An Ontological Representation of the Digital Library Evaluation Domain

Giannis Tsakonas and Christos Papatheodorou

Database and Information Systems Research Group, Laboratory on Digital Libraries and Electronic Publishing, Department of Archives and Library Sciences, Ionian University, 72 Ioanni Theotoki Str., 49100, Corfu, Greece. E-mail: {gtsak, papatheodor}@ionio.gr

Published version appearing at:

Abstract

Digital library evaluation is a complex field, as complex as the phenomena it studies. The interest of the digital library society still remains vibrant after all these years of solidification, as these systems have entered real-life applications. However, the community has still to reach a consensus on what evaluation is and how it can effectively be planned. In the present article, an ontology of the digital library evaluation domain, named DiLEO, is proposed, aiming to reveal explicitly the main concepts of this domain and their correlations, and it tries to combine creatively and integrate several scientific paradigms, approaches, methods, techniques, and tools. This article demonstrates the added value features of the ontology, which are the support of comparative studies between different evaluation initiatives and the assistance in effective digital library evaluation planning.

Introduction

Although there has been more than 15 years of intense research regarding the solidification of the domain of digital libraries (DLs), these systems continue to be challenging in terms of modeling. Many reasons are inferred, such as the interdisciplinarity that governs the domain, the dependency on technologies, and the strong association with the, wide or narrow, social context. Indeed, DLs are complex sociotechnical systems and their operation is subject to many interactions, either internal—in the sub-system level—or external—in the organization or the general environment level. DL evaluation constitutes a significant part of the domain, not only for the obvious reasons, such as the contribution on critical information systems development parameters such as accountancy or documentation, but also for the provision of a discourse field for the understanding of their nature. As with anything else revolving around the globe of DLs, evaluation is a multicultural construct that has provided multiple and extensively varied forms of the DL operation.

To study the outcomes and the functions of given systems, researchers often deconstruct DL models, either empirically or cognitively, and propose conceptualizations that are comprehensible, possibly as a response to a reasoning challenge. These processes of deconstruction and consecutive reformation have a rich nomenclature, such as formal, conceptual, or reference models or frameworks. These are mainly expressed by means of finding the principle entities, identifying the relationships developed among them, and, finally, tracing and assessing the produced outcome. This is a process that has multiple initiation points, such as a view from the object’s side or an effect of the outcome in the environment. The ultimate goal however remains to remove any technical dependency or system bias to forward an abstraction of a system’s architecture, processes and connecting bond and to provide to all collaborating parties the means of semantic convergence.

In this article, an ontological representation of the DL evaluation domain, named DiLEO, is proposed. The development of DiLEO is motivated by the need to address one important requirement of the DL evaluation community as raised by Führ et al. (2007), namely, the community building in
evaluation research. According to Führ et al.: “The lack of globally accepted abstract evaluation models and methodologies can be counterbalanced by collecting, publishing and analyzing current research activities” (p. 35). This requirement cannot be satisfied without a conceptual model that contains a certain formality that can meaningfully correlate evaluation data and cover the wide spectrum of the respective initiatives. As a response to this, the purpose of the development of this ontology is to create a schema for collecting and analyzing existing research activities in a formal way and to transform this information into a reusable corpus of knowledge. This knowledge can then strengthen the decision-making concerning evaluation planning by fostering reasoning paths, as well as the benchmarking of different evaluation initiatives. By modeling and representing formally DL evaluation activities, one can reason about the flow of processes and the respective requirements and constraints. Moreover, one can use the ontology to organize and associate real life instances of the evaluation concepts into a knowledge base, which can be queried to retrieve subgraphs of activities providing useful insights about existing studies.

Related Work

Digital Library Modeling

Because of the constant DL transformations and developments, one of the most challenging tasks in this field is to find a comprehensive way of modeling the main constituent parts and the relationships between them. Studies using classifications for the description of DLs are useful to reduce ambiguity and to structure more specific perceptions. For example, Markscheffel, Fischer, and Stelzer (2008) provide a classification of DLs according to the business models they employ. Similarly, Sandusky (2002) has organized the attributes of DLs into six categories, namely, audience, institution, access, content, services, design, and development. His effort focuses on the creation of a comparison tool that would reveal advantageous design parameters of DLs, such as usability. However, these faceted classification approaches have limited exploratory strength, because they focus on particular parameters of phenomenon under investigation. Furthermore, they lack inter-class associations and a complete enumeration of their properties, mainly because of their informality.

The terms “formal models,” “conceptual models,” “reference models,” and “frameworks” have been interchangeably used to describe initiatives of modeling the DL space. Through logics-based or any other formalism, these propositions of abstract views of an operation are not used as a means of comparison to the real world, but rather as a means of explicit description of the main concepts of the field and the relationships and constrains that exist among them in a given context. For instance, the conceptual model of the Joint Working Group between DELOS and NSF for the Digital Imagery for Significant Cultural and Historical Materials (Chen, Wactlar, Wang, & Kiernan, 2005) defines three interacting entities, i.e., the user, the technologies, and the content (cultural in their case), that communicate together in several stages from the creation and preservation of content to its usable provision and its retrieval through technologies.

One of the most comprehensive models of DLs is the 5S model, proposed by Gonçalves, Fox, Watson, and Kipp (2004). 5S encompasses the process of data streaming inside spaces and structures to serve the scenarios of use of a given society. The classification of recorded concepts and practices into self-excluded and compact categories, such as active agents, processes, activities, constituent parts, socioeconomic, legal contexts, and environment, is what this approach accomplishes. Therefore, the 5S model exploits the terminological basis of the DL domain, sorts the mined concepts, and proposes a set of formal statements to declaratively express instances of DLs.

Another kind of modeling is the development of reference models, which aim to describe a system or a phenomenon free from the conditions imposed by the environment. Thus, DL personnel can forward specific suggestions of development and application and can focus on specific parts of a
hypersystem, such as human agents, to understand the networking and the interaction of the stems of the DL (Borbinha et al., 2005). The DELOS Reference Model (Candela et al., 2006), bearing the title “Manifesto,” considers the content, the functionalities, architecture, quality, and the policy as integral components of DLs. These components do not belong in one single abstraction of a DL, but rather correspond to different conceptual constructs, such as the digital library, the digital library system, and the digital library management system. This distinction serves the conceptual convergence between the variant approaches, which all originate from the different communities’ background and the advancement of common and coordinated action. In its latest version, Digital Library Reference Model (RM; Athanasopoulos et al., 2010), it portrays with the help of topic maps—the relationship between six main concepts, namely, DL resource, user, content, functionality, policy, and quality. The concept of quality represents all activities that have “the goal of judging and evaluating them (note: meaning the DL systems) with respect to specific facets” (p. 50). These facets reflect the main concepts of the RM, achieving an overview of the DL system through categorized sets of criteria.

Zachman’s Framework aims to describe enterprise architectures by classifying and organizing on a two-dimensional plane the concepts needed to homogenize the variant perspectives in the enterprise design phases, first by actor (naming alternative views) and then by function (posing questions). This model has been used in the DL domain for many reasons, for instance, by Borbinha (2007) in a proposal to revisit the goals of DL and to suggest a movement of the field from research to practice.

