Mining Digital Library Evaluation Patterns Using a Domain Ontology

Angelos Mitrelis[1], Leonidas Papachristopoulos[1], Michalis Sfakakis[1], Giannis Tsakonas[1,2], Christos Papatheodorou[1,3]

[1]Department of Archives and Library Science, Ionian University, Corfu, Greece
[2]Library, Neapolis University Pafos, Pafos, Cyprus
[3]Digital Curation Unit, IMIS, "Athena" Research Center, Athens, Greece
{l11mirt, l11papa, papatheodor, sfakakis, gtsak}@ionio.gr

Abstract. Scientific literature is vast and therefore the researchers need knowledge organization systems to index, semantically annotate and correlate their bibliographic sources. Additionally they need methods and tools to discover scientific trends and commonly acceptable practices or areas for further investigation. This paper proposes a clustering-based data mining process to identify research patterns in the digital libraries evaluation domain. The papers published in the proceedings of a well known international conference in the decade 2001-2010 were semantically annotated using the Digital Library Evaluation Ontology (DiLEO). The generated annotations were clustered to portray common evaluation practices. The findings highlight the expressive nature of DiLEO and underline the potential of clustering in the research activities profiling.

Keywords: research trends discovery, clustering, digital library evaluation

1 Introduction

By nature a digital library (DL) is "a very complex and challenging proposition" [1] and therefore its evaluation on aspects of quality, effectiveness and excellence employs many researchers and concerns different domains. Each contributing field of DLs brings in evaluation its own background, terminology and implementation methods. Moreover, many studies can be conducted for the same DL differing at the goals and the used methods. Consequently this diversity indicates the need of decision making mechanisms, based on the existing knowledge, to assist the evaluation experiments planning in terms of its scope, aims, methods and instruments.

This paper exploits the Digital Library Evaluation Ontology (DiLEO) [2], to semantically annotate the literature of the European Conference on Digital Libraries (ECDL, Theory and Practice of Digital Libraries - TPDL since 2011) in the period of the 2001-2010 decade. Clustering techniques are then applied on the derived DiLEO instances in an effort to harvest usage patterns of the ontology in order to investigate how they are actually used so as to suggest evaluation practices to DL researchers. The present study is structured as follows: the next section gives a review of the ma-

jor accomplishments in DL and evaluation conceptual modeling. Section 3 presents the research settings and describes DiLEO ontology. Section 4 gives an overview and discusses the derived results, while in the last section the conclusions are drawn.

2 Background

Through the use of reference models and ontologies, researchers describe the whole life cycle of a DL consisting of services and processes. For instance, Kovacs and Micsik [3] suggested a four-layered ontology consisting of content, services, interface and community which apply as the most important elements of a DL, while Gonçalves' et al. [4] ontology developed the relation among them and introduced the concept of quality as a constituent. The Digital Library Reference Model (DLRM) [5] consolidates a collective understanding of DLs by abstracting the central concepts of the domain. The model defines a set of classes and properties, some of which are related to the evaluation through the concept of Quality. Recently, Khoo and McDonald [6] proposed a model for the evaluation of DL, acknowledging the effect of organizational communication. DiLEO [2] has been developed aiming to conceptualize the DL evaluation domain by exploring its key entities, their attributes and their relationships.

Ontologies are used extensively for the description, analysis and evaluation of the scholarly communication activities. For instance, OntoQualis [7] focuses on the assessment of the quality of scientific conferences, while MESUR [8] provides a framework for the analysis of scholarly communication data, such as references and citations. Furthermore the developments in the nanopublications area [9] provide tools for the semantic integration of the knowledge recorded in the literature enabling its organization and correlation.

In parallel, bibliometric techniques have been used to analyze the evolution of DLs [10] and other related domains, such as information retrieval [11, 12]. Biryukov and Donk investigated the interests of computer science researchers analyzing data from DBLP [13], while Reitz and Hoffman explored the connections and the evolution of topics through time [14]. This paper combines the ontology-based topical analysis of the literature with data mining techniques to reveal the trends of the DL evaluation domain.

