DCPAPERS
Proc. Int'l Conf. on Dublin Core and Metadata Applications 2013

With a Focused Intent: Evolution of DCMI as a Research Community

Jihee Beak
University of Wisconsin-Milwaukee, USA
jbeak@uwm.edu

Richard P. Smiraglia
University of Wisconsin-Milwaukee, USA
smiragli@uwm.edu

Abstract

The Dublin Core Metadata Initiative (DCMI) has played a pivotal role in developing and nurturing a metadata domain. DCMI's conference has become an international venue for metadata researchers and professionals. The purpose of this study was to discover the epistemological consensus and social semantics, if any, of a metadata domain based in DCMI conferences. Specifically, we identified the patterns of emergent and evolving themes over time. To do so we used bibliometric tools including co-word analysis and author co-citation analysis of the DCMI conferences from 2001-2012. The results showed a domain with an underlying teleology (Dublin Core metadata) and with social semantics, represented by semantic coherence in the use of terms. Social semantics also demonstrates shared epistemology as revealed by the co-citation perceptions of the domain. The domain clearly has a focused intent, albeit with a limited focus. User groups are missing from the domain's definition as it emerges in this analysis. There is much room for the domain to nurture so-far under-represented research topics.

Keywords: domain analysis; DCMI conference proceedings; author co-citation analysis; co-word analysis, metadata domain

1. Background

1.1. DCMI Conferences

The Dublin Core Metadata Initiative (DCMI) has played a pivotal role in developing and nurturing a metadata domain. DCMI's conference has become an international venue for metadata researchers and professionals. There have been 12 conferences from 2001 at Tokyo, Japan to 2012 at Kuching, Malaysia. Through the International Conferences on Dublin Core and Metadata Application, various metadata standards and technologies have been studied and introduced. The call for papers for DCMI 2013 refers to participants as "the community of metadata scholars and practitioners" (DC-2013 Call for participation). The body of literature and research on metadata presented at DCMI conferences has grown significantly, but there has not yet been a moment to retrospectively analyze the contents and epistemology of the domain that DCMI has generated. In other words, it is hard to recognize themes of research that have been studied, or to identify subjects or topics that DCMI has not yet broached. This situation suggests the potential usefulness of a broad domain analysis of DCMI conferences.

1.2. Domain Analysis

Domain analysis as a research toolkit was introduced and popularized in Library and Information Science (LIS) by Hjørland and Albrechtsen (1995, pp. 400), for whom a domain-analytic approach is "to study the knowledge-domains as thought or discourse communities, which are parts of society's division of labor." Smiraglia (2012a, pp. 111) notes that domain analysis in Knowledge Organization (KO) is "the act of defining the conceptual knowledge base of a community." As these definitions of domain analysis show, domain analysis has been used to understand thematic patterns as well as to identify emergent ontologies and recognize as pillars the pioneers in a domain. From an epistemological analysis of domain-analytic studies, Smiraglia

https://doi.org/10.23106/dcmi.952136168

(2012a, 114) derived an operational definition of a domain, such that it is: "a group with an ontological base that reveals an underlying teleology, a set of common hypotheses, epistemological consensus on methodological approaches, and social semantics." This paper is an attempt to apply domain-analytical techniques to the proceedings of DCMI conferences to ascertain whether an underlying teleology—i.e., metadata—has generated epistemological consensus and social semantics. The 2013 DCMI conference is titled "Linked to the Future," and has invited contributions that speak to the maturity and persistence of the products of a metadata community, thus suggesting that this historical moment is appropriate for domain analysis.