Digital Library Evaluation Modeling

In this article, DL evaluation is considered to be the purposeful procedure of estimating the value of a process or a product for the benefit of a given unit through formal and/or informal means and processes. This definition is used to support a modeling process, which requires a synthesis of perspectives to encompass user-centered, system-centered, and content-centered expressions of evaluation. This generic approach is used to abstractly cover the constructs of DL evaluation, as well as the most important modeling approaches. From the literature, it is concluded that understanding can be achieved after either an object-oriented process, for instance, by defining the main agents/constructs/elements in the DL domain, or a workflow-oriented process, by describing the stages for performing and the requirements needed to be satisfied in each of them. This section summarizes some important modeling efforts that have been recorded. These efforts share the common aim of providing a clear view of the domain and assisting decision making in future evaluation campaigns.

Kovács and Micsik (2004) created a theoretical device, the Evaluation Computer, to depict evaluation instances. The descriptive advantage was based on a faceted classification of aspects, which allowed a gradual construction of an evaluation instance, called an “evaluation atom.” This conceptual construction allowed the selection of attributes from five respective levels, namely, system, content, organization, user, and evaluation, thus producing multiple views of the evaluation activity. Apart from the flexibility it offers to view the same instance from multiple points, it provided an additional advantage—that of the measurement of distances between two different activities within a vector space, which is a useful process for a comparison on a meta-evaluation level.

Two of the above-mentioned modeling schemes, 5S and DELOS RM, have been widely articulated and have been worked together to synthesize a new reference model regarding the concept of quality, which is associated to the evaluation of interaction in the various stages of information processing (search, disseminate, etc.; Agosti, Ferro, Fox, Gonçalves, & Lagoeiro, 2007). The Zachman framework has been also used to assist evaluation activities. In an exhibition of the exploitation of the actors’ dimension, Abdullah and Zainab (2008) have used the framework as a tool to extract activity requirements of the owner, planner, and designer agent, while Xu (2009) has taken advantage of the framework by finding functionalities that an evaluation tool for DL planners should support.
Saracevic (2004), in an attempt to classify evaluation studies in the DL domain, followed a procedural approach by identifying four main classes of entities and posing several questions to fundamentally explore their nature and properties. The classes of constructs, context, criteria, and methodology give a substantial reflection of the DL evaluation domain by adding definitions that are important aids for designing an evaluation activity. Saracevic’s classification identified the most critical dimensions of evaluation and the factors that influence it and distinguished those studies that model evaluation processes and those that report findings. Blandford et al. (2008) presented the PRET A Rapporter framework for evaluating DLs from the HCI aspects of their operation. The framework was intended to be used for designing evaluation activities but, at the same time, provided an opportunity to model DL evaluation instances that were compared together. The PRET A Rapporter framework is seeking answers to critical questions, such as the purpose of the evaluation, the available resources and apparent constraints, the techniques for collecting, analyzing, and reporting findings, and the ethical parameters of the study. Extending the work on 5S Gonçalves, Moreira, Fox, and Watson (2007) have proposed a set of measurement units for assessing quality in DLs. They proposed a minimal set of DL constructs that first corresponded to the 5S concepts and then were associated to dimensions of quality, such as accessibility, accuracy, efficiency, effectiveness, reliability, completeness, preservability, and relevance. These dimensions were then translated to criteria and measurement units. Gonçalves et al. exhibited the validity of their proposition by providing examples from operating DLs in each respective construct, as well as for conducting a qualitative research study which involved information professionals.

A model of DL operation with the main agents, the interactions between them, and the contributing communities was presented by Führ, Hansen, Mabe, Micsik, and Solvberg (2001). Based on the three agents, i.e., users, technology, and data/content, the proposed framework addressed several questions, such as who, why, and how is evaluating. The model was used to gather insights on DL evaluation from researchers as well as to establish a benchmark for comparison purposes. One of the findings by Führ et al. (2001) was the validation of the multidisciplinary nature of the field, through the employment of diverse terminology and different conceptualizations.

Ontologies and Digital Library Evaluation Modeling

The term ontology often refers to the philosophical discourse about the categorical schemata of entities (Lowe, 2005). However, in computer and information science it is considered to refer to a systematic treatment of objects (concepts, relationships, constrains etc.) to represent a specific domain and to facilitate machine reasoning. The definition of the term, by the W3C Web Ontology Working Group (Heflin, 2004), is as follows: “An ontology defines the terms used to describe and represent an area of knowledge.” Gruber (1993) has provided a more typical definition, mentioning that ontology is “a formal, explicit specification of a shared conceptualization,” or else it is the typical representation of a knowledge domain’s primitives, such as classes and relationships, that facilitates the sharing of information among the members of community. Guarino (1998, p. 7) further refined this statement by saying that “an ontology is a logical theory accounting for the intended meaning of a formal vocabulary, i.e., its ontological commitment to a particular conceptualization of the world” (emphasis on the original).

Ontologies, encoded by languages like OWL, have been developed to model the DL domain. For instance, Kovács and Micsik (2005) have proposed a generic ontology model using a layered approach that defines four important DL components, such as content, services, interfaces, and community. An extension of this ontological representation was presented by Mitrelis, Antonakis, Siochos, and Papoutsis (2006), who integrated the conceptual classification of the 5S in pursuit of delivering a higher degree of expressiveness. Following the work on 5S, Gonçalves, Fox, and Watson (2008) have proposed a formal ontology to develop a theoretic approach to DLs. With their ontology, they extend
the formalization of workflows inside DLs, provide instances of implementation on the DL architectural field, and verify the intrinsic properties of the ontology by applying it to a conceptual variation of the DL quality. Recently, Ohren (2009) translated the conceptual modeling of the DELOS RM to OWL and created the respective vocabulary (DLRM Vocabulary).

FIG. 1. An overview of the ontology.

With this vocabulary, which followed the principles of DCMI Abstract Model, Ohren provided examples of DL functions and profiled services.

Ontology Development

In this section, the primitives of DiLEO are presented analytically. The ontology comprises two main layers: the upper or strategic includes the main concepts that define the purpose of the evaluation. Furthermore, the strategic layer includes classes that place the evaluation in the domain space, i.e., they denote its scope and relate it to other evaluation studies. The lower layer, the procedural layer, comprises classes that model the practice of an evaluation initiative. These classes specify an evaluation activity in terms of exact processes, constraints, and requirements. Figure 1 presents the two layers and the main classes that are included in each of them.

The ontology was developed through an extended literature review, which included studies and surveys from multiple venues, such as journals, conferences, and workshops. According to Saracevic (2004) the DL evaluation literature can be divided in two main categories: the first one is the meta or
“about” literature that presents conceptualizations of evaluation processes and the second one is the object or “on” literature that presents realistic evaluation activities. Thus, our empirical approach is mixed: taking a top-down thread, which traced important concepts in studies that discussed evaluation models, while simultaneously a bottom-up approach allowed the findings of particular evaluation cases to contribute to the solidification of the ontology’s vocabulary and structure. The description of the ontology follows the principal structure of an ontology, which means that firstly the classes and subclasses are outlined and then their relationships are specified.

Entities of the Strategic Layer

This section presents the main classes⁠¹ that reflect the strategic plan of the evaluation, which can be considered as a set of decisions about the purposive view of an evaluation and its aims.