3 Research Setting

3.1 The Digital Library Evaluation Ontology

The methodological fundamentals of DiLEO rely on the analysis of various DL evaluation models and the structural composition of the identified characteristics under a unified logic, depicted by a set of constraints.

DiLEO is a two-tiered ontology (Figure 1). The upper, the strategic, layer consists of a set of classes related with the scope and aim of an evaluation while the lower level, the procedural, encompass classes dealing with practical issues. The strategic layer consists of the classes *Goals*, *Dimensions*, *Dimensions Type*, *Research Ques-*

tions, *Levels*, evaluation *Objects* and *Subjects*. Each evaluation initiative is stimulated by a *Goal*, which may be a *description* of a state, the *documentation* of several actions or the enhancement of a *design*. The class *Dimensions* refers to the scope of an evaluation, measuring its *effectiveness*, *performance*, *service quality*, *outcome assessment* and *technical excellence* while the phase in which an evaluation is conducted is characterized as *summative*, *formative* or *iterative*. The *Research Questions* are directly related with the methodological design of the research and the expected findings, while the class *Levels* are the aspects of a DL which are assessed and include *content, engineering, processing, interface* or *individual, institutional* and *social* levels. An *Object* in the evaluation process is either a product or an operation, and a *Subject* is a *human* or *machine agent* participating in the process.

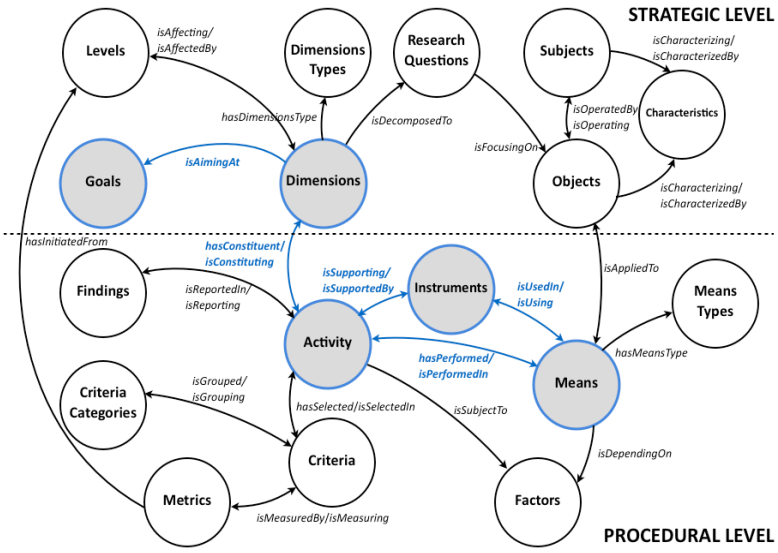

Fig. 1. The DiLEO classes and properties

The procedural layer includes classes that specify practical issues faced by every evaluation exertion. The *Activity* class includes the operations used to denature data to information after their collection by *recording* or *measuring*. Consequently, data are processed through *analyzing*, *comparing* and *interpreting* and finally are *reported* and *recommended*. The activities are affected by the *time*, *cost*, *infrastructure* and *personnel*, constituting the *Factors* class. A variety of *Means* are available, such as *logging studies*, *laboratory studies*, *expert studies*, *comparison techniques*, *field studies* and *survey studies*. DiLEO indicates the *Instruments* that are used, such as statistical analysis software, recording devices etc. Evaluators adopt or develop as reference points the *Objects* that can be valued while the *Criteria* are considered as controlling mechanisms of a measurement as well as a benchmarking process via standards and principles divided in certain *Criteria Categories*. The fulfillment of a criterion is a matter of measurement. *Metrics –user-originated*, *content-originated*, or *system-originated–* illustrate current conditions and indicate the gap between current and ideal states. Finally, the nature of *Findings* is determined by *Research Questions*, but is not

is not specified and not predicted. DiLEO classes are co-related through a set of properties, while for each property particular constraints have been defined to support precise reasoning.

