An Exploration of the Digital Library Evaluation Literature Based on an Ontological Representation

Giannis Tsakonas, Angelos Mitrelis, Leonidas Papachristopoulos, and Christos Papatheodorou

Database and Information Systems Research Group, Department of Archives and Library Science, Ionian University, 72 Ioanni Theotoki Str., 49100, Corfu, Greece. E-mail: {gtsak, l11mitr, 11papa, papatheodor} @ionio.gr

Evaluation is a vital research area in the digital library domain, demonstrating a growing literature in conference and journal articles. We explore the directions and the evolution of evaluation research for the period 2001–2011 by studying the evaluation initiatives presented at 2 main conferences of the digital library domain, namely the Association for Computing Machinery and the Institute of Electrical and Electronics Engineers (ACM/IEEE) Joint Conference on Digital Libraries (JCDL), and the European Conference on Digital Libraries (ECDL; since 2011 renamed to the International Conference on Theory and Practice of Digital Libraries [TPDL]). The literature is annotated using a domain ontology, named DiLEO, which defines explicitly the main concepts of the digital library evaluation domain and their correlations. The ontology instances constitute a semantic network that enables the uniform and formal representation of the critical evaluation constructs in both conferences, untangles their associations, and supports the study of their evolution. We discuss interesting patterns in the evaluation practices as well as in the research foci of the 2 venues, and outline current research trends and areas for further research.

Introduction

Evaluation is one of the most critical areas of the digital library (DL) domain. In general, evaluation covers all aspects of DL’s development and operation, including components, products, and processes, to provide rich information about DL’s effectiveness, quality, and excellence. The diversity of the DL applications field means that evaluation is a very sophisticated procedure. DL is admittedly a multi-disciplinary domain, comprised of different research disciplines, such as information retrieval (IR), information science and librarianship, human computer interaction, etc. and scholars use divergent paradigms, methods, or tools. This creates a “Tower of Babel” effect; each contributing field introduces to the evaluation its own background, terminology, and implementation methods.

As a research area attached to a highly evolving domain, evaluation in DLs has matured and developed, experiencing variations in its conceptualization, the role it has in DL design, and the ways it is applied. To trace these changes, studies need to be implemented to provide the historical view of specific constructs and to reflect the generation of event patterns, such as research practices and associations. A second aim would be to provide tools for the planning of DL evaluations and the profiling of such initiatives in terms of their scope, aims, methods, and instruments. To be truly effective, any such product should be solidly grounded in existing knowledge about the DL evaluation literature.

In this article, research trends in the evaluation of DLs are examined, as revealed in the literature of two of the most important conferences of the field. The conferences, namely the Association for Computing Machinery and the Institute of Electrical and Electronics Engineers (ACM/IEEE) Joint Conference on Digital Libraries (JCDL), and the European Conference on Digital Libraries (ECDL; since 2011 renamed to the International Conference on Theory and Practice of Digital Libraries [TPDL]), are considered important venues for the expression of areas of interest, methods, and instruments of the field of DLs. The literature was annotated using the Digital Library Evaluation Ontology (DiLEO; Tsakonas & Papatheodorou, 2011), a domain ontology for DL evaluation. This study exploits DiLEO as a means to semantically annotate this segment of the literature, to restructure the landscape of evaluation research, and to present a rigorous analysis of research trends in this given period. The following research questions were posed:

- How were DL evaluation efforts oriented? What was the relative significance among critical evaluation constructs in the two conferences? How strongly were these constructs related?
We used network analysis techniques to explore the semantic network created from the annotation of the literature. Our analysis is based on the betweenness centrality measure, which places each ontology concept in a position of relative importance.

- How did the areas of interest (instances of the class Dimensions in DiLEO, to be later outlined) and the methods (instances of the class Means in DiLEO) evolve during the period 2001–2011 in both conferences?

Temporal analysis is used to investigate the evolution of these two concepts introducing three measures to estimate the diversity degree of the evaluation studies.

The ontological schema aims at a formal and consistent representation of evaluation patterns that cross over abstract and concrete concepts, such as the goals of the evaluation initiatives and the expression of data gathering and analytical research activities. The study also sheds light on the foci of DL evaluation as revealed in these two conferences, which took the form of performance measurement, effectiveness, and technical excellence. These foci highlight a preference for system-centered research carried out through evaluation experiments inside laboratories and through surveys. Here we validate DiLEO as a tool that conceptualizes the DL evaluation landscape, without speculating about the contextual characteristics and factors that differentiate the experimental settings of the various studies. The article is organized as follows. In the next section, we provide a background on analytical studies and ontologies as semantically rich descriptive schemas, then we present the setting of our research, including a concise presentation of the instrument and the corpus. This is followed by a section on the major findings of the analysis, then a discussion section. We present our conclusions in the final section.

Background

Evaluation has drawn the attention of many researchers in the DL field. In most cases, their background, approaches, and research settings differ, comprising a synthesized domain that addresses classic retrieval problems in the DL environment, such as precision and recall, as well as more user-centered issues, such as usability and user experience. Saracevic’s (2004) meta-analysis of the DL evaluation literature distinguished two large groups of studies; the “meta,” those discussing evaluation models, and the “object” studies, those presenting actual evaluation activities. Based on this, Saracevic articulated vital procedural questions regarding the constructs, context, criteria, and methodology of evaluation. Führ et al.’s (2007) seminal study addressed these questions, but also proposed a blueprint for the evaluation constructs that outlined future directions for DL evaluation research, whereas Gonçalves, Moreira, Fox, and Watson (2007) presented a quality model, including constructs and metrics, in a functional way. A vivid discourse on the most significant evaluation criteria, mostly in the user-centered approach, has followed since then with most prominent the works by Xie (2008), Zhang (2010), and more recently by Heradio, Fernández-Amorós, Cabrero, and Herrera-Viedma (2012), whereas Khoo and MacDonald (2011) combined previous models, such as Porter’s (1998) value chain and logic models, to create and validate a new one that takes into account the organizational effects on DL evaluation.