A central concept of the ontology is the evaluation dimensions (D), which reflects the scope of evaluation activities, as well as designating its purpose and outcomes. Also through dimensions, evaluators purposefully attempt to observe a certain aspect of a phenomenon that takes place in time and space and concerns particular objects. This upper layer concept encapsulates the key dimensions of DLs operation studied by evaluation initiatives, namely, effectiveness (efct), performance measurement (prfm), service quality (serq), outcomes assessment (outc), and technical excellence (txc). Many of these dimensions have their origins in the library domain and the interaction between physical libraries, electronic resources suppliers, and DL developers. In a review of the major challenges in the provision of electronic services by physical libraries, Bertot (2004) considers five kinds of assessment. More specifically, he recognizes (a) output assessment, (b) service quality, (c) outcomes assessment, (d) performance measurement, and (e) other, as the kinds of assessment that one can perform on the outputs of a library service. Each one of these concepts represents a space of diverse theories and practices.

- **Effectiveness**: Effectiveness is the field that studies the outcomes of a process regarding the inputs. This type of evaluation provides a description of the state of a system altered after feeding it with a certain amount of resources, portraying thus aspects of cost-effectiveness. It can be considered one of the first stages of an evaluation, as the hermeneutics it can provide are quite limited.

- **Performance measurement**: Performance measurement indicators are recording data for the efficiency of an organization or a system during its operation. In the organization context, the calculation of costs required to achieve the outputs and the estimation of usage ratios in the serving community are typical performance measurement processes. In information retrieval (IR), performance measurement is associated with the precision and recall estimation of the search mechanisms and, according to Voorhees (2002, p. 355), is “an abstraction of the retrieval process that equates good performance with good document rankings.” Despite the obvious systemic differences, performance indicators have as common elements the pursuit of effective distribution and the efficient consumption of resources.

- **Service quality**: Service quality explores the effective delivery of services and deployment of functionalities, seeking evidence of effectiveness through the eyes of the “recipients.” A typical instance of this class is the DigiQUAL protocol (Kyrillidou & Giersch, 2005) that assesses the quality of services of DLs in several areas, such as navigation, interoperability, accessibility, and sustainability. However, the concept of quality is central in the evaluation literature and there are different interpretations that affect other integral components of DLs. For instance, Moreira, Gonçalves, Laender, and Fox (2007) propose the 5SQual tool to conduct evaluations that compare the quality of certain criteria. The tool is based on the formalized association of DL constructs with evaluation criteria. While DigiQUAL is associated with services, 5SQual focuses primarily on the

---

¹ The classes of the ontology are abbreviated by capital characters, while their subclasses by lowercase ones.
quality aspects of content (digital objects and metadata) and services. • Outcomes assessment: Outcomes assessment seeks to find traces of impact of the DL operation on a long-term basis. According to Bertot and McClure (2003, p. 596), outcomes assessment research focuses “on the extent to which a library’s services and/or resources made a difference in the life of the library’s individual, group, or institutional users” and this effect spreads to several aspects, such as economic, learning, and research.

• Technical excellence: The technical excellence dimension refers to the improvement of the DL design. The evaluation initiatives of this dimension target to the advancement of the performance of DLs, the adjustment to specific requirements of the serving audience, the enhancement of their behavior, and the upgrade of their interaction with the system, e.g., improving their own performance and effectiveness, and therefore ensuring the results are stable, reliable, useful, and usable.

Dimensions have different types (DT) and are related to the phases during which any evaluations take place. An evaluation can be formative (fmtv) or summative (smtv), while both these types can be also iterative (itrv). An evaluation is formative when it is performed synchronously to the development of a DL and its prime aim is to correct design errors or omissions before the latter’s actual use. For instance, Hill et al. (2000) present the process of integrating user requirements in the design of Alexandria Digital Library as extracted by a multifaceted formative evaluation process. Summative evaluation, on the other hand, focuses on a definite product, which may be a prototype version or the full working DL. It addresses how the users accept it and what flaws are observed in its operation. Theng, Chan, Khoo, and Buddharaju (2005) used both qualitative and quantitative means to evaluate the adoption of eLibraryHub, the DL of the National Library of Singapore. Their evaluation employed questionnaire surveys to collect opinions regarding the user effectiveness and satisfaction, as well as Claims Analysis techniques to identify problems with the interface design of fully working DL. A common feature in both types, each one from its respective stage, is to be iterative, avoiding thus the fragmentary hue of evaluation and the generation of disjointed results (Bertot, Snead, Jaeger, & McClure, 2006). Dimensions and their types link together through demand to proactively monitor and evaluate the development and performance of DL. Evaluation design tools, such as logic models (Khoo & Giersch, 2009), have proven useful instruments, in either formative or summative stages, addressing questions related to the scope of evaluation, such as the quality of short-term and long-term outcomes.

**Research Questions**

The dimensions outlined above are usually decomposed to some research questions (RQ), a class which comprises statements that guide the research design and frame the evaluation process. The Research Questions often reflect the Criteria (C) that are applied and are addressed by the Findings (FN) of the study (both classes belong to the procedural layer and are outlined below).

**Goals**

Any of the evaluation dimensions aims utterly at the improvement of a service or a product. Goals (G) are the, explicit or implicit, reasons to carry out an evaluation activity and can be listed in a nonexclusive and a supplemental manner. These goals concern the description (dscr) of the state in a given time fracture, the documentation (docm) of the actions performed and the optimization of design (desg).

• Description is not an evaluation goal per se, but it is considered a stage of it. It serves this purpose and to describe the state of a system or a class of users, one follows principal elements of scientific research. For instance, Sumner, Giersch, and Jones (2003) conducted a questionnaire survey to identify strengths and weaknesses in eleven DL projects, while McMartin et al. (2008) used a survey
to find the reasons for use and misuse of educational DLs in U.S. institutions. Several user studies can be ranked into this category as they report on the characteristics and behaviors of certain class of users, while reviews of DL systems perform the same task, but also forward the critical views by conducting comparisons. An instance of the former is the study of Ke, Kwakkelaar, Tai, and Chen (2002), who performed a transaction analysis log study to describe the state of usage habits of Taiwanese researchers, while the latter is expressed through studies like Open Society Institute’s (Crow, 2004).

- Documentation is the process during which an agent evaluates part or whole of the operations/products to prove the validity of the choices and it is heavily related to the notion of accountability. Reeves et al. (2003, p. 2) argue that evaluation serves justified decision making and that “decisions should be informed by timely, accurate information.” Franklin, Plum, and Kyrillidou (2009, p. 32) mention that usage metrics can lead to “informed collection development decisions and to better justify collections management choices,” while Barton (2004, p. 140) has argued on the necessity of “measurement-based management” by exhibiting the effect of several UK projects.

- The optimization of design is undoubtedly related to technical aspects of the DL development or provision. It needs to be complemented by the above-mentioned, such as description and documentation, but the main requirement it addresses is the creation of new knowledge, an organizational learning effect (to avoid further mistakes) and critical omissions. Zabed Ahmend, McKnight, and Oppenheim (2006) applied a range of methodologies to iteratively evaluate an IR interface and to result in a set of recommendations for the design, which were grounded on valid empirical findings, while Plaisant Marchionini, Brun, Komlodi, and Campbell (1997) followed a similar approach to refine the design of the Library of Congress National Digital Library and to uncover problems.

Levels

The evaluation processes and their dimensions affect the operation of a DL. According to Saracevic’s approach of stratified contexts (2000), there are seven levels (L) of the DLs operation that can be affected by evaluation findings and benefited by design interventions. They range from the lower levels of content (cntl), engineering (engl) and processing (prcl), to the mid level of interface (infl) and the upper ones of individual (indl), institutional (insl), and social (socl) levels. For instance, the work of Jaballah, Cunningham, and Witten (2005) on the assessment of user performance in personalized features of DLs refers to the individual level, while the work of Park (2000) concentrates on the role of interfaces for the enhancement of user interaction and Borgman (2002) presents the role of ADEPT DL for geography-related education at the user communities level.