3.2 Semantic Annotation Process

The first step concerned the selection of the ECDL papers' that deal with evaluation. Two researchers worked independently judging the ECDL papers' relevance to evaluation by examining their title, abstract and author keywords (Figure 2). The absolute inter rater's agreement was estimated to be 78%. This result corresponds to a Cohen's Kappa measure –*"the proportion of agreement after chance agreement has been removed"* [15] – equal to κ=0.58, which indicates an acceptable level of agreement. For all the papers that a disagreement was identified a third researcher provided additional rankings. The selection process resulted to identify 119 out of 400 papers (29.5%) as those having evaluation interest.

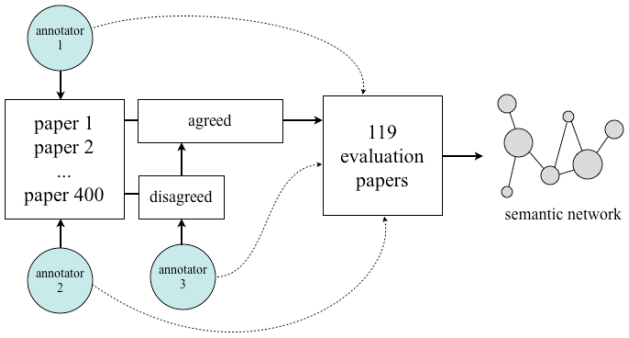

Fig. 2. Selection and annotation processes

The full text of the selected papers was semantically annotated manually by three experts with the goal to identify instances of the 21 subclasses of the DiLEO classes *Goals*, *Dimensions*, *Activities* and *Means* correlated with the properties *isAimingAt* (domain: *Dimensions*/range: *Goals*), *hasConstituent* (domain: *Dimensions*/range: *Activity*), *hasPerformed* (domain: *Activity*/range: *Means*). The annotators were familiar with both the instrument and the domain, agreeing beforehand on several thresholds, similarly to [16], and ensuring a common annotation method. To assess the correctness of annotation, random crosschecks were performed during that phase and any disagreements were resolved through discussion. The result of this process was the generation 1558 triples in the form of domain subclass – property – range subclass.

3.3 Clustering

In our experimentation we opted for K-Means, a well-known algorithm for partitioning *M* data points to *K* clusters. The 21 DiLEO subclasses, used as the vocabulary to annotate the documents, were considered as the features of the data points to be clustered.

Our dataset consists of 119 vectors of 21 features, representing the manually as-

signed annotations from the DiLEO vocabulary to the 119 ECDL papers related to DL evaluation. More specifically, the annotated documents are defined as vectors $A_m = (f_1, f_2, \ldots, f_n)$, where A_m, m= 1, ...,119 denotes an annotated document representation and f_n, n= 1, ..., 21 denotes a feature representing a DiLEO subclass.

The annotation vectors were partitioned to 11 clusters by applying the K-means algorithm. For the selection of the number of clusters K, the K-means was applied for the values of K between 1 and 25 recording the results of the objective (*cost* or *error*) function. Plotting the objective function results against the number of the clusters K the rate of decrease peaked at the values of K in between 9 and 13. Thereafter, we examined the clustering resulted by the K values equal to 11 and 12. The clustering resulted for a K value equal to 11 outperforms the other clustering, according to the evaluation procedure that follows. Moreover given that the K-means is highly dependent on the initialization of the centroids, it was run 200 times using different randomly initialized centroids. Then the clustering with the lowest cost was selected.

Two alternative formulations were tested for the annotation vectors A_m assigning different types of values to the features. In the first formulation, the values of the features were binary, therefore the feature f_i has the value one if the respective to f_i subclass was assigned to the document m, or zero otherwise. In the second formulation the value of the feature f_i was determined according to a variation of the *tf-idf* weighting scheme. To formulate this representation it should be remarked that every DiLEO subclass is annotated at most once to a document. Therefore, the *feature frequency* ff_i of the feature f_i in all vectors will be equal to one when the respective subclass was annotated to the respective document, or zero otherwise. In order to differentiate the values of the frequency of the same feature f_i in the different vectors and to correlate each value to the number of the annotations assigned to the document, the ff_i value is normalized by the number of the annotations to the respective document. Thereafter, given that the ff_i value corresponds to the *tf* part of the weighting scheme and its value is either 0 or 1, the feature f_i in the vector A_j is scored according to the following variation of the *tf-idf* schema:

$$nff_{i,j} = \frac{ff_{i,j}}{|A_j|}, \quad idf_i = \log\left(\frac{M}{1+df_i}\right), \quad tf-idf_{i,j} = nff_{i,j} \times idf_i$$

where, $nff_{i,j}$ is the normalized feature frequency of the feature f_i in the vector A_j, $|A_j|$ is the number of annotations to the document j, M is the number of documents, df_i is the document frequency of the feature f_i and idf_i expresses the inverse document frequency of the feature f_i.