Scholars from the knowledge organization domain have engaged in both the analysis of specific domains and the meta-analysis of the domain analytical research. Smiraglia has generated an ongoing approach to the analysis of the knowledge organization domain as it is represented in the journal *Knowledge Organization,* as well as in the international and regional conferences of the *International Society for Knowledge Organization* and its chapters (Smiraglia, 2013, 2012b, 2011a, 2011b, 2009). Smiraglia (2012a) presented a meta-analysis of domain-analytical studies, demonstrating especially their usefulness for tracking the emergence of new theoretical domains such as music information retrieval and nanotechnology. To date there has not yet been any study that surveyed the domain of DCMI conferences. Because DCMI conferences play an important role as an international venue and in building a scholarly and professional community for metadata studies, it is important to analyze the DCMI conference domain in order to provide insight into the "intellectual structure of a field" (Zhoa & Strotman, 2008, pp. 15).

2. Methods

The purpose of this study was to discover whether epistemological consensus (represented by common methodological approaches) and social semantics, if any (represented by common vocabulary), could be associated with a metadata domain based in DCMI conferences. Specifically, we identify the patterns of emergent and evolving themes over time using bibliometric tools including co-word analysis and author co-citation analysis. Hjørland (2002) suggested 11 domain-analytic approaches including bibliometric techniques. Smiraglia (2012a, pp. 118) points out that "co-word or term analysis can provide triangulating evidence about the emergence of trends in scholarly domains."

The proceedings of the DCMI conferences are not indexed by Web of Science. We manually collected 4430 citations from 350 papers in the online proceedings of DCMI conferences from 2001 to 2012. DC archives all proceedings in a database of DCMI International Conference on Dublin Core and Metadata Application (http://dcpapers.dublincore.org/pubs/). We studied separately the citations for the 350 papers and the 4430 works cited in them. Data analysis was problematic because of inconsistent citation style and the absence of control of authors' names. Thus we were required to clean the data manually.

3. Results

3.1. Co-Word Analysis of Title keywords

We began by looking for social semantics, a form of coherence in the domain that is represented by term usage. That is, we wanted to know whether the authors had an identifiably common vocabulary. We used the titles of the 350 contributions to DCMI conferences, which we entered into WordStat™ software. Words in the titles were filtered through both an English-language exclusion list of articles, prepositions, etc., and a categorization dictionary that we developed based on keywords in the 350 titles. Nine-hundred and eight terms occurred in the 350 titles. A frequency distribution revealed 18 core terms that were used 15 or more times, each of

https://doi.org/10.23106/dcmi.952136168

which represents more than 2% of the terminology in the entire list. These terms are shown in Table 1.

TABLE 1. Most frequently used title keywords in DCMI papers

	FREQUENCY	% SHOWN
METADATA	191	30.90%
CORE	45	7.30%
DUBLIN	40	6.50%
DIGITAL	35	5.70%
APPLICATION	32	5.20%
BASED	31	5.00%
WEB	29	4.70%
DATA	27	4.40%
SEMANTIC	25	4.00%
LIBRARY	23	3.70%
LEARNING	20	3.20%
FRAMEWORK	19	3.10%
MODEL	18	2.90%
PROFILE	18	2.90%
SYSTEM	18	2.90%
INFORMATION	17	2.70%
RESOURCES	16	2.60%
DESCRIPTION	15	2.40%

A multidimensionally-scaled (MDS) plot of the co-occurrence of these terms helps visualize the semantic coherence of the domain. This is shown in Figure 1. Multidimensionsal scaling uses co-occurrence data to produce visualizations that show proximity among entities in a variable set. In domain analysis MDS is used to help understand how clusters of related work are populated. The map produced from the 350 titles is shown in Figure 1. Here you see how the 18 most-used terms are co-related, or used in relative proximity to each other, within the domain of the 350 titles.