Evaluation Objects

As has been declared in several definitions, the Objects (O) of an evaluation can be a product (prdc) or an operation (oprt). This means that one can evaluate a full working system, a prototype, a new service, a new collection of items, or can measure the speed of response of a system, as well as the stability in transactions through machine interfaces, the number of downloaded items, performance of users in information searching activities, and so on. In general, DLs consist of content and system (product), which can be used (operation) by some agent. Content has data (dat, textual, visual, audio) and metadata (mdt, statically or dynamically created), while system has technologies (tec) that define the power and the outreach of the DL, functionalities (fnc) for the delivery of services, such as searching, annotating, sharing, and interfaces (inf) for the facilitation of communication of the system parts with human or mechanic agents. Furthermore, these products can be operated, which is to say one can use the system (uos) or use the content (uoc) of a DL.
Subjects in Evaluation Procedures

Objects are operated by Subjects (S), human (hmn), or machine agents (mcn), which take part implicitly or explicitly in their evaluation. These entities participate in the evaluation activity as essential constituents of the operation or the product and are indirectly assessed, as every kind of recording seeks to reflect potential system or content improvements.

Characteristics

Subjects and objects have their own Characteristics (CH and hmnc, mcnc respectively for each subclass), which refer to typological variations. For instance, in each evaluation, agents are differentiated by their age (hmnca), experience (hmnce), profession (hmncp), or discipline (hmncd) for samples of human agents and by their role (mcncr) for machine agents, while a common characteristic of both classes is their count (hmnc and mcnc). Tenopir (2003) summarizes the findings of a wide range of usage and user studies regarding electronic information, in which the diversity of human agents, such as variations in disciplines is outlined. Objects can be described by their size (ocz), the type of items (oct) they hold, and their number (occ) and kind (ock) of services. Westell (2006) presents a list of evaluation measures for institutional repositories, which incorporates such typological features, while the Repository Support Project gives a comprehensive matrix of digital repository system functionalities and services (www.rsp.ac.uk/software/surveyresults).

Entities of the Procedural Layer

This set of classes outlines the main features of the process of evaluating a DL. The classes define the requirements one has to satisfy to perform an evaluation initiative.

Activity

Activity (A) is a central concept in the process of evaluating a DL. The term encapsulates all actions needed to be performed to process data and transform them into meaningful information that will improve the design and operation of a DL system. Activity is divided into several stages:

- Record (rcrd), measure (mesr): The first stage regards the capturing of the current image of product or the process by either a record act, such as video recording users or keeping logs of keystrokes, or a measure act, like counting new entries in the collection.
- Analyze (anlz), compare (comp), interpret (intp): The second stage includes the activities of data comparison, analysis, and interpretation in terms of comparing the data collected to the ideal image that the product should have, describing the data, tracing relationships between variables, and transforming the data into meaningful findings.
- Report (rprt), recommend (recm): By the way of a report, the evaluator presents in a clear and communicable way the major findings of the study and, hopefully, if the kind of evaluation permits it, recommends resolutions, asks for improvements, and frames future work. As a subclass of activity, it is carrying out the major findings of the evaluation study, which in turn are expected to address the research questions.

Factors

The performance of an activity is subject to several Factors (F) that influence the process of evaluating a DL and include Time (tim), Cost (cst), Infrastructure (ifr), and Personnel (prs). These factors are
coherently connected. As Doubleday et al. (1997) mention, there are quite striking differences in time, depending on the personnel that is employed in the evaluation. They report that to finalize a usability evaluation with the heuristic method 35 hours were needed, while inspecting users required 125 hours. Of course, expert evaluators can be characterized and described by their count (prsc), experience (prse), and skills (prss). Yu and Apps (2000), in their SuperJournal study, stated that log analysis is a low-cost method to describe the behavior of users of electronic information, with, however, the disadvantage of limited explanatory strength, while heuristic evaluation techniques have been characterized as “discount” or low cost (Blandford, Keith, Connell, & Edwards, 2004). Covey (2002) reports that the DLF members of the 2002 survey believe that developing evaluation skills is essential for conducting well-crafted evaluation studies. The HCI field has investigated in-depth the issue of evaluators’ performance with Hertzum, Jacobsen, and Molich (2002) introducing the term “the evaluator effect” to highlight the discrepancies in the evaluators’ results.

Means and Their Types

All evaluation activities are performed utilizing particular Means (M). Means is the enumerative set of techniques and methods with which one collects, analyzes, and reports findings in an evaluation activity. Among the means that evaluators have are logging studies (lgst, transaction log analysis, deep log analysis), laboratory studies (lbst, usability inspection, IR examinations, eye-tracking), expert studies (exst, cognitive walkthroughs, collection quality), comparison techniques (cmst, consistency inspection, checklists), field studies (flst, observation studies), and survey studies (svst, questionnaires, interviews, focus groups).

The means are distinguished in two different Types (MT), namely, qualitative (qltv) and quantitative (qntv). One can make a spontaneous hypothesis that the two types of means are differentiated only on the data they collect. Weiss (1997, p. 82) gives the norm by stating that the information in qualitative studies is “usually collected in case studies, ethnographies, or other non-standardized designs,” while data in their quantitative counterparts are collected “in experimental or quasi-experimental designs.” According to Gorman and Clayton (1997), quantitative evaluation is deductive, using theories and hypotheses to end—through a road of austere and firm methodology—in a conclusion, if not in a generalization and prediction. Contrary, qualitative evaluation collects evidence to theorize states and event arrays and is therefore inductive and synthesizes data to interpret the various phenomena.

Instruments

Finally, an evaluation activity is supported by Instruments (I), which are artifacts, cognitive constructs, or software that assist the specific processes. These instruments can be software (sftw), like statistical packages or quality/ethnography study analysis software, devices (devc), like microphones and cameras, narrative items (nrvi), like questions and statements, research artifacts (rsar), like prototypes, scales (scal), such as ordinal scales and nominal scales, as well as statistics (stat). Typical instances of this class include questionnaires, such as QUIS for solicitation of usability opinions (Fox et al., 1993), processing packages like SPSS for statistics analysis (Nicholas, Huntington, Jamali, & Tenopir, 2007), or ActivityLens for qualitative analysis (Papachristopoulou, Tsakonas, & Papatheodorou, 2008) and Likert scales for gathering rates of preferences and opinions (Jeng, 2005).

Criteria and Their Categories

Evaluators select formal or informal Criteria (C) to assess a perspective of the DL operation or a product. In the case of formal criteria, these can be standards (stnd), principles that ascertain a level of
quality, or toolkits (tlkt), sets of commonly agreed means and methods for benchmarking and comparison of certain objects with stable measurement units. For instance, in the case of usability, one can use a set of formal criteria as defined by the ISO 9241/11 (1997) or apply less formal, as well as standardized through wide acceptance, criteria of usability, such as the 10 heuristic principles (Nielsen, 1994). Another example is the employment of criteria via toolkits, such as the eVALUEd toolkit (Thebridge, 2004), that once again proposes various criteria for the evaluation of DLs and electronic resources within physical libraries. In the case of informal criteria, these refer to context-specific measurements (spca) to respond to local conditions and project idiosyncrasies that cannot be covered by generic-wise criteria. In practice, this is acceptable as long as it results to meaningful and usable findings, but it has also to be remarked that as a result of this “criteria for digital library evaluation fluctuate widely from effort to effort” (Saracevic, 2004, p. 18).