Meta-analyses of research studies, including also evaluation ones, have various subjects, such as the extent of use of specific methods, topical shifts, temporal trends and the evolution of communities. Fidel (2008) studied the use of mixed methods research (MMR) in librarianship as it was reflected in four key journals and concluded that the MMR approach is more likely to occur during the research design and analysis stages. Similarly, Hider and Pymm (2008) investigated the use of empirical methods in 20 library and information science (LIS) journals. They reported that most researchers employed surveys and experiments in their studies, with a growing trend to the use of ethnographical methods in LIS.

Studies that emphasize shifts of interest often use network analysis to illustrate the associations between thematic areas. Sugimoto and McCain (2010) used pathfinder network (PFNNet) analysis to visualize and explore topical changes in the IR domain across three decades. They identified structural changes in the research interests of the community that were explained by contextual developments and technological shifts. Similarly, Reitz and Hoffmann (2010) used data from the DBLP bibliographic service to investigate the centrality of thematic areas and their in-between associations in the database over time.

Larivière, Sugimoto, and Cronin (2012) undertook a temporal analysis of the LIS field by examining data from 1900 to 2010 and combining the creation of literature’s historical profile with lexical analysis to highlight journals’ and authors’ activity and term and topic evolutionary behavior, and to underline the interdisciplinary nature of its scientific environment. Chang and Huang (2012) synthesized three different bibliographic techniques, namely direct citation, bibliographic coupling, and coauthorship analysis, to explore the progress of LIS publications during the last 30 years in terms of interdisciplinary associations. Lee, Kim, and Kim (2010) used data mining techniques to cluster representations of DLs in LIS journals and database descriptors over a period of 15 years (1994–2008) and identified three main phases of development as well as the journals of critical importance per phase.

Many studies, characterized by their bibliometric orientation, concentrate on the evolution of scientific communities, venues, and actors often using as a data source the emerging repository of proceedings publications, as in this study. Hofer, Smajkal, Bilgin, and Wuehrer (2010) perceive “conference proceedings as a source of and basis for the measurement of scientific knowledge, ensuring a holistic and detailed perspective of current research” (p. 846). Additionally, Glänzel, Schlemmer, Schubert, and Thijs (2006) described conference proceedings as a valuable supplement to journal publishing and their bibliometric analysis based on the Thomson Reuters proceedings database aimed at
highlighting the rich information flow that exists within this specific scholarly communication channel. Drott (1995) has argued that conference papers are more than “ephemeral literature” (p. 299). Proceedings should not be considered as an inferior type of scientific publication, but as a parallel sphere where a different kind of scholarly speech is communicated and as such they hold their own research interest for the scholarly community. Smeaton, Keogh, Garrin, McDonald, and Sodring (2003) analyzed the articles of 25 years of the Special Interest Group on Information Retrieval (SIGIR) conference to trace the evolution of topics and to predict areas of interest for the succeeding years, whereas Martins, Gonçalves, Laender, and Pappa (2009) exploited several conference characteristics, such as tradition, acceptance rate, and program committees, to automatically generate quality indicators. Biryukov and Dong (2010) explored the performance of computer science researchers in related conferences by analyzing data from DBLP and found interesting patterns about their career length, performance, and work distribution in both top and nontop venues.

