A Framework to Conduct and Report on Empirical User Studies in Semantic Web Contexts

Catia Pesquita[1], Valentina Ivanova[2], Steffen Lohmann[3], Patrick Lambrix[2]

[1]LASIGE, Faculdade de Ciências da Universidade de Lisboa, Lisbon, Portugal
clpesquita@fc.ul.pt
[2]Linköping University, Linköping, Sweden
patrick.lambrix@liu.se
[3]Fraunhofer IAIS, Sankt Augustin, Germany
steffen.lohmann@iais.fraunhofer.de

Abstract. Semantic Web technologies are being applied to increasingly diverse areas where user involvement is crucial. While a number of user interfaces for Semantic Web systems have become available in the past years, their evaluation and reporting often still suffer from weaknesses. Empirical evaluations are essential to compare different approaches, demonstrate their benefits and reveal their drawbacks, and thus to facilitate further adoption of Semantic Web technologies. In this paper, we review empirical user studies of user interfaces, visualizations and interaction techniques recently published at relevant Semantic Web venues, assessing both the user studies themselves and their reporting. We then chart the design space of available methods for user studies in Semantic Web contexts. Finally, we propose a framework for their comprehensive reporting, taking into consideration user expertise, experimental setup, task design, experimental procedures and results analysis.

Keywords: Semantic Web · empirical evaluation · user study · user interface · literature review · design space · protocol · reporting.

1 Motivation

The Semantic Web enables intelligent agents to create knowledge by interpreting, integrating and drawing inferences from the abundance of data at their disposal. It encompasses approaches and techniques for expressing and processing data in machine-readable formats. Semantic Web technologies are being applied to increasingly diverse areas where user involvement is crucial. Providing carefully designed user interfaces, visual representations and interaction techniques has the potential to foster a wider adoption of Semantic Web technologies and to lead to higher quality results in different application contexts where ontologies and Linked Data are employed.

As the number of user interfaces for Semantic Web systems is growing, one important step is to evaluate their capabilities and features in order to reveal their usefulness together with their advantages and disadvantages. As organizers of the VOILA

Preprint of a paper to appear in the proceedings of the 21st International Conference on Knowledge Engineering and Knowledge Managemen (EKAW '18) to be published by Springer

workshop series[1], we noticed that both the assessment of interactive Semantic Web approaches as well as the reporting on conducted user studies still suffer from weaknesses.

We can basically distinguish at least three evaluation approaches: i) *formal evaluation*, based on defined models, for instance, a cost-based model where costs are assigned to different user actions executed in order to achieve a certain goal; ii) *automated evaluation*, aiming to reveal computational—as opposed to visual—scalability and efficiency of approaches and algorithms, and iii) *empirical evaluation*, based on the observation of users who interact with a system. In this paper, we focus on this latter category—empirical evaluation.

In short, empirical evaluation refers to the testing of user interfaces by real users. The various methods are usually categorized into quantitative methods (e.g., controlled experiments) and qualitative methods (e.g., inspection methods). Drawing from relevant literature [1], all common evaluation methods exhibit to a different extent the following factors: i) *generalizability* (or external validity, i.e., the extent to which the results apply beyond the immediate setting, time and participants), ii) *precision* (or internal validity, i.e., the degree to which one can be definite about the measurements that were taken and about the control of the factors that were not intended to be studied) and iii) *realism* (or ecological validity, i.e., the degree to which the experimental situation reflects the type of environment in which the approach will be applied); they serve different purposes and are eventually conducted during different stages of user interface development (e.g., formative vs. summative evaluations).

Regarding external validity, one differentiating aspect in conducting empirical evaluations for the Semantic Web versus other fields of study is that users of Semantic Web tools can typically not be categorized along a single axis of expertise. Considering that Semantic Web tools are often used in domains where information complexity is an issue (e.g., life sciences, governance, health care), it becomes essential to be able to understand user expertise both with the domain that underlies the data being used and explored by the tool, but also with knowledge modeling and representation concepts. This poses challenges in assessing population validity, since both expertise axes need to be considered.