Criteria often are clustered in Criteria Categories (CC), which, as the name implies, classify evaluation criteria and principles. Several categories of criteria can be found in the literature that work as an umbrella, under which different propositions can be covered. In the proposal of Interaction Triptych Framework (Tsakonas & Papatheodorou, 2008), which is a user-centered framework, the criteria categories of usability, usefulness, and performance have been identified. An extended set of criteria has been proposed by Zhang (2010). Zhang has concluded on a model of DL evaluation based on a grouping of amalgamated criteria. This grouping reflects six different levels, namely, technology, content, user, context, interface, and service, which portray properties of these levels as constructs of measurement. Xie (2008) has also studied the criteria used in DL evaluation, under a user-centered perspective, identifying five categories of metrics. The criteria were interface usability, service quality, collection quality, system performance efficiency, and user feedback solicitation. Both research initiatives group criteria according to some DL constructs or processes inside the DL and the perspective they have on the topic differentiates them. These categories can be assessed on the basis of comprehensiveness and applicability.

**Metrics**

To measure or record criteria, one has to define or adopt metrics (MTR). Metrics are the measurement units that are needed to establish a distance between the two states, the expressed state, and the ideal state. Metrics can be user originated (usro), content originated (cnto), or system originated (stmo). While in the later two cases the nature is obviously numerical (see Nicholas, Huntington, & Watkinson, 2005 for content originated in the field of content usage metrics or McClure, Lankes, Gross, and Choltco-Devlin (2002) for system originated in the field of digital reference metrics), in the user originated case of metrics there is a distinction between perception metrics (prcm) and report metrics (rptm, i.e., opinions, feelings). Theng, Yin, Binte, Ismail, and Ahmad (2007) gathered the perception of users regarding the validity of heuristic principles application on the design of NLBDL, while Davis and Connolly (2007) gathered the opinions of the faculty members of Cornell University on self-archiving practices in DSpace.

**Findings**

Study Findings (FN) are the data that are generated or discovered as a result of the enquiring process. These Findings are often predefined to a great extend by the Research Questions of the evaluation, which means that some of the Findings are unanticipated and probably not predicted.
Entities—Labels and Names

The ontology includes also an Appellations (AP) class that provides names and identifiers to the evaluation studies and initiatives that constitute the instances (individuals) of this ontology. For each study, a record is generated that holds information for the study identifier (stid), study reference (strf), and object identifier (obid). The class of Appellations is not pictured in Figures 1 and 2 for economy reasons.

Relationships—Weaving the DL Evaluation Intelligence

The second important primitive feature of an ontology is properties. This associates classes together and defines the essential dependencies in an evaluation initiative. The properties form the intelligence of the ontology, as, to establish reasoning paths, one must describe the initiative in terms of clear statements, rules, and constraints. In the ontology, the property constraints refer to their range, (a) like restricting certain subclasses of the domain to commune with certain subclasses of the range (for instance specific dimensions are affecting specific levels, as in the case of out-comes assessment, which refers to the personal, institutional and community levels), and (b) their cardinality (quantity of values that can be assigned to each class/subclass). It is important to observe that there exist particular properties that connect the classes of the strategic and procedural layers of DiLEO, denoting that decision making concerning the DL evaluation is based on an integrated set of reasoning paths and the activities, methods, and procedures performed during an evaluation experiment depend on (are deduced by) upper layer decisions. An instance is the Research Question class, which is a strategic layer class and is connected with the Criteria class through the isReflectedOn property as well as with the Findings class through a isAddressedBy property. Table 1 presents the relationships used in the ontology.

<table>
<thead>
<tr>
<th>Property</th>
<th>Domain</th>
<th>Range</th>
<th>Constraints</th>
</tr>
</thead>
<tbody>
<tr>
<td>isCitedIn</td>
<td>Appellations/study identifier (AP/stid)</td>
<td>Appellations/study reference (AP/strf)</td>
<td>max cardinality=1</td>
</tr>
<tr>
<td>inverse: isCiting</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hasDimensionsType</td>
<td>Dimensions (D)</td>
<td>Dimensions Type (DT)</td>
<td>min cardinality=1, ∃ (fmtv ∪ smtv ∪ itrv)</td>
</tr>
<tr>
<td>isAffecting</td>
<td>Dimensions (D)</td>
<td>Level (L)</td>
<td>min cardinality =1, ∃ (cntl ∪ engl ∪ prcl ∪ infl ∪ indl ∪ insl ∪ socl)</td>
</tr>
<tr>
<td>inverse: isAffectedBy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hasConstituent</td>
<td>Dimensions (D)</td>
<td>Activities (A)</td>
<td>min cardinality =1, ∃ (rcrd ∪ mesr ∪ anlz ∪ comp ∪ intp ∪ rprt ∪ remc)</td>
</tr>
<tr>
<td>inverse: isConstituting</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>isSupporting</td>
<td>Instruments (I)</td>
<td>Activities (A)</td>
<td>min cardinality =1, ∃ (rcrd ∪ mesr ∪ anlz ∪ comp ∪ intp ∪ rprt ∪ remc)</td>
</tr>
<tr>
<td>inverse: isSupportedBy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hasPerformed</td>
<td>Means (M)</td>
<td>Activities (A)</td>
<td>min cardinality =1, ∃ (rcrd ∪ mesr ∪ anlz ∪ comp ∪ intp ∪ rprt ∪ remc)</td>
</tr>
<tr>
<td>inverse: isPerformedIn</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hasSelected</td>
<td>Means (M)</td>
<td>Criteria (C)</td>
<td>min cardinality =1</td>
</tr>
<tr>
<td>inverse: isSelectedIn</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hasMeansType</td>
<td>Means (M)</td>
<td>Means Type (MT)</td>
<td>min cardinality =1, ∃ (qltv ∪ qntv)</td>
</tr>
<tr>
<td>isMeasuredBy</td>
<td>Metrics (MTR)</td>
<td>Criteria (C)</td>
<td>min cardinality =1</td>
</tr>
<tr>
<td>inverse: isMeasuring</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
To demonstrate how properties define statements that describe precisely an evaluation initiative, one evaluation case from the Khoo et al. (2008) collective work on the use of web metrics in evaluation of DLs as an instantiation of the ontology is used. Figure 2 presents this instantiation as a semantic network. Each box corresponds to an ontology class and includes the names of the class and its subclasses—the classes and subclasses for which an instance exists as well as the corresponding instances appear in bold characters. Moreover, Figure 3 presents a fragment of the OWL file that

---

2 The notation used for the statements is as follows: Class/subclass: [instance_identifier] – property – Class/subclass:[instance_identifier]. The notation of instance identifiers follows the syntax: study identifier – underscore – instance name with capitalised lexical units.
implements the ontology with a number of evaluation studies as instances.