In this study, a domain ontology is used to semantically annotate the literature and to explore the dynamics of the DL evaluation domain. The strength of ontologies in expressing explicitly and formally the concepts of a domain has enabled the modeling of semantically dispersed knowledge areas, such as DLs. Gonçalves, Fox, and Watson (2008) extended the work on the 5S framework and proposed a formal ontology to define workflows inside DLs. In their elaboration of the practical implementation of their modeling schema, they presented an example focused on the development of quality indicators. The constructs of ontology, as its vocabulary, can be used to address the multidisciplinarity of a field or the abstraction of modeling techniques. Ohren (2009) translated the DELOS reference model (Athanasiopoulos et al., 2010), an abstract, yet robust conceptual model of DLs, to web ontology language (OWL) and created the related vocabulary (DLRM [Digital Library Reference Model] vocabulary). Ontologies have also been used to represent domains and activities related to scholarly communication, such as scholarly event description (SEDE) (Jeong & Kim, 2010) for scientific events, OntoQualis (Souto, Warpechowski, & Oliveira, 2007) for the assessment of the quality of the latter and metrics of scholarly use of resources (MESUR) (Rodriguez, Bollen, & Sompel, 2007) such as references and citations. These examples portray the expressive power of ontologies in formally explaining domain parameters, as well as their reasoning strength in decision-making circumstances. Other ontologies have been developed to represent and to guide research synthesis in evolving domains, such as software engineering. Biolchini, Mian, Canadila, Conte, and Travassos (2007) state that the usefulness of an ontology “. . . ranges from providing semantic and terminological support to the researcher, during the planning and execution phases of a systematic review in the SE field, up to supplying a formal conceptual framework to help examining the results of the studies under analysis” (p. 135). Pepe, Mayernik, and Borgman (2010) have illustrated the declarative power of the open archives initiative object reuse and exchange (OAI-ORE) model to reflect relationships between research entities so as to present semantically enhanced versions of common scientific workflows. Soldatova and King (2006) developed the EXPO ontology to express domain independent research processes in their attempt to provide a formal description of experimental designs, while Brahaj, Razum, and Schwichtenberg (2012) focused on the information emerged from the scientific laboratory context in their Core Scientific Model Ontology (CSMO) ontology. Recently, developments in the nanopublications area (Groth, Gibson, & Velterop, 2010) showcased the expressivity of ontologies to support the mining of semantically enriched information in literature corpora. Using scholarly communication-enabling ontologies, researchers can create assertions (descriptive statements of scientific information), extract knowledge, and share them both in human and machine-readable formats—something that can be facilitated by means of semantic annotation. The use of ontologies in semantic annotation has been previously illustrated in content classification, query expansion, automatic metadata descriptors extraction, interoperability facilitation, and so on (for a concise review refer to Uren et al., 2006). The application of ontologies covers various forms of content, such as text (Schreiber et al., 2008), images (Schreiber, Dubbeldam, Wielemaaker, & Wielinga, 2001), and video (Athanasiadis et al., 2005). Ontologies and related constructs, such as vocabularies and taxonomies, have been proven to be important research tools for creating semantically enhanced versions of the literature, especially in the field of biomedicine. Roberts et al. (2009) illustrated the case of generating a semantically enhanced corpus of biomedical information using domain vocabularies. Without the implementation of typical knowledge organization systems, such as ontologies, Roberts et al. (2009) employed the Cross-Language Evaluation Forum (CLEF) annotation schema—a fine-grained modeling effort for annotations—to develop a methodology, as well as to create a semantically tagged body of the biomedical literature for reuse and analysis. Cui (2010), following a reverse approach, used the semantically tagged literature of the biodiversity domain to assess the quality characteristics of four ontologies/glossaries in terms of completeness and semantic consistency. These examples underline the need for knowledge representation schemas as aid mechanisms for the review and analysis of the constantly growing—and in diverse semantic forms—literature.

Research Setting

Annotation Schema

DiLEO (Tsakonas & Papatheodorou, 2011) constitutes an instrument for the unified conceptualization of the DL evaluation domain reflecting the main concepts of DLs and their evaluation. The development of this ontology was motivated by the need to (a) define and clarify the main concepts of the DL evaluation field as well as their relationships; (b) address the lack of globally accepted abstract evaluation models, as
suggested by Führ et al. (2007), enabling the analysis of existing research activities in a formal way; and (c) transform this information into a reusable corpus of knowledge. DiLEO was created by exploring key evaluation models, identifying principle concepts, and structuring them in a hierarchical schema. This top-down process was supported by the development of a vocabulary through the mining of other constructs, such as instruments and means as they were reflected in a large number of evaluation studies. This mixed approach led to iterative development phases and the creation of a coherent first version, which was later validated and refined in two qualitative studies. The first one was a panel that involved DL practitioners and the second a focus group comprising information science experts. This validation and refinement stage highlighted several issues, like the ability to conduct evaluation plans in the absence of pre-defined entry points, the assembly of instances in more complex initiatives, and the clarity level of some concepts. Since then, the ontology has been subject to several refinements in pursuit of a wide, yet generic representation of the domain.

DiLEO provides a two-tiered conceptualization of the DL evaluation domain. The structure of the ontology is illustrated in Figure 1, which highlights the classes and the properties that are within the scope of this study. In the upper level, named strategic, abstract concepts of the evaluation studies are identified, such as Goals, Dimensions, Dimensions Type, Levels, Research Questions, Subjects, Objects, and Characteristics. A DL evaluation study is characterized by a Goal, which may be either the description of the current state of the DL, or the documentation of actions taken or the DL design optimization. The Dimensions class refers to the reasons an evaluation is conducted and its anticipated outcomes and consists of five subclasses, namely effectiveness, performance measurement, service quality, outcome assessment, and technical excellence. Dimensions Type characterizes the phase in which an evaluation is conducted as summative, formative, or iterative. Research Questions are related to Dimensions, documenting the evaluation study and setting its scope, while the class Levels defines the aspects of the DL that can be affected by the evaluation findings, which can be the content, engineering, processing, interface, individual, institutional, and social levels. Every evaluation process involves Objects, which may be a product or an operation, as well as Subjects, such as a human or a machine agent, who take part in various ways in the Objects evaluation. Each of them has specific Characteristics which play significant role in the evaluation process.

FIG. 1. The two-tiered organization of DiLEO.