Another issue is the generalizability to other situations, for instance, when applying an approach to different datasets, especially those with varying degrees of semantic complexity. Ecological validity (i.e., the degree to which the experimental situation reflects the type of environment in which the approach will be applied) is also of particular concern in Semantic Web contexts, since both population and dataset characteristics need to be accounted for.

In this paper, we present a review of empirical evaluations published in the Semantic Web community in recent years (Sec. 2). We then discuss the design space of evaluation methods for interactive Semantic Web systems (Sec. 3), and use this as a springboard to outline a protocol for reporting on user studies in Semantic Web contexts (Sec. 4). The design space and protocol together constitute a framework for conducting and reporting on empirical user studies in Semantic Web contexts. In Sec. 5 and Sec. 6, we summarize related work and provide a discussion, before we conclude the paper in Sec. 7.

[1] VOILA: International workshop series on "Visualization and Interaction for Ontologies and Linked Data", see `http://voila.visualdataweb.org`

2 Literature Review of User Studies in Semantic Web Contexts

We define a user study in the context of the Semantic Web (SW) as any user-based empirical evaluation of a system, tool or method that employs SW technologies. The purpose of the evaluation may span different aspects, such as the assessment of graphical user interfaces, ontology and Linked Data visualizations or user interaction techniques.

2.1 Methodology

We conducted a literature review covering the following four conference and workshop series dedicated to the SW, in their 2015, 2016 and 2017 editions: i) ISWC (International Semantic Web Conference), ii) ESWC (Extended Semantic Web Conference), iii) VOILA (International Workshop on Visualization and Interaction for Ontologies and Linked Data) and iv) IESD (International Workshop on Intelligent Exploration of Semantic Data). The first two venues were selected given their primacy in SW-dedicated conferences, whereas the latter two for their specific targeting of user interaction and visualization in SW contexts.

We restricted the review to papers where the expressions "user study", "user evaluation", "empirical evaluation", "interaction" and/or "visualization" appeared in the abstract—also taking into account spelling differences (e.g., "Visualization" and "visualisation") and word form variations (e.g., plural forms). This resulted in a total of 87 papers. All papers were analyzed in their entirety and split into three groups: i) papers that include a report of a user study (46 papers); ii) papers that do not report on a user study but present a SW approach addressing user interactions (35 papers); iii) papers that do not report on a user study and do not present an interactive approach, such as position papers (six papers).

This distribution can already be seen as an indicator of the lack of user studies in SW publications that report on a system, tool or method concerned with user interaction.

Each paper of the first group was further categorized within three aspects: i) *purpose*, ii) *users* and iii) *evaluation methods*. We followed an inductive analysis approach to identify the major categories or themes within each aspect [2]. Each paper was assigned a category and code to reflect a relevant characteristic. For instance, Mitschick et al. [3] write that their "[...] interface provides an expressive but still approachable way of querying for specific entities and their accompanied information" and thus was assigned the code *querying* under the *purpose* category.

A running list of codes was shared between all four coders (all researchers, namely the authors of this paper) to ensure code reuse when possible. After each paper was coded by one researcher, the full list of codes was edited to ensure coherence and remove any remaining duplicates, and codes were organized in a hierarchy. For *purpose*, we defined two broad categories: *learning & understanding* and *creating & managing*; for *users*, we defined *participant number*, *participant expertise* and *participant recruitment*; and for *evaluation method*, we defined *quantitative* and *qualitative*. Finally, the code assignment for each paper was revised by at least two of the other researchers.

The classification resulting from the literature review and coding is available as a table on the Web, published under a Creative Commons license.[2]

[2] The classified papers can be accessed at: `http://survey.visualdataweb.org`.

2.2 Purpose

Purpose describes the general intent of the operations[3] supported by the evaluated approach. We induced a list of operations from the reviewed papers, which fit into two broad categories: *learning & understanding* and *creating & managing*. *Learning & understanding* is concerned with information and knowledge acquisition needs, whereas the purpose of *creating & managing* operations is the creation of new content, its manipulation and lifecycle support. These categories are further discussed in Sec. 3. The results of the classification are listed in Table 1. Note that several of the users studies reported in the 46 papers looked at more than one operation type and purpose.