https://doi.org/10.23106/dcmi.952136168

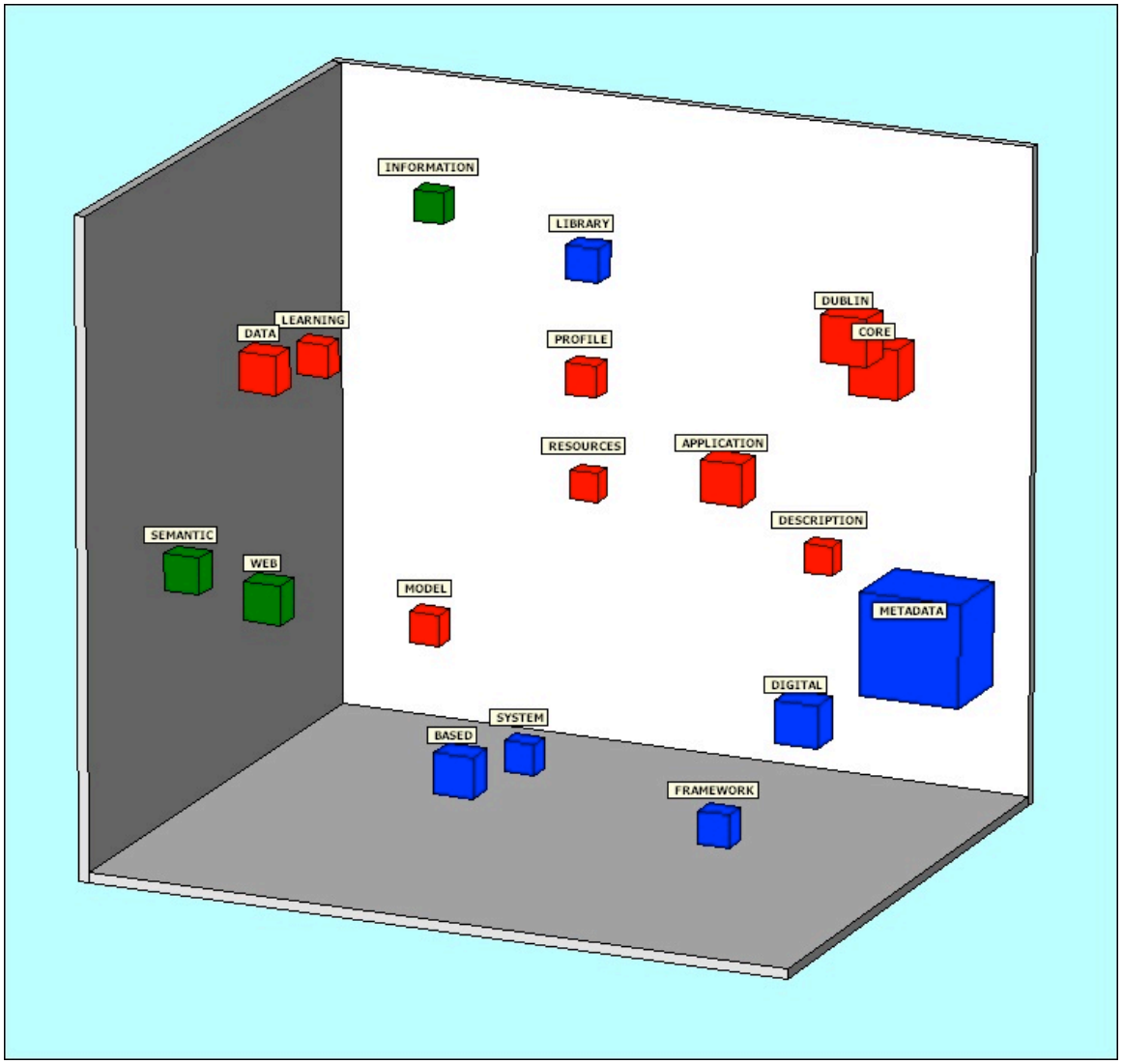

FIG. 1. MDS plot of DCMI title terms (stress = .29512 RSQ = .7891)

For comparison we tried producing plots of keywords from different DCMI conferences, but there was not much gained by doing so; the plot in Figure 1 is the most informative view of the entire domain. The colors demonstrate clusters, the blue representing library metadata, the red representing Dublin Core applications, and the green representing the semantic web. Similarly, we analyzed the titles from the 4430 works cited by DCMI contributors. Words in the titles were passed through the same two filters as before, yielding 4639 terms. Table 2 shows the top of the frequency distribution of terms, in this case terms used more than 100 times, representing 2% of the total or more. The distribution is quite similar, not surprisingly, to the terms in the DCMI paper titles.