1In the rest of the article, the DiLEO classes, subclasses, and properties are in italics. The first letter of a class label is capitalized.
The lower part of the ontology, named procedural, illustrates the procedural aspects of the DL evaluation. This level includes the classes Activity, Instruments, Findings, Factors, Criteria, Criteria Categories, Metrics, Means, and Means Types. Activity categorizes all the actions taken for the data collection (recording) and measurement, their analysis, comparison and interpretation, and finally the results and the derived recommendations. The Means class represents all the methods used by the DL evaluation activities, such as logging studies, laboratory studies, survey studies, expert studies, comparison techniques, and field studies. There are two general Means Types, namely qualitative and quantitative means. For the application of Means, research tools are needed constituting the Instruments class. These instruments can be research artifacts, such as paper prototypes or test collections, narrative items, such as questions, devices, such as cameras, and software, such as analysis software. A crucial parameter of an evaluation experiment is Criteria, which can be standards, principles, or commonly agreed methods used as stable units of measure for comparison and classified under certain Criteria Categories. Metrics may be user-originated, content-originated, or system-originated. These metrics illustrate the current condition of a system and indicate the distance between current and ideal states of its operation. Furthermore, an evaluation study is affected by several Factors, such as time, cost, required infrastructure, and personnel. Finally, Findings are the data gathered after the implementation of the evaluation process and its analysis and they are directly related to Research Questions.

The ontology provides a set of properties and each of them has a domain class and a range class. The properties formulate concrete paths, enabling the coherent description of evaluation experiments and revealing the semantics of particular evaluation practices. Specific constraints on the properties have been defined to express exceptions and to provide precise reasoning possibilities. Some properties correlate the strategic with the procedural classes. For instance, the property hasConstituent (domain class is Dimensions and range class is Activity; the inverse property is isConstituting,) denotes that an evaluation process with a particular scope is consistent with some activities that address this scope. The ontology instances constitute a knowledge base, in the form of a semantic network, to which a researcher can pose queries and extract knowledge on evaluation initiatives.

Corpus

The data set consists of evaluation studies presented in two significant venues for scholarly communication in the area of DLs, namely the ACM/IEEE Joint Conference on Digital Libraries and the European Conference on Digital Libraries (now TPLD) from 2001–2011. Only short and full-length articles were used, and all documents that reported posters, panels, and demos were excluded from the sample. Furthermore, user studies that were not linked with evaluation initiatives or system development were also excluded, as well as articles with evaluation interest, such as the presentation of evaluation models or instruments, but without reports of experimental results (articles corresponding to the “meta” literature as defined by Saracevic, 2004).

The screening process was based on three common descriptive elements, namely the article’s (a) title, (b) abstract, and (c) author keywords, a set of attributes that has been proven adequate for selection processes (Chang & Huang, 2012). During this stage, two domain experts worked in parallel to identify the articles that held DL evaluation interest. The experts were acquainted with DiLEO through training sessions to acquire knowledge about the content of the corpus and develop a high level of annotating ability. Any disputes between the experts were recorded and resolved by the intervention of a third expert who provided additional ratings. The third expert resolved disagreements that concerned 25.45% of the ECDL articles and 19.07% of the JCDL articles. However, if ambiguity about an article’s evaluation nature persisted, then the three researchers discussed the case to reach a consensus. Finally, 98 articles out of 624 were identified for the JCDL conference (a number representing 15.7% of all articles) and 128 out of 440 for ECDL (a percentage of 29.09).

The full text of the selected articles was then semantically annotated manually by the two experts in order to identify instances of subclasses of five DiLEO classes, namely Goals (G), Dimensions (D), Activity (A), Means (M), and Instruments (I). To achieve an optimum annotation consistency and assure the interannotator agreement at an acceptable level, the ontology developers set precise and clear guidelines following the approach of Roberts et al. (2007). The correctness of the annotation procedure was additionally assessed by random crosschecks, performed by a third expert and any disagreements were either moderated by him, or resolved through discussion.

The result of this process was the formation of a semantic network with nodes being instances of the ontology subclasses and edges the correlating DiLEO properties. In total, 4,248 triples of ontology instances with the form (domain C/sbc)—property—(range C/sbc) were generated. According to the notation syntax of DiLEO instances, a path of four instances is expressed as follows:

\[
\text{G/docm : } \{ \text{jcdl364 } _\text{Document}\} \text{—isAimingAt—} \\
\text{D/prfm : } \{ \text{jcdl364 } _\text{PerformanceMeasurement}\} \text{—} \\
\text{hasConstituent — A/rcrd : } \{ \text{jcdl364 } _\text{Record}\} \text{—} \\
\text{isPerformedIn— M/lbst : } \{ \text{jcdl364 } _\text{LaboratoryStudies}\} \\
\text{—isUsing— I/resr : } \{ \text{jcdl364 } _\text{TestCollection}\}
\]

This path refers to the article identified as AP/std: {jcdl364}, that corresponds to a study by Medelyan and Witten (2006).

The data set for the current study can be found at http://www.ionio.gr/labds/dbis/addendum.html

In the remainder of the article, the notation C/sbc is used, where C stands for a class label and sbc stands for a subclass.
The path expresses that the goal (G) of the study was to document (G/docm) actions taken to improve performance measurement (D/prfm). These actions included a recording activity, such as the subclass record (A/rcrd), which took place in a controlled setting (a laboratory study, M/lbst), using research artifacts (I/resr) in the form of a test collection.

Results

The analysis aims to investigate what and how researchers evaluate, which corresponds respectively to two DiLEO paths: (a) G—isAimingAt—D—hasConstituent—A and (b) A—isPerformedIn—M—isUsing—I. These paths are essential to reflect the orientation and the strategies of the evaluation research, as well as the practices used to realize them.