Purpose	Operation	N. of user studies
learning & understanding	exploration	17
	navigation	1
	search	8
	querying	10
	question answering	1
	explanation	2
creating & managing	modeling	10
	editing	9
	validation	1
	mapping	1
	annotation	5

Table 1: Purposes and operations reported in papers that included user studies

Our systematic literature review revealed that the majority of works aim to support information exploration and seeking behaviors. These behaviors differ from navigation and information retrieval where users' information needs and questions of interest can be specified and expressed in advance before an interaction with a SW approach. Information exploration activities are usually more open-ended with evolving (on the basis of current observations) information needs, personal experience, motivation and context. These are high-level complex activities characterized by uncertainty and acquiring unexpected findings as the exploration progresses. Users may lack knowledge in the area of interest (often referred to as *exploratory search* [4]) or may possess domain expertise, without being familiar with a particular multi-dimensional dataset (and employing an exploratory environment to understand and use it).

2.3 Users

Regarding users, we classified the papers according to three aspects: i) *participant number*, ii) *participant expertise* and iii) *participant recruitment*. Table 2 shows the result of this classification. The categories are non-exclusive, i.e., in several studies, users with diverse areas and levels of expertise are recruited. Further, a couple of papers included more than one user study with different numbers of participants.

[3] We use the term *operation* instead of *task* here to differentiate it from *evaluation tasks*.

User aspect	Category	N. of user studies
participant number	not reported	3
	1-9	9
	10-19	16
	20-29	18
	30+	8
participant expertise	not reported	4
	SW	14
	IT	10
	domain	8
	non-expert	12
	diverse	14
participant recruitment	not reported	25
	researchers	4
	students	11
	clients/users	4
	crowdsourcing	1

Table 2: Distribution of the reviewed papers according to user aspects

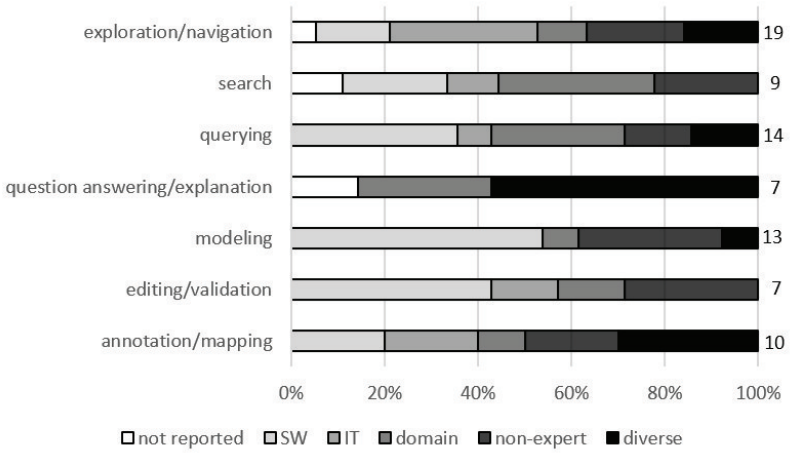

Fig. 1: User expertise by type of purpose and operation. The numbers on the right indicate the total numbers of user studies within each operation.

The majority of user studies were conducted with sample sizes between ten and 29 participants. Of the 46 reported studies, 14 employ a diverse mixture of participants, while only four do not mention any user area and/or level of expertise. The situation is less encouraging with regards to the used recruitment method. Nearly none of the papers explicitly reports the method used to recruit the study participants. Some do (more or less detailed) indicate the population to which the participants belong (e.g., students, researchers, etc.), so we took those as our (broad) recruitment categories. Still, more than half of the study reports do not include a description of the recruitment method they employed. Further, students were recruited as participants in eleven studies, which can be critical, as this may result in a sampling bias and can negatively affect the population validity when the participating students do not match well the needs and characteristics of the target population (i.e., the actual users of a SW tool, method or system).

To further understand the impact of participant expertise across user studies in SW contexts, we looked into each *purpose* category and determined the area of *participant expertise* within it (cf. Figure 1). This revealed that the participant expertise is not uniformly distributed across purposes. One interesting finding is that domain expertise is more frequent in querying, search and question answering/explanation, and SW expertise clearly dominates in modeling, but also has a high prevalence in editing/validation and querying.