Fig. 2. The instances of the DiLEO classes as a semantic network
The study by Khoo et al. has an identifier wm2008c (AP/std:{wm2008c}) in the knowledge base and is part of a set of evaluations with common denominator the application of web metrics for different evaluation purposes. The NSDL study presented in this work has the scope to gather essential information to advance the technical excellence of the system and constitutes a summative evaluation. This dimension is affecting the individual level, as it measures user performance, behavior and satisfaction. Thus, this relationship would acquire the formal expression of D/texc:{wm2008c_TechnicalExcellence} – isAffecting – L/indl:{wm2008c_IndividualLevel}. The study focuses on metadata and the use of functionalities and appears to aim at the documentation and the enhancement of design, producing therefore the pair of expressions D/texc:{wm2008c_TechnicalExcellence} – isAimingAt – G/dscr:{wm2008c_Description} and D/texc:{wm2008c_TechnicalExcellence} – isAimingAt – G/docm:{wm2008c_Documentation}. The two objects of the study were operated by human agents that in the field of usability evaluation were characterized by their profession, namely, K12 Educators and University TA.

(a) 
<isMeasuredBy>
  <SystemInitiated rdf:ID="wm2008c_SessionTime">
    <isMeasuring rdf:resource="wm2008c_SizeOfUsage"/>
    <hasInitiatedFrom>
      <ProcessingLevel rdf:ID="wm2008c_ProcessingLevel">
        <isAffectedBy rdf:resource="wm2008c_TechnicalExcellence"/>
      </ProcessingLevel>
    </hasInitiatedFrom>
  </SystemInitiated>
</isMeasuredBy>

(b) 
<isUsing>
  <ResearchArtifacts rdf:ID="wm2008c_PaperPrototypes">
    <isSupporting rdf:resource="wm2008c_MeasureOpinions"/>
    <isUsedIn rdf:resource="#wm2008c_UsabilityTesting"/>
  </ResearchArtifacts>
</isUsing>

(c) 
<isReporting>
  <Findings rdf:about="#wm2008c_MeanTimeToClick=25.8sec">
    <isAddressing rdf:resource="wm2008c_ExplorationOfSessionLengthAsMetric"/>
    <isReportedIn rdf:resource="wm2008c_ReportFindings"/>
  </Findings>
</isReporting>

Figure 3: Instances expressed in OWL

The activities of measurement (for web usage), recording (for the usability and the survey/interviews legs of the study), analysis, and reporting are constituents of the technical excellence dimension. These activities were performed in the frame of logging and laboratory means: A/mesr:{wm2008c_Measure} – isPerformedIn – M/lgst:{wm2008c_LoggingStudies}, A/recr:{wm2008c_Record} – isPerformedIn – M/lbst: {wm2008c_LaboratoryStudies}, A/recr:{wm2008c_Record} – isPerformedIn – M/svst: {wm2008c_Interviews} and A/recr:{wm2008c_Record} – isPerformedIn – M/svst: {wm2008c_Questionnaires}, respectively.
These means had qualitative and quantitative means type and had specific aim-related criteria, such as the size of usage and types of use: M/lgst:{wm2008c_LoggingStudies} – hasSelected – C/spca:{wm2008c_SizeOfUsage}, M/lbst:{wm2008c_UsabilityStudy} – hasSelected – C/spca:{wm2008c_TypesOfUse}, M/svst:{wm2008c_Interviews} – hasSelected – C/spca:{wm2008c_TypesOfUse} and M/svst:{wm2008c_Questionnaires} – hasSelected – C/spca:{wm2008c_TypesOfUse}. These criteria were grouped by two main categories, usage and usability, and were measured by system-originated metrics, such as interface clicks and session time (Figure 3a), as well as user-originated metrics, like the reported types of use. During their logging study, the authors were using software to log and visualize data, which supported the respective activities, namely, measuring and reporting, while in the interviews and the surveys, the authors used question items to measure the opinions of the participants. In the usability testing study, the research team used paper prototypes to measure and analyze the participants’ preferences (Figure 3b). The findings that were reported concerned the mean, the mode time to click, and the clicks on the distracting graphics and poor navigation that addressed the second research question of the usability of the search results page.

Knowledge-Based Evaluation

The proposed ontology can serve as a schema to extract knowledge on evaluation initiatives and support the decisions. In particular, DiLEO could support the analysis of the data of previous initiatives, by comparing examples that share common characteristics and by finding their differences. Moreover, it provides a decision tool for designing evaluation initiatives based on retrieved statements that provide sequences of goals, research questions, objects, followed methods, and findings.

To demonstrate this potential a knowledge base has been developed in OWL, using the Protégé ontology editor; an evaluator can pose queries to the knowledge base and get assistance in planning experimental procedures.3

(a) SELECT DISTINCT ?Instruments WHERE { ?Means a <Surveys>. ?Instruments <isUsedIn> ?Means }

(b) SELECT DISTINCT ?Instruments WHERE { ?Means a <Surveys>. ?Instruments <isUsedIn> ?Means. ?Means <hasPerformed> ?Activity }

Figure 4: Indicative SPARQL queries for the first scenario

The demonstration of the ontology exploitation comprises two indicative scenarios. The first refers to the comparison of evaluation initiatives that share common properties and assumes that an evaluator needs to find information about the instruments used in questionnaire studies and compare their use (Figure 4a). The knowledge base would return instances such as (a) I/sftw:{eco2002_AccessDatabase}, I/sftw:{eco2002_SPSS}, I/sftw:{eco2002_SurveySolutions}, I/sftw:{nric2009_Software} for the software used in the studies of Cherry and Duff (2002) (AP/std:{eco2002}) and the search box, which addressed the research question of the exploration of session length as metric (Figure 3c) and Kostkova and Madle (2009) (AP/std:{nric2009}), I/nrvi:{eco2002_Questions}, (b) I/nrvi:{nric2009_Questions} and I/nrvi:{nric2009_Tasks} for the narrative items of these two studies, and (c) I/stat:{eco2002_Statistics} and

3 The OWL file is available at http://dlib.ionio.gr/~gtsak/dileo.
Given these results, the evaluator could pose the query presented in Figure 4b and observe that these instruments are used in different activities, identifying alternative evaluation patterns (practices). Following the path I – isUsedIn – M/svst – hasPerformed – A, the evaluator can identify that the activities of the study AP/std:{eco2002} were: (a) to measure (A/mesr:{eco2002_Measure}) the participants’ opinions through narrative items and software solutions, (b) to analyze them (A/anlz:{eco2002_Analyze}) with the help of a database, a statistical processing software and statistical principles and techniques, and (c) to compare (A/comp:{eco2002_Compare}) their data with the latter. In the study (AP/std:{nric2009}) the tools were used: (a) to measure (A/mesr:{nric2009_Measure}) via questions, tasks and software, and (b) to analyze (A/anlz:{nric2009_Analyze}) via software. In conclusion, the evaluator can trace elemental differences between the studies as in one of them (AP/std:{nric2009}) the questionnaire survey was complimentary to other methods and thus it is not assigned with equal gravity to the instruments of the other study (AP/std:{eco2002}) that conducted exclusively a questionnaire survey.