The 27 subclasses of these classes with their in-between properties generate a weighted graph, whose nodes correspond to ontology subclasses and its edges correspond to their respective properties. The edge weights correspond to the property frequencies, while the node weight is equal to its betweenness centrality ($C_b$). Although the frequencies of ontology subclasses can provide a good estimation of overall trends, betweenness centrality was chosen as the principal metric for the remainder of the analysis. Betweenness centrality is defined as:

$$C_b(v) = \sum_{s\neq t\neq v} \frac{\sigma_s(t)}{\sigma_{st}}$$

where $\sigma_s$ is the total number of shortest paths from node $s$ to node $t$ and $\sigma_s(v)$ is the number of paths that pass through the node $v$. Betweenness centrality is often used to determine the significance of concepts in sequential or networked association schemas in challenging settings, such as the investigation of the dynamic nature of information behavior tactics, strategies, and interactions (Xie & Joo, 2010). In Figures 2 and 3, the size of the nodes varies according to their betweenness centrality values, while their shape relates to their class. The edge color varies according to its weight (from maroon denoting high values to grey denoting low values). To enhance graph readability only edges with values greater than 2.5 are visible.

Significance Concentration

Evaluation Strategies

The graphs in Figure 2 correspond to the path G—isAimingAt—D—hasConstituent—A and aim to reflect the
adopted strategies in the articles of the JCDL and ECDL conferences. In the case of JCDL, the prioritized goals were to assist design (G/dsgn, \( C_B = 1.402 \)) and to describe the DLs status (G/dscr, \( C_B = 1.402 \)), while the purpose of document attracts less interest (\( C_B = 0.775 \)). In ECDL, however, it was discovered that all the three subclasses share the same centrality degree (\( C_B = 1.195 \)). Regarding Dimensions, performance measurement (D/prfm, \( C_B = 19.234 \)) was favored in the environment of JCDL, while technical excellence (D/tecx, \( C_B = 15.186 \)) held exactly the same centrality value with effectiveness (D/efct, \( C_B = 21.541 \)). In ECDL though, technical excellence had significantly lower value (D/tecx, \( C_B = 13.796 \)) than effectiveness (D/efct, \( C_B = 21.541 \)), which was the dimension in the most nodal position. Furthermore, in ECDL studies the betweenness centrality value of performance measurement (D/prfm, \( C_B = 13.796 \)) is equal to service quality. On the contrary, in JCDL service quality (\( C_B = 11.045 \)) seems to attract less interest than effectiveness, performance measurement, and technical excellence, with outcomes assessment holding the least nodal position in the network (D/outc, \( C_B = 9.721 \)). Similarly, outcomes assessment is an evaluation dimension of reduced interest in ECDL (D/outc, \( C_B = 8.361 \)).

Regarding evaluation activities, scholars publishing in JCDL mainly measure (A/mesr, \( C_B = 30.554 \)) or record (A/rcrd, \( C_B = 30.554 \)) and secondarily analyze (A/anlz, \( C_B = 17.750 \)). As the weights of their in between edges indicate, all three activities are associated with performance measurement, effectiveness, and the pursuit of technical excellence of DL components and services. In the ECDL community, researchers follow a similar approach; they rank first the record (\( C_B = 30.505 \)), second the measure (\( C_B = 22.404 \)), and third they analyze (\( C_B = 18.404 \)) activity. These activities are correlated with the same Dimensions in both conferences, namely performance measurement, effectiveness, and technical excellence. Researchers from both venues behave again similarly in the case of the fourth rank activity, compare (A/comp, \( C_B: \text{ECDL} = 17.612, \text{JCDL} = 17.368 \)), which is mainly associated with the effectiveness dimension.

**Evaluation Practice**

Figure 3 presents the graphs for the path A—\( \text{isPerformedIn} \)—M—\( \text{isUsing} \)—I, which reflects research practices in the two venues.

According to the graph, the most popular means that scholars in JCDL employ are logging (M/lgst, \( C_B = 9.984 \)) and laboratory studies (M/lbst, \( C_B = 9.572 \)) along with survey studies (M/svst, \( C_B = 9.572 \)). These three means are followed by comparison (M/cmst, \( C_B: \text{JCDL} = 7.289, \text{ECDL} = 5.303 \)) and expert studies (M/exst, \( C_B = 3.284 \)). Laboratory (M/lbst, \( C_B = 11.534 \)) and survey studies (M/svst, \( C_B = 11.534 \)) are similarly considered important means in ECDL as well, whereas field studies (M/lfst, \( C_B = 9.134 \)) hold a more central position than in JCDL. Surprisingly enough, logging studies (M/lgst, \( C_B = 7.149 \)) have
TABLE 1. Means co–occurrence (upper half for JCDL, lower half for ECDL).

<table>
<thead>
<tr>
<th></th>
<th>Lbst</th>
<th>Lgst</th>
<th>Extst</th>
<th>Cmst</th>
<th>Svst</th>
<th>Flst</th>
</tr>
</thead>
<tbody>
<tr>
<td>JCDL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lbst</td>
<td>–</td>
<td>11</td>
<td>5</td>
<td>19</td>
<td>24</td>
<td>3</td>
</tr>
<tr>
<td>Lgst</td>
<td>25</td>
<td>–</td>
<td>1</td>
<td>4</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>Extst</td>
<td>9</td>
<td>3</td>
<td>–</td>
<td>2</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Cmst</td>
<td>25</td>
<td>11</td>
<td>5</td>
<td>–</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>Svst</td>
<td>24</td>
<td>12</td>
<td>11</td>
<td>11</td>
<td>–</td>
<td>10</td>
</tr>
<tr>
<td>Flst</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>–</td>
</tr>
<tr>
<td>ECDL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lbst</td>
<td>–</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>–</td>
</tr>
<tr>
<td>Lgst</td>
<td>25</td>
<td>–</td>
<td>11</td>
<td>5</td>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td>Extst</td>
<td>9</td>
<td>3</td>
<td>–</td>
<td>2</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Cmst</td>
<td>25</td>
<td>11</td>
<td>5</td>
<td>–</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>Svst</td>
<td>24</td>
<td>12</td>
<td>11</td>
<td>11</td>
<td>–</td>
<td>10</td>
</tr>
<tr>
<td>Flst</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>–</td>
</tr>
</tbody>
</table>