2.4 Evaluation Methods

Regarding the evaluation methods, ten studies report using qualitative strategies, while 39 employ quantitative approaches (cf. Table 3; note that the categories are non-exclusive again, i.e., a number of user studies applies more than one quantitative method). The most popular evaluation approach is the use of questionnaires. Surprisingly, standard questionnaires are only used in eight studies, whereas the authors of 20 studies apply their own custom questionnaires. More than half of the studies report using tasks in some fashion to support the evaluation. Of these, the majority records time to complete a task and uses success metrics or both (i.e., classical time and error measures). Few of the studies include comparative evaluations (eight in total), of which six use within-subjects designs and two use between-subjects designs. Although some study reports include detailed descriptions of the used evaluation methods and tasks (e.g., [5,6]), in most cases, only little space is dedicated to describing the design and procedure of the user evaluation (these descriptions often take up less than a page of the papers).

3 Design Space

The literature review we conducted revealed several limitations of empirical user studies in SW contexts that are shared by a majority of works. In the following, we chart the design space of such studies with the aim of giving SW researchers a guide to help them design and successfully conduct user studies. We have structured the design space into six dimensions: i) purpose, ii) users, iii) tasks, iv) setup, v) procedure and vi) analysis and presentation of data. For each dimension, we identify the main concerns to be taken into consideration, define categories to systematize procedures and techniques,

Type of evaluation	Category	N. of studies
qualitative	unspecified	7
	observation	1
	open interview	2
	total number of studies	10
quantitative	standard questionnaire	8
	custom questionnaire	20
	task success	13
	task time	2
	task success and time	9
	non-tracked task	3
	comparative within subjects	6
	comparative between subjects	2
	total number of studies	39

Table 3: Distribution of reviewed papers according to evaluation method aspects

and offer guidelines to support the design of adequate user studies. Based on these six dimensions, we then define the minimum information required to report on a user study in SW contexts.

3.1 Purpose

As already introduced, interactive SW approaches can typically be classified in two general categories based on their purpose: *learning & understanding* and *creating & managing*. Both categories include several high-level operations, which could also be present in the other category as supporting operations:

1. *Learning & understanding*: The main purpose of these approaches is to provide means for satisfying information needs and acquiring knowledge. This might be done due to different reasons, such as generating or validating a hypothesis, using a dataset or ontology in application development, looking up particular information or exploring a topic of personal interest due to curiosity. This category encompasses high-level tasks, such as exploration, navigation, search, querying and question answering (cf. Table 1). All of them aim at satisfying information needs, however, they differ in the extent the information need is defined at the beginning. Exploration is usually a more open-ended activity with vague initial goals which are evolving as it progresses (cf. Sec. 2.2). In comparison, direct search (as opposed to "exploratory search" [4] often mentioned in information retrieval) and querying have more clearly specified initial goals which might change as results are retrieved.

2. *Creating & managing*: the main purpose of these approaches is to provide means for creating and editing content. This might be authoring an ontology or creating a dataset by, for instance, creating mappings to publish content in various formats in RDF (high-level operations: modeling, editing and publishing data). Other operations might include linking datasets or ontologies, and discovering and resolving quality issues. This category also covers high-level tasks, such as documentation and annotation which create meta-data.

3.2 Users

More than half of the reviewed user studies have a reported *number of participants* too low to support quantitative analysis for user testing [7]. In fact, this may be one of the reasons why user studies are not common even in publications presenting systems for end-users. However, 10 ± 2 users have been reported to be often enough to detect 80% of the issues at least in qualitative usability studies [8]. Using crowdsourcing platforms may alleviate this issue, but it limits the type of evaluations that can be conducted [9].

Beyond concerns about the number of participants, SW researchers also need to take into consideration the skills and experiences of participants, since they can have an impact on the performance when using a SW approach to solve problems. Assessing user competence is a general concern in information systems research [10,11], and its importance is magnified in SW contexts [12,13], given that both *expertise of the participants* with SW concepts and with the domain at hand can impact user performance and experience. Due to the specificity and complexity of some of the domains where SW applications are applied (e.g., biomedicine, earth sciences), understanding the level of domain expertise required of target users and mapping it to the study participants should be a major focus of SW researchers designing user studies, since more significant conclusions can be drawn when there is a more significant overlap between the characteristics of study participants and target users, thus ensuring *population validity*. This is an increased concern in crowdsourced studies, where finding the 'right crowd' is still a challenge [14]. Also, recruitment strategies can have an impact on the results, due to bias (e.g., *selection bias* and *sampling bias*) as well as researchers who are testing their own tools (*experimenter bias*).