TABLE 2. Top of the frequency distribution of terms in citations

	FREQUENCY	% SHOWN
METADATA	936	18.20%
CORE	350	6.80%
INFORMATION	323	6.30%
WEB	309	6.00%

https://doi.org/10.23106/dcmi.952136168

DUBLIN	307	6.00%
DIGITAL	301	5.80%
DATA	251	4.90%
DESCRIPTION	198	3.80%
SEMANTIC	196	3.80%
APPLICATION	188	3.70%
LIBRARY	173	3.40%
FRAMEWORK	157	3.10%
LEARNING	157	3.10%
MODEL	149	2.90%
RDF	141	2.70%
ONTOLOGY	124	2.40%
SCHEMA	124	2.40%
RESOURCE	115	2.20%
XML	113	2.20%
STANDARD	112	2.20%
KNOWLEDGE	108	2.10%
LANGUAGE	108	2.10%
BASED	106	2.10%
PROJECT	101	2.00%

Figure 2 is an MDS plot of these terms.

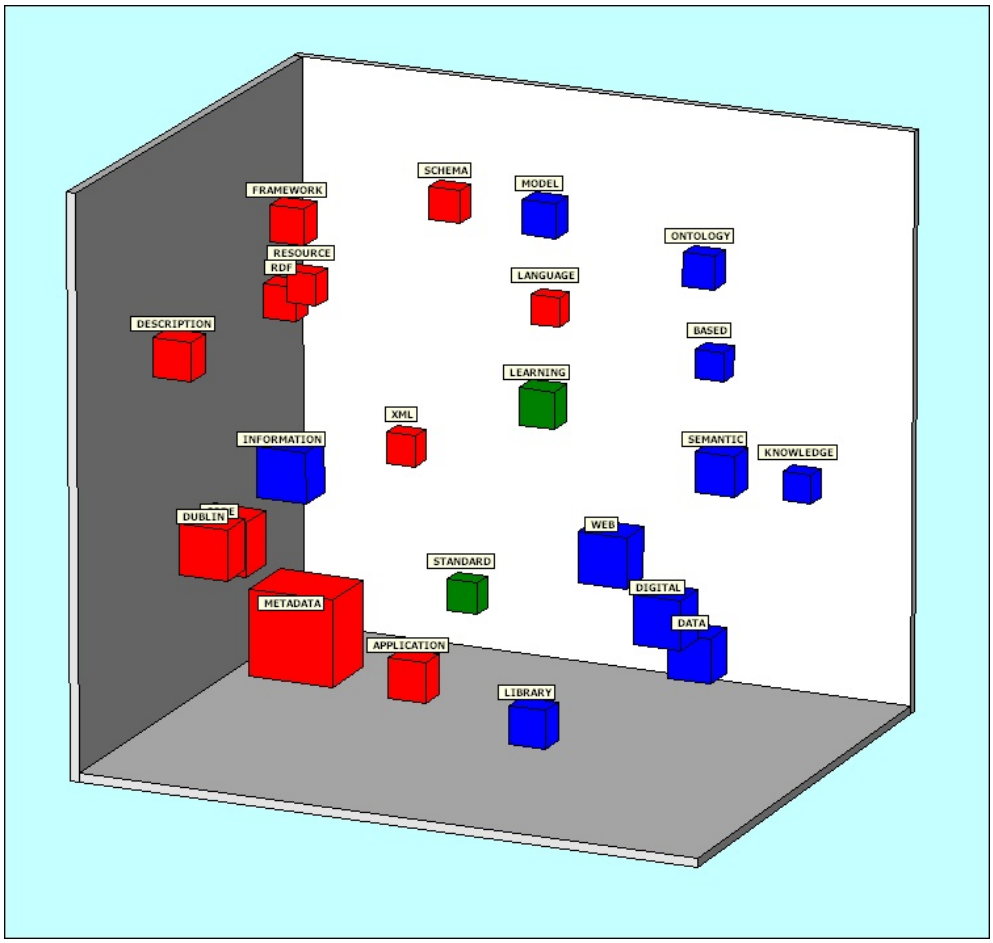

FIG. 2. MDS plot of cited title terms (stress = .27583 RSQ = .7265)