Note. JCDL = Joint Conference on Digital Libraries; ECDL = European Conference on Digital Libraries.

decreased attention in the frame of ECDL evaluation research, followed closely by comparison studies (M/cmst, C2 = 6.899). The instruments mainly used are research artifacts (I/rsar, C2: JCDL = 2.459, ECDL = 2.508), such as test collections or algorithms, and software (I/sftw, C2: JCDL = 1.834, ECDL = 1.376), followed by statistics (I/stat, C2: JCDL = 1.367, ECDL = 1.031) and narrative items (I.nrvi, C2: JCDL = 0.971, ECDL = 0.762).

The examination of the edge weights showed that there are strong links between the subclasses measure and laboratory studies (A/mesr—isPerformedIn—M/lbst). It was also identified that the paths starting from analyze to laboratory or survey studies (A/anlz—isPerformedIn—M/lbst or A/anlz—isPerformedIn—M/svst) are important in the ECDL setting, while a moderate link was traced for the activity compare with laboratory studies in JCDL (A/comp—isPerformedIn—M/lbst). Quite naturally survey studies are associated with narrative items (M/svst—isUsing—I/nrvi), which include questions or statements and help researchers to measure perceptions and/or preferences of the evaluation subjects, while the most preferable instruments in laboratory settings are software and statistics (M/lbst—isUsing—I/sftw and M/lbst—isUsing—I/stat respectively). The latter pattern was more evident for ECDL, while for JCDL the use of statistics in laboratory environments was found limited.

Triangulation patterns in both conferences were explored during second-stage analyses, based mainly on frequencies. Although survey and laboratory studies are considered central in the evaluation landscape, they are occasionally used in isolation: 73.08% of the ECDL articles that use survey studies are also using other means, whereas 57.82% of studies conduct laboratory studies with other means. For instance, laboratory studies in ECDL are combined with survey (mainly questionnaires and interviews), comparison, and logging studies. According to Table 1, most ECDL articles reporting laboratory studies, also report the co-occurrence of some kind of logging and comparison studies (25 articles in both cases), while 24 articles refer to the mixture of laboratory with survey studies.

A similar pattern was identified in JCDL conferences as well, in which 80.49% of survey studies and 61.54% of laboratory studies are combined with other means, respectively. More specifically, Table 1 shows that survey studies co-occur with comparison studies in 12 articles and with logging and field studies in 10 articles. Furthermore, in 24 articles laboratory studies co-occur with survey studies and in 19 articles with comparison studies. The co-occurrence of survey studies with laboratory studies could indicate that researchers validate or extend the findings from laboratory experiments by performing user surveys, or refine the results from surveys by conducting laboratory studies. The Means that mostly co-occurred with other ones in both ECDL and JCDL at a rate within the range of 71.43 to 100% are expert studies and comparison studies.

Furthermore, to reveal the type and orientation of the research, key instances of the survey studies subclass were further explored as they usually enclose new subclasses (more specific methods) thus enriching the ontology concepts. It was found that there is wide co-occurrence of the questionnaires’ and the interviews’ instances in both ECDL and JCDL conferences (in 13 and 8 articles, respectively). However, the co-occurrence of questionnaires with focus groups, as well as of focus groups with interviews is relatively limited. In ECDL, there were found only five articles reporting the concurrent use of focus groups and questionnaires and two articles reporting focus groups and interviews, while in JCDL there were found three articles reporting each of the above pairs.

Temporal Evolution of DL Evaluation

The second phase of analysis focused on the evolution of the evaluation processes for the period 2001–2011. Indicative issues that were investigated deal with how research interests evolved, as well as synchronization of the two conferences.

Temporal Characteristics of Key Classes

In this phase the temporal evolution of the centrality values of the Dimensions and Means classes were explored; both classes are considered key ones in the ontology, the former for the strategic and the latter for the procedural level. With regards to the Dimensions subclasses, it is shown that in both conferences research is oriented towards performance measurement and effectiveness issues (Figure 4). In ECDL, technical excellence also has significant centrality values to exhibit, most notably the years 2008–2009, which are the highest in the conference set. The same subclass, D/tecx, records significant centrality values in JCDL as well, especially in the years 2001, 2005, and 2011. The high centrality values of this subclass in combination with the high values of performance measurement and the decrease of the effectiveness values could verify the system-oriented character of the research in DL evaluation. This research orientation could also be inferred by the prevalence of performance measurement in cases that two dimensions coexist in a study; usually the pairs that are built are either
performance measurement and effectiveness, or performance measurement and technical excellence. However, this phenomenon altered in the last 3 years, when all three subclasses have a more balanced presence.