3.3 Tasks

Most of the reviewed SW user studies employ tasks as the basis for empirical evaluations. Defining tasks is typically an integral part of designing a user-based evaluation, more so in SW user studies where task complexity should be aligned with the different aspects of user expertise. To ensure *ecological validity*, the evaluation tasks should mirror typical tasks target users are expected to perform using the system, and their definition needs to be articulated with care, taking into account the user characterization as well as experimental setup and procedure. In SW user studies, it is particularly important when considering task performance to be able to discern if failure was due to the user's unfamiliarity with the domain of the SW resources being used, i.e., to ensure *construct validity* or, in other words, that the evaluation is measuring what it is supposed to measure. It is also crucial to take into consideration the characteristics of the datasets

used in the evaluation, including their domain and semantic complexity. Understanding how these articulate with user expertise is necessary for an adequate interpretation of results.

A task needs also to coordinate with the evaluation method employed in the study, e.g., open-ended tasks may be best employed with *think-aloud* techniques, case studies or observation techniques, whereas specific action-based tasks will allow quantitative measures, such as time to complete and accuracy. The majority of the studies we reviewed report on approaches that have exploration as their main target purpose. Designing such environments involves computational and algorithmic approaches entangled with visualization and interaction techniques to foster information seeking behaviors. Due to the nature of exploratory behavior, designing tasks to support the evaluation of such environments presents a particular challenge [15].

3.4 Setup

Two aspects should be addressed regarding experimental setup: setting conditions and study design. Setting conditions can influence the result of a test, thus special attention should be devoted to minimizing the variance of non-tested conditions (room, lighting, display size, etc.), i.e., maximizing *internal validity*.

It is often advisable to perform a comparative study when there are similar approaches available that can be compared to. Such comparisons against a baseline are often well suited to show the benefits and limitations of a new piece of work. Only eight of the reviewed SW user studies used a comparative approach, with different designs. Most popular is a within-subjects design where the study participants are exposed to both approaches, i.e., they first interact with one approach and then with the other. In this setting, it is important to control for *order effects* (e.g., by *counter-balancing*).

An alternative is a between-subjects design, where the study participants are split into two groups and each group sees and evaluates only one of the approaches. However, a main drawback of the between-subjects design is that it usually requires a much larger number of study participants to get useful and reliable data. This might be the main reason why it was applied in only two of the reviewed studies.

If more than two approaches (or conditions) are compared, counterbalancing quickly multiplies and study designs using Latin Squares and other incomplete counterbalanced measures need to be applied. One paper in our study used a Latin Square design to compare different visual querying interfaces [5]. Naturally, the more complex the study design gets, the more complex the analysis and the higher the probability that errors are made in the analysis of the results. Thus, keeping the study design as simple as possible is usually advisable.

3.5 Procedure

Many of the reviewed user studies evaluate the usability of the proposed tool or system, and while most do so using custom questionnaires, eight studies use standard questionnaires, such as the popular System Usability Scale (SUS) or the Post-Study System Usability Questionnaire (PSSUQ). Usability commonly comprises of effectiveness, efficiency, and satisfaction, which can be evaluated in different ways: effectiveness is

typically measured through task success, efficiency is measured through task speed, and satisfaction is measured through user feedback, as discussed in [16].

However, we observe that several of the reviewed user studies are limited to the evaluation of the usability of a tool or system, which is an important aspect but often not sufficient to fully evaluate an interactive system or tool and answer the research questions addressed in the work. In particular for exploratory tasks, which were very common in the reviewed user studies (cf. Table 1), cognitive measures are often required [17], such as looking at insights obtained while using an exploration tool [18] or associated metrics based on engagement, novelty, task success and time as well as learning [19].

Other ways to study exploratory behavior and cognitive processes is via eyetracking and the aforementioned think-aloud method. For instance, Fu et al. use eyetracking to compare indented lists and graphs as two different types of ontology visualizations [20]. Mitschick et al. applied a think-aloud method to learn about the cognitive model of the study participants [3].