https://doi.org/10.23106/dcmi.952136168

We see three clusters again this time, but with a more meta-level contour; the red cluster contains library applications, Dublin Core, and resource description; the blue cluster contains ontology-based data-models, digital libraries and the Semantic Web; and the small green cluster represents learning standards. Because the clusters represent term co-occurrence, and because we saw similar patterns in both visualizations, we can be fairly certain that there is semantic coherence in the domain on the major clusters: library metadata, Dublin Core, and Semantic Web. There also is an emphasis on applications, with very small clusters representing theoretical research in the DCMI titles, and RDF in the cited reference titles.

3.2. Authors in the research front

We then looked for both the most prolific and the most cited authors in the domain by sorting the citations according to first-named authors. Forty-eight authors were represented with 2 or more contributions to DCMI conferences over time. Apps and Greenberg were the most prolific authors among the top 12 (See Table 3). To identify the most frequently cited authors we excluded web resources and organizations such as W3C. Logoze, Heery, Berners-Lee, Greenberg, and Powell were the most frequently cited authors (See Table 4). We merged the two lists to form what we consider to be a representation of the research front—the most influential contributors.

TABLE 3: Most prolific first-named authors in DCMI conferences.

Authors	Frequency
Apps	6
Greenberg	6
Nevile	5
Francesconi	4
Nagamori	4
Qin	4
Eckert	3
Harper	3
Hillmann	3
Nagarkar	3
Tonkin	3

TABLE 4: First-named authors most frequently cited in DCMI conferences.

Authors	Frequency	Authors	Frequency
Lagoze	47	Brickley	17
Heery	43	Hunter	17
Berners-Lee	38	Weibel	17
Greenberg	37	Bizer	15
Powell	32	Duval	13
Baker	30	Johnston	12
Hillmann	30	Bearman	11
Nilsson	27	Godby	11
Apps	24	Chan	10
Miles	24	Day	10
Coyle	19	Doerr	10
Van de Sompel	18	Lassila	10

https://doi.org/10.23106/dcmi.952136168

3.3. Author co-citation analysis

One way to visualize shared hypotheses and methods, and therefore to uncover epistemological coherence, is to employ author co-citation analysis (ACA). ACA uses the perceived similarities of pairs of co-cited authors to reveal the contours of a domain. "Author co-citation analysis (ACA) is frequently used in domain analysis to help identify active nodes within a domain. ACA measures the perceptions of the authors who are most productive in the domain, about relationships among the researchers they cite, based on the assumption that there is some likelihood that two researchers who are co-cited might be working on similar problems sets" (Smiraglia, 2011a, pp. 6). We began with the authors from the research front represented above, gathering co-citation data from the *Web of Science*™. IBMSPSS™ was then used to generate an MDS plot, which serves as a topical map of the domain. After several filterings of authors with few or no co-citation occurrences, we arrived at the map shown in Figure 3. In ACA, MDS plots visualize the perceived proximity of the named authors based on their theoretical premises. Put another way, the MDS plot creates an image of how citing authors perceive the similarities among the authors they cite. It is important to grasp that we are looking at how the authors named in the plot are perceived by the domain.

The visualization shows two major clusters. A group to the right side includes Greenberg, Day, Qin, Harper, Mason, etc., and another group to the left side includes Doerr, Riely, Van de Sompel, Lagoze, Heery, etc. Based on what we learned from the co-word analysis, we think the two clusters represent the Semantic Web and Library Metadata (the left cluster) and RDF and theoretical research (the right cluster). We think the left cluster represents the more dynamic or predominant cluster in the domain.