Figure 5 presents the temporal evolution of the Means’ subclasses. It can be observed that laboratory studies and survey studies have a stable appearance in both conferences. These seem to be the most preferred types of studies by the scholars, despite the fact that they almost never synchronize. When the centrality values are high for one conference, the values are lower for the other. It is interesting to note the stable presence of field studies in JCDL for the first half of the period. Although field studies have low or zero centrality values in ECDL, with the exception of 2009, in JCDL this kind of study is strongly represented in the first 5 years of the period with an impressive return in 2010. Logging studies also have a stable presence at moderate levels in ECDL, while in JCDL they are rather arrhythmic in their presence. In almost every case, the logging studies centrality values are lower than those of laboratory studies, a finding that was further investigated; it was found that in many cases log analysis was performed in laboratory settings. There was only one exception though, in JCDL in year 2002, where logging studies had higher centrality values than laboratory studies. Finally, although comparison studies have very low betweenness centrality values during that period, these kind of studies scored the highest centrality value in 2004 in JCDL.

Temporal Characteristics of Evaluation Practices

An evaluation study is a complex process and entails several instances of the DiLEO ontology concepts. To compare the complexity, a new measure, named Coverage $CO_i$, is introduced and is defined as the sum of the instances of the DiLEO classes (and their subsumed subclasses) appeared in a study. This is expressed as:

$$ CO_i = \sum_{c=1}^{C} n_{c,i} $$

where $n_{c,i}$ is the number of the instances of a class (or subclass) c ($c = 1, 2, \ldots, C$) in a article $i$ presented in year $y$ ($y = 2001, 2002, \ldots, 2011$). Given these definitions, the class frequency $cf_i$ of a study, representing the weight of a class c in a article $i$, is defined as follows:

$$ cf_i = \frac{n_{c,i}}{CO_i} $$

The average normalized coverage of a class c in the articles published in a period of time $Y$ ($CA_y$) is defined as:

$$ CA_y = \frac{\sum_{i=1}^{P_y} \sum_{c=1}^{C} cf_{i,c}}{\sum_{i=1}^{P_y}} $$

where $P_y$ is the number of articles in the year $y$. In the case that $Y = 1$, $CA_y$ provides the average class coverage in a year; then a time series is generated to examine the variations of coverage in a period of $Y$ years.

The $CA_y$ analysis concentrates again on Dimensions and Means. Figure 6a presents the values for the class
Dimensions, which range from 0.11 to 0.18 for ECDL and from 0.9 to 0.16 for JCDL. In general, ECDL provides more instances of the Dimensions subclasses and this results in the overall value of CA<sub>Dimensions</sub><sup>2001–2011</sup>, which is equal to 0.14 (SD = 0.022); for JCDL it is equal to 0.12 (SD = 0.024).

In the case of Means, the CA<sub>cy</sub> values vary between 0.14 and 0.23 for JCDL and between 0.12 and 0.21 for ECDL and are presented in Figure 6b. It can be noticed that the ECDL articles utilize more means in years 2004, 2005, and 2006. However, overall the CA<sub>Means</sub><sup>2001–2011</sup> for JCDL is equal to 0.18 (SD = 0.027), while for ECDL equals to 0.17 (SD = 0.029), denoting that evaluation studies are slightly richer in the context of JCDL.

Despite the emphasis on Dimensions and Means outlined above, it was found that the CA<sub>cy</sub> values for the Activity class are higher. Overall, CA<sub>Activity</sub><sup>2001–2011</sup> value is 0.33 for both ECDL and JCDL. While in ECDL, the values increase in the second half of the period, the JCDL values decrease. In particular, the CA<sub>Activity</sub><sup>2001–2006</sup> for JCDL equals to 0.36, and for ECDL equals to 0.32, while for the period 2007–2011 the CA<sub>Activity</sub><sup>2007–2011</sup> equals to 0.29 for JCDL and 0.35 for ECDL.
To provide a general view of the evolution of the number of instances, a third measure is introduced. It is the average coverage of the articles published over $Y$ years, and is defined as follows:

$$PAC_y = \frac{\sum_{i=1}^{Y} \sum_{t=1}^{P} CO_{yt}}{\sum_{y=1}^{Y} P_y}$$

For $Y = 1$, $PAC_y$ provides the average coverage in a year; a time series is then generated to examine the variations of articles’ coverage in $Y$ years.

The $PAC_y$ values for the period 2001–2011 are presented in Figure 7 and they range from 7.86 to 13.11 for ECDL and from 9 to 12.25 for JCDL. In general, studies in JCDL include more DiLEO subclasses than those in ECDL, with the exception of the years 2004 ($PAC_{2004} = 12.69$), 2005 ($PAC_{2005} = 12.69$), 2006 ($PAC_{2006} = 11.60$), and 2011 ($PAC_{2011} = 13.11$). On average, an evaluation article in JCDL needs 10.59 subclasses ($SD = 1.16$), whereas in ECDL 10.37 subclasses are needed ($SD = 1.83$), while an increasing trend of the average coverage is evident. Additionally, in JCDL, the $PAC_y$ value presents a steady increase during the last 5 years of the period ($PAC_{2001–2006} = 10.16, PAC_{2007–2011} = 11.11$). On the other hand, the articles in ECDL follow a rather uniform distribution, as $PAC_{2001–2006}$ equals to 10.34 and $PAC_{2007–2011}$ equals to 10.40.