To enable the discovery of patterns by characterizing each cluster with respect to the terms of DiLEO, we used the frequency increase measure (henceforth *FI*), which calculates the increase of the frequency of a feature f_i in the cluster k as compared to its document frequency df_i in the whole dataset. The frequency increase of the feature f_i in the cluster k, $FI_{i,k}$ is defined as the difference of the squares of the two frequency measures as specified by the formula:

$$FI_{i,k} = df_i^2 - ff_{i,k}^2$$

Intuitively, a representative feature for a cluster should be used to annotate a large number of documents in the cluster; hence its frequency in the cluster will be increased. In contrast, for a non representative feature the values of its *FI* measure will

be low and probably negative. The definition of a representative feature for a cluster, is that $FI_{i,k}>\alpha$, where α is the required extent of frequency increase. If $\alpha>0$ then the frequency of a feature within a cluster is greater than the frequency of that feature in the initial dataset. If the *FI* value of a feature in a cluster becomes negative or lower than α then the feature will be filtered out from the representative features of that cluster, even if this feature has been assigned to several documents. The question that arises is to determine objectively the value of the threshold parameter α so that to define a set of features that characterizes a cluster.

In order to estimate the impact of the threshold parameter α, two indicators are defined: the *Coverage* and the *Dissimilarity Mean*. Hence, the selected value for the threshold α is the value that maximizes the combination of the *Coverage* and *Dissimilarity Mean* measures.

Coverage is defined as the proportion of the features participating in the clusters, for a particular *FI* value, to the total number of features used for annotation and is specified by the formula:

$$Coverage = \frac{\left|\bigcup_{k=1}^{K}\left\{f_i: \ FI_{i,k} \geq a\right\}\right|}{N}$$

Dissimilarity Mean expresses the average of the distinctiveness of the clusters and is defined in terms of the dissimilarity $d_{i,j}$ between all the possible pairs of the clusters. Specifically, the dissimilarity $d_{i,j}$ between the clusters i and j, the mean of dissimilarity $DMean_i$ of a cluster i and the *Dissimilarity Mean* are specified by the following formulas respectively:

$$d_{i,j} = 1 - \frac{\left|\left\{f_n: \ \left(FI_{n,i} \geq \alpha\right) \wedge \left(FI_{n,j} \geq \alpha\right)\right\}\right|}{\left|\left\{f_n: \ \left(FI_{n,i} \geq \alpha\right) \vee \left(FI_{n,j} \geq \alpha\right)\right\}\right|} \quad DMean_i = \frac{1}{K-1}\sum_{j=1}^{K} d_{i,j} \quad , \quad i \neq j$$

$$DissimilarityMean = \frac{1}{K}\sum_{k=1}^{K} DMean_k$$

F1-measure is the harmonic mean of the *Coverage* and the *Dissimilarity Mean* and expresses an acceptable balance between them. It is specified as follows:

$$F1-measure = 2 \times \frac{Coverage \times DissimilarityMean}{Coverage + DissimilarityMean}$$

The highest value of the *F1-measure* determines the desired value for the threshold parameter α.

Concluding, given the manually produced annotations based on the DiLEO ontology for the DL evaluation-related ECDL papers, the workflow for discovering patterns of evaluation practices consists of the following steps: (a) representation of the document annotations by two alternative vector models, the binary and the weighted *tf-idf*, (b) clustering of the vector representations of the annotations by applying the K-means algorithm, (c) assessment of the features of each cluster using the frequency increase metric, (d) selection of the threshold α that maximizes the *F1-measure* between the *Coverage* and the *Dissimilarity Mean* indicators, and (v) identification of the evaluation profiles based on the representative features of the clusters.