3.6 Analysis and Presentation of Data

In a first step, data analysis concerns the collection and organization of data. There are several methods to compile and analyze both qualitative [21] and quantitative data [22].

When participants have diverse backgrounds and expertise levels and areas, it is useful to report separately on results for each group. In fact, comparative studies (either of several systems or of several user groups) pose additional challenges for experimental design and data analysis, especially when obtaining statistically significant results is a goal of the study. This requires a rigorous experimental design, in a much more controlled setting and with a larger sample of participants.

4 Reporting on Semantic Web User Studies

Our guidelines take inspiration from the molecular biology field where 'minimum information' guidelines to describe experiments were proposed quite early [23]. Our goal with defining the minimum information required to describe user studies in SW contexts is to ensure that the recorded information is sufficient to: (i) support the *interpretation* of the conducted user study; (ii) enable the *comparison* to similar evaluations; and (iii) permit the *replication* of the user study. These requirements imply that a detailed description of several aspects of the user study needs to be produced, and that the description should be as unambiguous as possible.

According to our guidelines, the minimum information about a user study in a SW context includes a description of the following six aspects:

1. Purpose: This aspect describes the general types of operations that are supported by the interactive approach under evaluation. We propose to categorize purposes into four non-exclusive types (the first two fall into the learning & understanding category and the last two in the creating & modifying category):

- *exploration*: includes operations such as exploratory search, browsing and navigation; it can be applied to operations where the information need is not clearly defined or the goal is general discovery and insight generation;
- *search*: includes search, querying and question answering; it applies to the more focused examination of SW content with a clear information need or specific target in mind;
- *creation*: includes operations such as modeling ontologies or RDF content, and creating mappings between SW resources; it applies to tasks where SW content is created;
- *management*: includes assessment, validation, annotation and editing of SW resources; it applies to operations where existing SW content is manipulated;

2. Users: This aspect contains information about the intended users of an approach, the participants of the user study and how well the two groups overlap. Many of the reviewed works do not describe their *target users*, which makes assessing the population validity nearly impossible.

A proper description of target users should include expected demographics but also *expertise levels in both SW and the domain* covered by the approach. Likewise, the demographics together with the SW and domain expertise of the study participants should also be reported, as well as information on the *participant recruitment*, i.e., which type of participants were recruited (e.g., domain experts recruited from a company, students of a university course, etc.) and how they were recruited.

3. Tasks: This aspect describes the tasks required of the participants. We separate tasks from the experimental procedure, because the same tasks can be utilized with different procedures and systems (and vice-versa). To support interpretation and reproducibility as well as allow for comparison, the report should include the exact task descriptions (e.g., the task form) given to the participants. For multi-purpose systems, tasks should further be categorized according to their purpose. Furthermore, the data used in each task should also be reported on and made available when possible. Exact descriptions of tasks and data are essential to support the assessment of ecological validity.

4. Setup: The setup should clearly describe the *type of evaluation* (controlled experiment, field study, etc.) and the *setting and interaction context* of the study participants. Further information of relevance is the exact *experimental design of the study*, such as the independent and dependent variables measured. Descriptions of design types and assignment procedures are, for instance, described by Field & Hole [24].

5. Procedure: The different *phases* of the evaluation should be described in chronological order to provide the reader with a clear picture of the procedure from the moment the participants arrived to the moment they left. Common phases to cover are: introduction, briefing of users (ethical issues), form filling and questionnaire (demographic information, etc.), instruction material, training (if any) and the actual testing session, post-test interview, and debriefing. Examples of *issues* to report on are: whether any assistance was given; how the tasks were presented to the participants (on paper/on screen, etc.) and how they were executed; how responses were given (clicking on a button, using the keyboard, etc.), and the overall participation time.

6. Analysis: This aspect should report on the selected *data analysis method* and show awareness of potential biases. For quantitative analyses, it must also be clarified which *response measures* (dependent variables) were used for analysis, which type of statistical test was used and how this relates to the study design. Findings should be reported, together with the *test statistics* and any other descriptive measures. Also, the results from any standardized or custom questionnaires as well as observations made and relevant responses to any interview questions should be reported as part of this aspect.