**Discussion**

This analytical work explored research directions in the DL evaluation domain. It was found that effectiveness and performance measurement are main research interests in DL evaluation, followed by technical excellence. This finding confirms the system-centric focus of the evaluation campaigns, which was further validated by the centrality of the researchers’ goals. The high frequency of the paths (a) from design to performance measurement and then to report (G/dsgn—AimingAt—D/efct—hasConstituent—A/rprt), and (b) from document to effectiveness and then to report (G/dcmnt—AimingAt—D/efct—hasConstituent—A/rprt) supports the claim that scholars are interested in the documentation of actions taken towards the increase of their systems’ effectiveness and the capturing of performance aspects to advance their DLs’ design. A limited interest is exhibited in service quality and outcomes assessment (D/srvq and D/octa, respectively), which was capitalized in a relatively moderate value of service quality in ECDL.

In terms of Means employed, the prevalence of laboratory and survey studies signifies a trend towards quantitative aspects of experimentation. This was demonstrated by the analysis of the instances of survey studies, where a consistent preference for questionnaires, was revealed, followed by interviews and focus groups.

The significant centrality of the data gathering, either measure, or compare, and analysis activities indicate that they are considered the dominant stages in the synthesis of evaluation initiatives. Specifically, the measurement and analysis activities that are correlated with the laboratory or survey studies constitute a sound research pattern in both conferences. The record activity is also important in JCDL, correlated again with the aforementioned means, while the betweenness centrality value of the compare activity is almost of the same importance in both venues (CO: ECDL = 22.929, JCDL = 20.894). The path from effectiveness to compare (D/efct—hasConstituent—A/cmpr) corresponds to a reasonable research pattern when the evaluation of effectiveness passes through comparison and benchmarking. Additionally, the path from performance measurement to compare (D/prfm—hasConstituent—A/cmpr) is also frequent in both conferences. The low centrality values of the recommend subclass (A/recm) highlight the absence of suggestions at the end of the reporting process. This reveals a preference of the scholars of the domain to focus on the presentation of the data gathering and analysis stages, to demonstrate research prototypes and to exchange experimental data, instead of showing how they addressed issues risen from their experimental study.

**Instruments, research artifacts (Irsar),** which include instances of test collections, prototypes, and algorithms, are used widely in laboratory studies, followed by survey and comparison studies. This pattern is identical in both conferences and it was found to co-occur with the use of statistics (I/stats) as a tool. Software (I/sftw) was also widely used in laboratories, followed by applications in survey studies for ECDL and comparison studies for JCDL, respectively. A focused look at the usage of narrative items (I/nrvi) showed that task statements and questions are the main instances, followed by statements and scenarios. While the two main narrative items are balanced in ECDL, there is a strong preference towards tasks in the JCDL conference. Finally, the frequencies of the two research artifacts, namely test collections and prototype systems, were balanced in both conferences.
The temporal analysis revealed that the practices of scholars in both conferences do not synchronize. Though a significant number of researchers publish in these venues and the general trends are similar, their practices differ. Moreover, the temporal analysis revealed that during the years 2004, 2005, and 2006, the evaluation initiatives presented in ECDL are richer in DiLEO instances than those in JCDL. In general, though, the JCDL evaluation studies incorporate more subclass instances in their structure than ECDL’s.

It must be noted that there is a difference in the number of selected and annotated articles between the two conferences. The selected JCDL articles cover 13.88% of the total number of articles presented in the conference, while the corresponding ECDL articles cover 29.5% of the articles in the conference history. This difference is justified by the exclusion of a significant number of JCDL articles, which while they reported research data, are concerned primarily with user studies instead of evaluation exercises. It might be the underlying cause of the asymmetry between the centrality values of the Mi/svst node in JCDL (37.219) and ECDL (11.534).

Concerning the limitations of the study, it can be noted that the annotation process is subject to the familiarization of the annotators with the DL evaluation domain. The annotators in this study were experienced with DiLEO conceptualization and acquainted with the literature. Although other studies, such as bibliometric ones, use more controlled elements, such as authors, citations, or subject/thesaurus headings, this study dealt with more sensitive data. As Janssens, Glänzel, and De Moor (2008) note “textual information can indeed indicate similarities that are not visible to bibliometric techniques” (p. 608) (where similarities in this case correspond to evaluation patterns) and it is argued that research practices can be revealed through annotated in-text information. Therefore, future work includes the evaluation, improvement, and further formalization of the annotation process. The implementation of automated semantic annotation tools is considered a challenge and part of this research is directed towards this end with the development of the respective ontology constructs that will enable less expert users to use the ontology effectively.

Conclusions

In this study, an ontology was used to annotate the DL evaluation literature published in two conferences of the domain and to present the conceptual associations between descriptions of the evaluation initiatives. The analysis of the ontology instances offered an insight into the logic and mentality that governed the design of evaluation experiments for the period 2001–2011. The DiLEO ontology provided a ground for defining critical evaluation constructs, representing evaluation initiatives, as well as for sharing and reusing their findings. The main advantage of DiLEO is that it can provide a common understanding of the DL evaluation domain, permitting the exploitation of analysis and inference mechanisms to derive new knowledge.

In our analysis we found particular strategic-level and procedural-level patterns, which sketch the crucial decisions for the design of evaluation initiatives. The research in both conferences focused mainly on performance measurement, effectiveness, and technical excellence. The most preferred methods for conducting evaluation experiments are inside laboratories and through surveys. Our findings also highlight the complexity of the evaluation initiatives, especially in terms of practice.

References