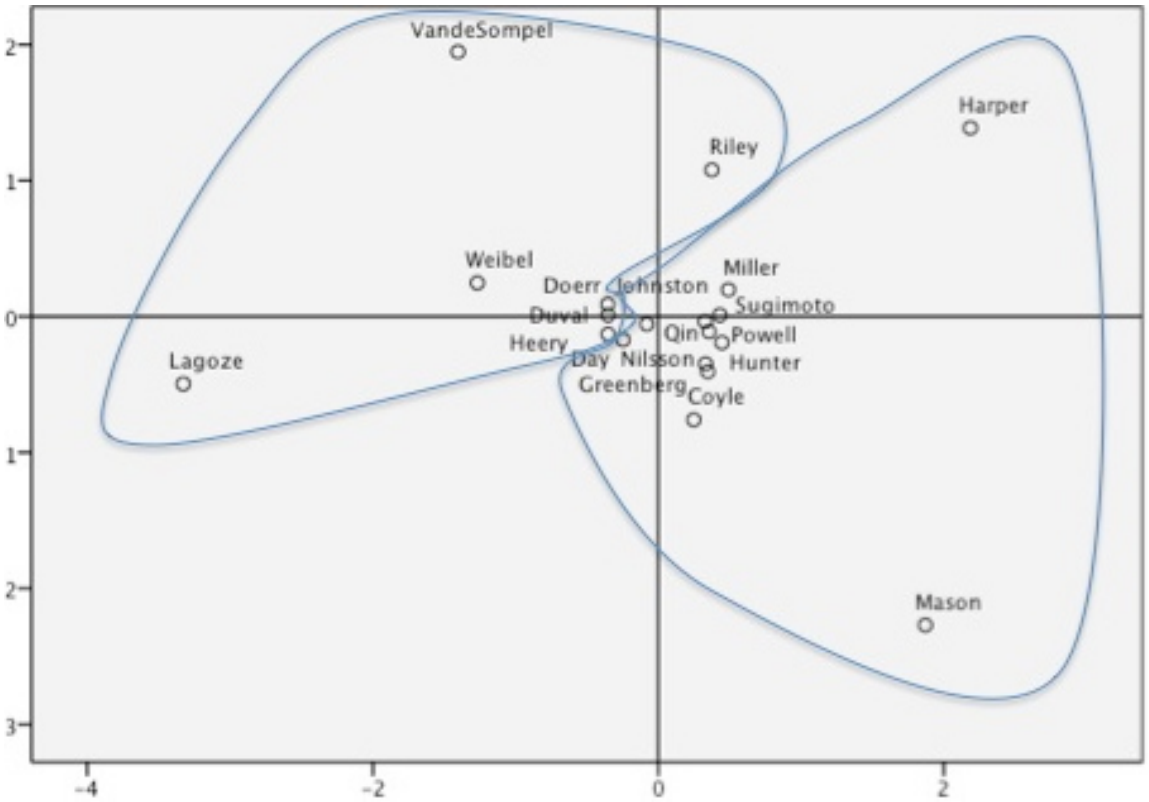

FIG. 3. MDS plot of Author Co-citation in DCMI Conferences (stress = .14510; RSQ = .94453)

https://doi.org/10.23106/dcmi.952136168

4. Concluding discussion

We were interested to reveal the contours of a metadata domain represented by DCMI conference proceedings. The results show a domain with an underlying teleology (Dublin Core metadata) and with social semantics, represented by semantic coherence in the use of terms. Social semantics also demonstrates shared epistemology as revealed by the co-citation perceptions of the domain. The domain clearly has a focused intent, albeit with a limited focus. Library metadata, the semantic web as a concept, and applied and limited theoretical research using RDF seem to be the glue that holds the domain together.