4 Findings and Discussion

In general the *tf-idf* weighted representation provides more precise results than the binary vector representation and this is evident by the K-Means objective function error for each representation; the error for the *tf-idf* representation is 0.03, while the error of the binary representation is 1.95. Moreover, for most of the threshold α values, both *Coverage* and *Dissimilarity Mean* indicate better performance and permit the selection of a higher value for α. As the value of α increases, the number of the representative features in the clusters decreases, while the clusters become more distinct, generating thus clear evaluation profiles.

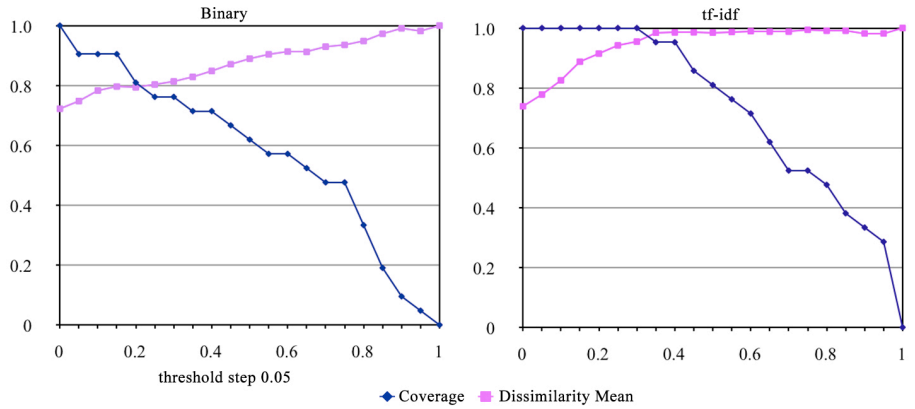

Fig. 3. *Coverage* and *Dissimilarity Mean* for the binary and *tf-idf* representations using different values for α.

Figure 3 presents the curves of the *Coverage* and the *Dissimilarity Mean* with respect to the parameter α, as it varies in the range from 0 to 1 at a step of 0.05. The left part of the figure corresponds to the binary representation, while the right depicts the values of the *tf-idf* weighted representation. It is obvious that the *tf-idf* representation outperforms the binary representation for both indicators and permits a higher value for α to be selected.

In detail, the *Coverage* of the *tf-idf* representation remains equal to 1 meaning that all the features are present in the clustering, up to the point α=0.3. The dissimilarity of the clusters increases quickly and its value is very close to its maximum at the same α value. Concerning the binary representation, the features are eliminated from the first step of α and the distinctiveness of the clusters reaches its maximum value near the highest α value. The outperformance of the *tf-idf* representation is also verified by the *F1-measure* values depicted in Figure 4.

The maximum score for the *F1-measure* is 0.98 when α=0.3 for the *tf-idf* representation, while for the binary representation the *F1* maximizes at 0.85 when α=0.15. Given these results we opted for the *tf-idf* representation and a *FI* value greater than or equal to 0.3 ($FI \geq 0.3$). The features having FI values less than the threshold 0.3 will be excluded from the pattern description. The features resulted by this setting and describe each cluster are presented in Table 1.

All the clusters have at least one strong representative feature with *FI* value greater than 0.69. In particular, in three clusters the highest *FI* is close to 0.70, while the strongest feature of the other eight clusters has a *FI* value greater than 0.94.