(b) SELECT DISTINCT ?Goals WHERE {?Goals <isAimedBy> ?Dimensions. }

(c) SELECT DISTINCT ?Factors WHERE {?Means <isDependingOn> ?Factors. }

Figure 5: Indicative SPARQL queries for the second scenario

The second scenario addresses the need to plan an evaluation initiative and is demonstrating how the knowledge extracted from the knowledge base supports the decision making. Assume that the specific aim of an evaluator is to plan a study for the improvement of the design of a DL. For this purpose, she decides to conduct a log analysis to analyze information about the usage of the system. A reasonable concern would be the identification of the research questions that an initiative like this should investigate. A query to the knowledge base (Figure 5a) will highlight the research statements that have been already expressed by other evaluators, such as the ones in the Khoo et al. (2008) study, which explored the session length (e.g., RQ:{wm2008c_ExplorationOfSessionLengthAsMetric}), or the Jones, Cunningham, McNab, and Boddie (2000) study, which focuses on the monitoring of user actions (e.g., RQ:{nzdl2000_DescribeUserActions}).
Additionally, the evaluator could unveil several alternatives, such as those for which other dimensions log analysis have been used. In this case, the knowledge base would return instances where logs have been used—together with other methods—to assess the outcomes of a DL’s operation. Furthermore, the ontology can provide support to the evaluator to understand the idiosyncrasy of the method, so as to understand that it is a primarily summative type of study, or to reveal alternatives to the evaluator on significant areas, such as what goals this method serves (Figure 5b). For instance, logging studies have not been used only for describing the use of systems and/or data sets, but also for providing valuable information on the design of the DL. Therefore, the knowledge base can expand the evaluator’s knowledge and probably revise the initial planning by widening its scope. Because many evaluation campaigns do not use only one method to collect and analyze data, the evaluator can also retrieve important information on how other methods can complement logging studies and what are the stages that that would need to be performed.

Furthermore, through the ontology the evaluator can understand the limitations that the method has to a greater extent. In particular, they could understand which timeframes are sufficient for conducting studies via logs by getting responses such as F/tim: {nzdl2000_30September1996-1December1996} and F/tim: {nzdl2000_April1996July1997}, which are the two timeframes that Jones et al. (2000) needed to collect specific log data (Figure 5c). Based on these results the evaluator can identify the research questions used in log analysis studies and could proceed with more queries to gain a deeper insight of the existing studies, identifying their criteria and criteria categories, as well as the used metrics, so that to decide on the main parameters of a new initiative.

Validation and Refinement

The aforementioned examples are hypothetical cases of the use of DiLEO indicating its expressiveness and reasoning strength. However, these scenarios do not constitute a valid representation of real users’ demands from an instrument like this. For the DiLEO to be validated, it was tested in two separate events involving researchers and practitioners who had either a strong interest in evaluation planning or significant experience in the performance of evaluations. The ontology was first explored in the context of a tutorial held as part of an international conference on DLs. The participants of the tutorial, approximately 15, represented different types of researchers and different domains of practice (for instance, computer scientists, librarians, subject librarians focusing on particular domains, like chemistry, etc.) and represented a wide range of institutions, such as academic, research, and national libraries, as well as known scientific information providers. Many of them had a strong background in DLs and IR systems and testified that evaluation is a multidimensional construct featuring synthesized requirements.

During the first stages of the tutorial, the ontology was presented thoroughly and then the participants were asked to use the ontology paths to describe a plan for evaluating their institution’s DL. The aim of this exercise was to trigger a discussion for the usefulness of the ontology and its acceptance as a tool for planning evaluation experiments.

The discussion revealed diverse comments about several properties that constituted a first round of validating and refining notes. Two particular findings were very informative, both referring to planning. The first finding regarded how DiLEO can model a set of evaluations that are assembled to a wider evaluation effort. Evaluations often comprise individual studies that align to a certain scope and serve the purpose of a bigger evaluation effort. It was thus of great interest to discuss about the way these evaluations are expressed in either one descriptive instance or many interlinked ones. The second finding referred to the ability of the researcher to indicate a steady workflow; the participants stated that the definition of entry points to the ontology graph would help the evaluator to formulate a course. For instance, one participant mentioned that one of the starting points he would suggest would
be the Factors class, because this designates all the limitations that will affect the consideration of options in other classes. According to his words “the most important factor was the available resources. It worked like “we have these resources, what research questions can we ask? This may not be a reasonable way to start the design of a study but I think it is a very important factor that comes up quite early in the planning.”

The second event was a focus group with five library officials (two of them are also researchers on information science) chosen for their deep experience in conducting evaluation experiments in DLs. The focus group lasted 2 hours, in which the moderators catalyzed a discussion that was recorded and later analyzed, while simultaneously they took field notes on a protocol. After an introduction to DiLEO, the participants stated their experiences in evaluation planning and performance and discussed about the challenges they face during these phases. They usually conduct evaluation experiments to record the current state of operation of a DL and to answer questions on system improvement. Their common goal is to identify whether the DL services are satisfactory, mainly in terms of collection and retrieval. The usual resources for planning an evaluation are the literature and the relative studies in their field. Additionally, they underlined the lack of planning tools that can guarantee them a holistic view of the process.

The participants believe that the classes of the ontology in general cover adequately the core concepts of the evaluation process and help them express a workflow. Similar to the first validation event, they marked the necessity of defining a particular class for starting planning an evaluation initiative. The participants agreed that even though the ontology is not recommending a starting point, one should start his evaluation planning from posing the appropriate Research Questions. They justified their rationale saying that Research Questions imply the high-level concepts, such as dimensions, guide the procedural steps of the study, and guarantee methodological soundness. In the end of this part of the discussion, one participant commented that the independence of the ontology from starting points “requires familiarity with the ontology.”

The participants stated that DiLEO could provide alternative paths and could assist them in considering all planning options. When the discussion was extended to the issue of clarity, the participants noted that there is a confusion about some terms, such as in the case of Goals and Research Questions, and asked for clarifications. After further elaboration, this confusion was resolved, but the issue of clarity remained as some classes of the ontology were not easily perceived during the presentation. It was noted that some concepts were not comprehensible unless being viewed in the graphical presentation of the ontology, where their meaning was clarified by their location in the graph and their adjacent concepts. They also suggested new relations between the existing classes. For instance, one participant suggested two new relationships, namely, isReflectedOn and isPresupposing. These suggestions were incorporated in the ontology as it was judged that they provide significant semantic connections between their domain (Research Questions and Instruments, respectively) and ranges (Criteria and Metrics, respectively, see Table 1). During the discussion, some of the participants thought that new classes could be introduced as well. One of the suggestions referred to the class “Limitations,” a proposal that was abandoned as the discussion progressed and its proximity to the Factors class was grounded.

Overall, the participants in both validation experiments stated that DiLEO could be a helpful and useful instrument, but they did not consider it friendly enough. They believe that a human-friendly interface would facilitate data entry and support researchers to plan effectively their evaluations. The data gathered from these two events were taken into consideration during the refinement of the ontology.
Ontologies are useful instruments that inflate the understanding in a domain on the basis of a commonly agreed vocabulary. One can use ontologies to express precisely the rich semantics of the DL evaluation literature and to formulate reasoning declaratives. Marchionini (2000, p. 311) states that evaluation “is a research process that aims to understand the meaning of some phenomenon situated in a context and the changes that take place as the phenomenon and the context interact.” Forwarding to Marchionini’s words one can find that “evaluation specify what is the research process (metrics and procedures), what is the phenomenon (its mission, and salient characteristics), and the context(s) in which the phenomenon operates” (pp. 311–312). The semantics of these important concepts, such as the concept of understanding a phenomenon and its context, could be represented by the ontology as paths. In particular, the understanding of a phenomenon can be described by the path D – isFocusingOn – O – isOperatedBy – S. Given that an evaluation experiment is a phenomenon, the path includes the participating entities in an interactional event, as well as the subjective and partial view and exploration of it. Additionally, context can be represented by the union of the paths L – isAffectedBy – D – isFocusingOn – O – isOperatedBy – S and D – isAimingAt – G. This path contains the conceptual space that nests the expression of both the view of the phenomenon and its impact, along with the interaction event itself.

