *'New educational methods'*, which is more relevant. This improvement is reflected on the rise of cosine similarity to 0.90.

The words of Topic 4 are relevant to the Concept *'Online data services'* (similarity 0.84), while the second experiment improved its assignment to the Concept '*Cloud computing'* (similarity 0.88). Furthermore, it is obvious that the words of Topic 10 are indeed similar to the concept *'Access to information'* (similarity 0.89).

5. Conclusions

The development of an effective workflow for automatic indexing seems not to be a 'chimera', but an issue of well designed workflow, which has to be based on the use of machine learning tools. Our proposed methodology includes tools for text classification, topic extraction and text representation, which are used either for the verification of human involvement (corpus selection and labeling), or for fully automatisation of the workflow.

Obviously, the success of an endeavour like this should be based also on well trained word embeddings. Future work includes an attempt to train the word embedding algorithm in a Digital Library's oriented corpus in order to generate more representational 300-dimensional word vectors.

References

Afiontzi, E., Kazadeis, G., Papachristopoulos, L., Sfakakis, M., Tsakonas, G., & Papatheodorou, C. (2013). Charting the Digital Library Evaluation Domain with a Semantically Enhanced Mining Methodology. *Proceedings of the 13th ACMIEEECS Joint Conference on Digital Libraries*, 125–134. Retrieved from https://doi.org/10.1145/2467696.2467713

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, *3*, 993–1022.

Brown, K., & Barrière, C. (2006). Indexing, Automatic. *Encyclopedia of Language & Linguistics*, 603–610. https://doi.org/10.1016/B0-08-044854-2/00963-9

Chu, C. M., & Ajiferuke, I. (1989). Quality of indexing in library and information science databases. *Online Review*, *13*(1), 11–35.

Dunham, G. S., Pacak, M. G., & Pratt, A. W. (1978). Automatic indexing of pathology data. *Journal of the American Society for Information Science*, *29*(2), 81–90. https://doi.org/10.1002/asi.4630290207

Fox, C. (1989). A stop list for general text. *ACM SIGIR Forum*, *24*(1–2), 19–21. https://doi.org/10.1145/378881.378888

Fuhr, N., Tsakonas, G., Aalberg, T., Agosti, M., Hansen, P., Kapidakis, S., … Solvberg, I. (2007). Evaluation of Digital Libraries. *Int. J. Digit. Libr.*, *8*(1), 21–38. https://doi.org/10.1007/s00799-007-0011-z

Hjørland, B. (2001). Towards a theory of aboutness, subject, topicality, theme, domain, field, content and relevance. *Journal of the American Society for Information Science and Technology*, *52*(9), 774–778. https://doi.org/10.1002/asi.1131

Lau, J. H., Grieser, K., Newman, D., & Baldwin, T. (2011). Automatic labelling of topic models. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, 1536–1545.

Lau, J. H., Newman, D., Karimi, S., & Baldwin, T. (2010). *Best topic word selection for topic labelling*. 605–613. Retrieved from http://dl.acm.org/citation.cfm?id=1944566.1944635

Li, Y., Xu, L., Tian, F., Jiang, L., Zhong, X., & Chen, E. (2015). Word embedding revisited: A new representation learning and explicit matrix factorization perspective. *IJCAI International Joint Conference on Artificial Intelligence*, *2015–Janua*(Ijcai), 3650–3656.

Magatti, D., Calegari, S., Ciucci, D., & Stella, F. (2009). Automatic labeling of topics. *2009 Ninth International Conference on Intelligent Systems Design and Applications*, 1227–1232.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2018). Advances in Pre-Training Distributed Word Representations. *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018).*

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 3111–3119.

Mimno, D. (2018). jsLDA: An implementation of latent Dirichlet allocation in javascript. Retrieved February 10, 2015, from https://github.com/mimno/jsLDA

Névéol, A., Shooshan, S. E., Humphrey, S. M., Mork, J. G., & Aronson, A. R. (2009). A recent advance in the automatic indexing of the biomedical literature. *Journal of Biomedical Informatics*, *42*(5), 814–823. https://doi.org/10.1016/J.JBI.2008.12.007

Papachristopoulos, L., Kleidis, N., Sfakakis, M., Tsakonas, G., & Papatheodorou, C. (2015). Discovering the Topical Evolution of the Digital Library Evaluation Community. In E. Garoufallou, R. Hartley, & P. Gaitanou (Eds.), *Metadata and Semantics Research SE - 9* (pp. 101–112). https://doi.org/10.1007/978-3-319-24129-6_9

Papachristopoulos, L., Tsakonas, G., Sfakakis, M., Kleidis, N., & Papatheodorou, C. (2016). *The "Nomenclature of Multidimensionality" in the Digital Libraries Evaluation Domain*. https://doi.org/10.1007/978-3-319-43997-6_19

Publications Office of the European Union. (2015). *EuroVoc thesaurus Volume 1 Alphabetical version Part B*. Retrieved from http://europa.eu

Pulgarín, A., & Gil-Leiva, I. (2004). Bibliometric analysis of the automatic indexing literature: 1956–2000. *Information Processing & Management*, *40*(2), 365–377. https://doi.org/10.1016/S0306-4573(02)00101-2

Thellefsen, T. L., Brier, S., & Thellefsen, M. L. (2003). Problems concerning the process of subject analysis and the practice of indexing. *Semiotica*, *2003*(144), 177–218. https://doi.org/10.1515/semi.2003.022