On the other hand, user groups are missing from the domain's definition as it emerges from this analysis. For instance, there is no single study addressing metadata for children's resources (which is not the same as K-12 educational resources or Learning Object Metadata). Metadata as a domain has been discussed in the larger domains of knowledge organization (KO) and information retrieval (IR). Given that there are many studies in both KO and IR concerning children's digital libraries, web portals, OPACs, classification and subject headings etc., it is obvious that children are an important user group in the online environment and a surprise that no such study appears among the DCMI conference proceedings.

We checked the long tails of the term distributions from the co-word analysis to see where terms indicating user groups might have appeared. Among the 908 terms from the 350 titles of DCMI papers the term "user" appeared ranked 61st among terms, and was employed 7 times; "users" fell at rank 201 and appeared 3 times; no other permutations appeared. Among the 4,639 terms from the 4430 cited reference titles the term "user" appeared ranked 96th and was employed 38 times (0.20%); "users" was ranked 442nd, "library user" is ranked 3213, "project user" 3769, and a series of permutations of the term "user generated" appeared ranked 4466 ff.

Limitations of this study included time and space constraints, which in turn prevented analysis of the immense granularity of the domain. That is, we were looking at the upper tier of frequency distributions of 350 papers contributed over a period of 12 years, in which 4,430 works were cited and from which a keyword list emerged containing 4,639 terms. There is ample evidence, then, of great granularity within the domain. Future studies could be designed to analyze that in greater detail. A particular focus on studies embracing users within the domain would certainly be informative. Another serious limitation emerged from the necessity to manually index the proceedings, including extensive cleaning of over four thousand often irregularly constructed citations.

Then again, although there was evidence of much granularity, it was not clear to what extent this represents as yet undeveloped potential for new research streams, as opposed to temporarily interesting concepts, or new ideas that were proposed and rejected by the domain. Time series analysis might yield better evidence here, although, as we stated above, separate analyses of proceedings from 2001, 2005 and 2011 failed to yield a different map of core terms. Our evidence is inconclusive on this point. This suggests, however, that a future direction for the domain might be to extend DCMI conferences by including diverse research topics and to continuously or more deliberately nurture the minor research topics.

References

DC-2013 Call For Participation. Available http://dcevents.dublincore.org/index.php/IntConf/dc-2013/schedConf/cfp

Hjørland, B. & Albrechtsen, H. (1995). Toward a new horizon in information science: Domain-analysis. *Journal of the American Society for Information Science, 46*(6), 400-425.

Smiraglia, Richard P. (2009). Modulation and specialization in North American knowledge organization: Visualizing pioneers. *Proceedings North American Symposium on Knowledge Organization, 2*, 35-46.

https://doi.org/10.23106/dcmi.952136168

Smiraglia, Richard P. (2011a). I Simposio Internacional sobre organizacion del conocimiento, bibliotecologia y terminologia: An editorial. *Knowledge Organization, 38*(1), 3-8.

Smiraglia, Richard P. (2011b). ISKO 11's diverse bookshelf: An editorial. *Knowledge Organization, 38*(3), 179-186.

Smiraglia, Richard P. (2012a). Epistemology of domain analysis. In R.P. Smiraglia and H. Lee (Eds.), *Cultural frames of knowledge*. (pp. 111-124). Würzburg : Ergon

Smiraglia, Richard P. (2012b). Universes, Dimensions, Domains, Intensions and Extensions: Knowledge Organization for the 21st Century. *Advances in Knowledge Organization, 13*, 1-7.

Smiraglia, Richard P. (2013). ISKO 12's bookshelf – Evolving intension: An editorial. *Knowledge Organization, 40*(1), 3-10.

Zhao, Dangzhi, & Strotmann, Andreas (2008). Information science during the first decade of the web: An enriched author cocitation analysis. *Journal of the American Society for Information Science & Technology, 59*(6), 916-37.

https://doi.org/10.23106/dcmi.952136168