DiLEO is the result of a modeling process that covers horizontally the DL evaluation domain attempting—through its abstraction—to address concepts from many disciplines. Its scope is broader in terms of describing evaluation instances and assisting planning purposes, instead of suggesting specific evaluation methods and criteria. Its discriminative feature is that it focuses on the standardization of the evaluation workflows instead of simply describing formally particular constructs, such as the DL components or evaluation criteria. Hence, it reveals the relationships between the main concepts of the DL evaluation domain and thus it covers both modeling approaches, either object-based or workflow-based. For instance, the 5SQual model by Gonçalves et al. (2007) concentrates on the specific concept of quality. The 5SQual model describes formally the DL objects and services and correlates their features to quality criteria. In DiLEO, the association between Objects, which encapsulates both digital objects and services, and Criteria is expressed by the path O – isOperatedBy – S – isParticipatingIn – M – hasSelected – C. This path reveals aspects of the evaluation context, such as the subjects that use these objects and the means these criteria are applied. On the other hand, PRET A Rapporter, workflow model with a planning perspective as well, shares critical concepts of the evaluation structure, such as the purpose, resources and constraints, techniques, and analysis. However, PRET A Rapporter lacks the relationships that conceptually bridge these concepts. A small fracture of the process in one of the cases that Blandford et al. (2008) present can be described by DiLEO as M/flst{pret2008b_fieldstudy} – isPerforming – A/rec:{pret2008b_Record_think-aloud} – isSupportedBy – I/dev{pret2008b_audiorecorder}, adding thus semantic value to these procedures.

Evaluations are very complex processes and researchers often follow equally complex ways to reach their end, often improvising to increase the value of their findings and recommendations. For example, in several evaluations the means types were a union of qualitative and quantitative. One evaluator can do different kind of processes than those “prescribed,” such as collect data via ethnographic methods and analyze them, both qualitatively and quantitatively. As Ingwersen and Järvelin (2005, p. 250) point out, the role of these approaches differs in the data analysis stages and not only in their collection. An instance is the use of sophisticated logging methods, such as the one mentioned by Nicholas (2009), which allow a parallel— to the dominant quantitative—qualitative analysis of the data. In this case, Nicholas selected three anonymized logging sessions and analyzed them to find referrals and feed them back into the study. This example could be expressed by the path A/anlz:{bll2009_Analyze} – isPerformedIn – M/lgst:{bll2009_Logs} – hasMeansType – MT{bll2009_Quantitative} and A/anlz:{bll2009_Analyze} – isPerformedIn – M/lgst:{bll2009_Logs} – hasMeansType – MT{bll2009_Qualitative}. The ontology can describe both qualitative and quantitative works, host
information that is derived from different standpoints and support both comparison and evaluation planning tasks. Tools like ontologies can address the problems that Urquhart (2010) states by saying that “trying to integrate findings from studies done from different standpoints is difficult – here be dragons of different viewpoints of what counts as valid knowledge, as well as the practical problems of trying to convert qualitative ‘important’ findings to something can be melded with quantitative data.” Thus, DiLEO serves systematic review procedures, as its abstraction facilitates a neutral approach to research paradigms and it can accommodate both quantitative and qualitative data and assist meta-analysis and meta-synthesis reviews.

Furthermore, an ontology, such as DiLEO, can be a schema for classifying instances in a structured network of assertions and axioms, thus supporting comparison, aggregation, and synthesis. One of the problems of the research community is the ability to produce comparable data between the evaluations—data that can be collated through time and space. However, because of the uniqueness of each DL system and the inherent properties of each evaluation campaign, these data cannot be easily exploited. The attempts to generate instruments and tools, such as logging schemes (Klas et al., 2006) or performance measurement indicators, are constrained by the processes of data collection and analysis, often ignoring or decreasing the importance of contextual parameters. Hence, the proposed ontology emerges as a system that can exploit the required primitives to portray a holistic image of an evaluation initiative and make each of its constituent instances comparable and interpretable.

The proposed ontology follows some structural conventions. Apart from the core classes that accommodate much of the domain’s vocabulary, it comprises specific classes that serve as classification mechanisms of instances into meaningful groups. An example is the class of Metrics (MTR), which uses the organization of metrics under three categories, namely, user-originated, system-originated, and content-originated, to group them semantically and channel them to the respective Levels (L). The concepts of user, content, and system, together with the concept of levels, as articulated by Saracevic and expressed as a class in the proposed ontology, are vital because they represent abstractions of the “reality” or else the domain. These are essential to the scope of merging the cultural differences and amplifying the understanding between agents from different backgrounds. The three constituents (also appearing in the ontology as objects, i.e., content and system, and subjects, for the purpose of representing evaluation acting entities) are considered integral parts of many DL architectural schemes, while Levels can be assessed as a concept that provides to different stakeholders a tiered representation to define the focus of efforts. The concept of Levels has also served the clustering of evaluation criteria that are crucial to different participants in the DL evaluation domain (Zhang, 2010). Furthermore, for the purpose of supporting the explicitness of the vocabulary, the class of Criteria classifies typologically the respective instances, while the class Criteria Categories categorizes them semantically. Finally, the classes Research Questions and Findings enumerate instances being seminatural language statements, as the current limited lexical units in the ontology vocabulary cannot convey their semantics.

Populating the knowledge base is not an easy task for many reasons. One of them is the openness to the various sources of information. The ontology does not dictate a set of quality criteria for initiatives to be included, as is done by systematic reviewing processes in other domains, like the Cochrane Collaboration (http://www.cochrane.org/cochrane-reviews). The aim of the ontology is not such. It can address from very detailed expressions of evaluations to ill-defined and incomplete descriptions, aiming to host and integrate different “cultures” that are subject of the idiosyncrasies of the reporter and communication channels. One proof of the cultural differences was exhibited in the validation phase, when two participants from different domains expressed the same need, a starting class, indicating however different classes. Hence, ontology-based planning concludes to the statement of Uschold and Gruninger (1996, p. 99) “by using an ontology to provide a normative model of the system, this integration (note: meaning of different user perspectives) can be achieved by assisting participants in communicating and coming to an agreement.”
Conclusions

Engineering an ontology for the DL evaluation domain is not a trivial task, because of the uniqueness of the applications and the various research backgrounds existing in each project. Therefore, the convergence range is wider and leaves vast spaces for coverage. However, the ontology presented here attempts to bridge the abstract levels of an initiative, such as specifying the scope of the evaluation, and the lowest ones, such as detailing the research parameters. The ontology scheme can be used as a benchmark between evaluations, and it can be exploited for research design recommendations using the reasoning strength of ontology. Compared with other modeling systems, ontologies transform abstractions of past events into significant aids for the design of future evaluations, but their potential allows more functionalities to be added by defining particular rules that orient the evaluation planning to (a) more precise options enabling reasoning mechanisms and (b) adaptive options based on the particular needs of an institution.

Acknowledgments

We thank Professor Maristella Agosti for her constructive recommendations, and Steven Reid, Paul Jerome Smith, and Euridice Skarmountzou for reviewing and suggesting improvements on the article, as well as the anonymous reviewers for their insightful comments.

References