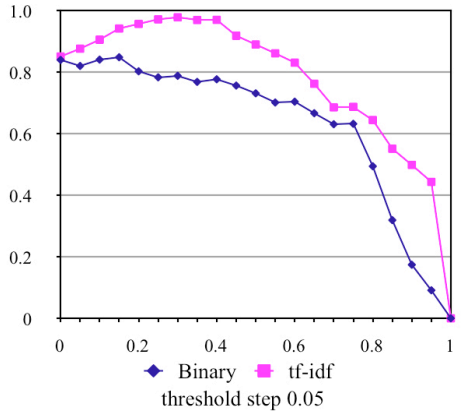

Fig. 4. *F1-measure* scores for the binary and *tf-idf* representations using different values for α

Furthermore in ten clusters the *FI* value of the strongest feature is the highest from all the other *FI* values of the same feature in the other clusters. Hence, these features could be the starting points for formulating the evaluation pattern of the cluster they belong. One exception occurs for the feature *expert_studies*, which holds the same *FI* value in the clusters 5 and 11. However, clusters 5 and 11 have no other common feature, implying that even though the patterns start from the same feature, they follow different paths and express different evaluation practices.

Table 1. The derived models of evaluation profiles and their frequencies

Cluster (Size)	**Features for *FI* ≥ 0.3**
1 (7)	*interpret*: 0.97, *survey-studies*: 0.34, *analyze*:0.32
2 (19)	*description*: 0.69
3 (11)	*log_analysis_studies*: 0.94
4 (12)	*recommend*: 0.98, *effectiveness*: 0.35
5 (5)	*expert_studies*: 0.99, *documentation*: 0.87, *effectiveness*: 0.79, *comparison_studies*: 0.59, *comparison*: 0.42, *service_quality*: 0.34
6 (29)	*design*: 0.70, *technical_excellence*: 0.63
7 (7)	*service_quality*: 0.98, *record*: 0.31
8 (4)	*field_studies*: 0.99, *record*: 0.80, *analyze*:0.32
9 (15)	*documentation*: 0.74, *performance_measurement*: 0.63, *mesure*: 0.43, *laboratory_studies*: 0.41, *comparison*: 0.32
10 (4)	*outcome_assessment*: 0.99, *survey-studies*: 0.83, *analyze*: 0.32
11 (6)	*expert_studies*: 0.99, *technical_excellence*: 0.58, *laboratory_studie*: 0.54, *design*: 0.52, *record*: 0.50, *report*: 0.49, *description*: 0.39

Observing the lowest *FI* values in Table 1, it would be possible to filter more features from the clustering without affecting its structure. By increasing the value of α to 0.5, the features *analyze*, *measure* and *report* are filtered. The document frequency df_i of these three features is higher than 72%, meaning that these subclasses of the

class *Activity* are widely adopted by the most papers. The removal of these features does not subtract any significant information, since it is known that almost every evaluation initiative includes these activities. The low impact of these features is implied by their *tf-idf* values, confirming, the outperformance of the *tf-idf* representation.

In general the evaluation profiles, presented in Table 1, are informative enough for investigating the DL evaluation trends in the last decade. Regarding the size of the clusters, it can be noted that the size of clusters 8 and 10 is small, each of them including four *tf-idf* weighted annotation vectors. However these clusters should not be merged with others, because the features with the highest *FI* value, *field_studies* and *outcome_assessment* respectively, correspond to subclasses not appeared in the patterns of the other clusters. Besides, these features hold the maximum *FI* value (0.99) of all the patterns in the other clusters, indicating their significance as starting points for formulating evaluation patterns.

The advantage of the presented method is that it reveals 'hidden' patterns not usually applied in the literature, such as the pattern of cluster 5, along with patterns representing frequently used evaluation practices, such as the pattern of clusters 8,9 and 10. For instance the pattern of cluster 9 implies that when a study aims to *document* the *performance* of a DL, then it consists of *measuring* activities, held –preferably– in a *laboratory setting*.

Some of the clusters are quite generic in the sense that the number of their representative features is quite small including a couple of features that usually refer to strategic level of the DiLEO schema, such as cluster 6 and cluster 2. This generic pattern corresponds to papers whose first priority is to present a new service or a system and for this purpose they provide evaluation results to describe the current state of that service or system. Cluster 3 consists also of a unique operational feature, the logging studies, indicating a significant trend in the literature.

The clustering process does not emerge any implications for modifying the ontology structure, in the sense that it does not reveal patterns that alter the structure of the ontology reasoning paths. Most of the patterns are identical to sequences of DiLEO triples (domain subclass –property– range subclass), such as the pattern of cluster 6, which is compatible to the DiLEO path *technical_excellence - isAimingAt- design*. This pattern is extended by the pattern of cluster 11, which emerges that when the goal is to improve the design of a DL, then the conduction of expert and laboratory studies contributes to the purpose of the technical excellence investigation; nevertheless the technical excellence investigation implies recording and reporting activities to be held. This pattern generates a set of DiLEO paths, indicatively *technical_excellence –hasConstituent– record* and *record –isPerformedIn– expert_studies*. Given these results, we consider that the application of clustering techniques on the instances of the DiLEO triples would discover complementary knowledge and therefore the future work will address this possibility.

5 Conclusions

Eleven groups of evaluation studies were generated after the semantic annotation of 119 papers and the application of the K-Means algorithm. These clusters, refined by

statistical measures, such as the *FI*, provide meaningful patterns for the evaluation activities presented in ECDL among 2001-2010. We conclude that DiLEO provides the possibility to express meaningful annotations of the DL evaluation literature. Regarding the research questions of our work, we could confirm that the proposed approach can discover solid profiles of the evaluation research landscape, which reflect common practices and identify areas of interest.

References

1. Saracevic, T.: Introduction: the framework for digital library evaluation. In: Tsakonas, G. & Papatheodorou, C. (eds.) Evaluation of digital libraries: an insight to useful applications and methods. Chandos Publishing, Oxford (2009)
2. Tsakonas, G., Papatheodorou, C.: An ontological representation of the digital library evaluation domain. JASIST 62(8), 1577-1593 (211)
3. Kovács, L., Micsik, A.: An ontology-based model of digital libraries. In: 8th International Conference on Asian Digital Libraries, LNCS 3815, pp. 38-43. Springer, Berlin (2005)
4. Goncalves, M.A., Fox, E.A., Watson, L.T.: Towards a digital library theory: A formal digital library ontology. International Journal on Digital Libraries 8(2), 91-114 (2008)
5. Athanasopoulos, G., Candela, L., Castelli, D., et al.: The Digital Library Reference Model. (2010)
6. Khoo, M., Macdonald, C.: An organizational model for digital library evaluation. In: International Conference on Theory and Practice of Digital Libraries 2011, LNCS 6966, pp. 329-34. Springer, Berlin (2011)
7. Souto, M.A.M., Warpechowski, M., Oliveira, J.P.M.D.: An ontological approach for the quality assessment of Computer Science conferences. In: Advances in Conceptual Modeling: Foundations and Applications, pp. 202–212. Springer, Berlin (2007)
8. Rodriguez, M.A., Bollen, J., Sompel, H.V. de: A practical ontology for the large-scale modeling of scholarly artifacts and their usage. In: 7th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 278–287. ACM, New York (2007)
9. Groth, P., Gibson, A., Velterop, J.: The anatomy of a nanopublication. Information Services & Use 3, 51–56 (2010)
10. Bolelli, L., Ertekin, S., Zhou, D., Giles, C.L.: Finding topic trends in digital libraries. In: 9th Joint Conference on Digital libraries, pp. 69-72. ACM, New York (2009)
11. Rorissa, A., Yuan, X.: Visualizing and mapping the intellectual structure of information retrieval. Information Processing & Management 48, 12-135 (2011)
12. Smeaton, A.F., Keogh, G., Gurrin, C., Mcdonald, K., Sødring, T.: Content analysis of SIGIR conference papers. ACM SIGIR Forum 37, 49-53 (2002)
13. Biryukov, M., Dong, C.: Analysis of Computer Science communities based on DBLP. In: 14th European Conference on Digital Libraries, LNCS 6273, pp. 228-235. Springer, Berlin (2010)
14. Reitz, F., Hoffmann, O.: An analysis of the evolving coverage of Computer Science subfields in the DBLP Digital Library. In: 14th European Conference on Digital Libraries, LNCS 6273, pp. 216-227. Springer, Berlin (2010)
15. Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20(1), 37-46 (1960)
16. Roberts, A., Gaizauskas, R., Hepple, M., et al.: The CLEF corpus: semantic annotation of clinical text. In: AMIA Annual Symposium Proceedings, pp. 625 - 629. (2007)