5 Related Work

This work intersects with both usability testing and information visualization evaluation. In both domains there is a considerable body of literature concerning best practices and challenges in the design of experimental evaluations [22,25]. However, user studies in SW contexts have specific characteristics that require tailored approaches for their evaluation.

Some recent works by the SW community have focused on building resources to foster the evaluation of user interfaces and interactive Semantic Web tools. A catalog of aggregated statistics on user interactions with over 200 BioPortal ontologies was recently released, containing information of user clicks, queries and reuse counts for over half a million users in a 3-year period [26]. Dragisic et al. [12] have created benchmarks that simulate different levels of user expertise to evaluate robustness of interactive ontology alignment systems. In [27], Ivanova et al. provide a set of requirements that foster the user involvement for large-scale ontology alignment tasks. Gonzalez et al. [28] developed a quality in use model for Semantic Web Exploration Tools (SWETs). A framework of exploration operations was proposed by Nunes & Schwabe [29]. Combining these operations would result in more complex exploration tasks. A follow-up work [30] then describes an analytical evaluation framework based on it. In [31], Garcia et al. present a benchmark for SW user interface evaluation that provides data, tasks and an environment to measure low-level performance metrics (keystrokes and clicks). All twelve tasks fall under the exploration and search categories, and are focused on the evaluation of SW browsers. This interest in supporting the evaluation of exploratory interactions matches the results of our literature review where the most popular operation supported by the systems was exploration.

6 Discussion

Nearly half of the papers we reviewed did not present a user study despite presenting a system or tool with support for user interaction and graphical user interface. Few of the works that did conduct a user study reported enough detail to allow an adequate interpretation or even the reproducibility of the experiments.

One challenge for experimental design for SW interactive evaluation is that many times tools need to support diverse users. A classical dichotomy is the Knowledge Engineer vs. the Domain Expert. Another dimension opposes SW novices/laymen to SW

experts. Of the 46 papers with user study reports we reviewed, only four conducted studies for SW experts and domain experts, and three for SW experts vs. novices. Designing experiments that take into account different areas and levels of expertise is necessary to support these types of tools. Determining expertise is another challenge, and only one study conducted a pre-assessment to stratify participants.

While ensuring both ecological and population validity is a concern in any user empirical evaluation, it is of particular importance in SW studies, where the impact of both user expertise and dataset characteristics can jeopardize the generalizability of conclusions.

One of the areas lacking detail was target user and participant description. Although demographical data was nearly always reported, recruitment strategies were generally not described. Getting students to evaluate systems is a common strategy, with well-known limitations [7]. Beyond bias issues, they represent a fairly homogeneous population that may not align well with the target users. We hypothesize that finding the right participants for the study may be one of the reasons behind the lack of user studies.

Another possible cause is lack of space in a publication. The thorough description of empirical evaluations requires a considerable amount of space, which can be difficult when faced with a page limit. When space is an issue, one might be forced to focus on the details most important to the outcome of the evaluation, and to those needed to enable the correct interpretation, replication, and comparison. However, the aim should always be to describe the conducted user evaluation as completely as possible, supporting the assessment and interpretation of the results. Although true reproducibility of user studies can be difficult—it is difficult enough to conduct a rigorous controlled experiment in one setting—providing a detailed description of the user study may allow for insightful comparisons between studies. This could, for instance, foster the evaluation of systems that have similar purposes and support similar operations, or the evaluation of the same system with different groups of users by a different research team. We would like to encourage SW researchers to make use of the publication of supplementary materials and other persistent data sharing options to provide detailed descriptions of their empirical evaluations if space does become an issue.

7 Conclusions

We have conducted a literature review of 87 papers published in Semantic Web venues between 2015 and 2017 that mention user interaction or visualization. Nearly half of these did not report on a user study, despite presenting approaches that supported user interaction. We classified the remainder according to the information they contained about the purpose and operations supported by the approaches under evaluation, study participants (number, expertise, recruitment) and evaluation methods employed.

The literature review served as the basis for charting the design space along six dimensions: i) purpose, ii) users, iii) tasks, iv) setup, v) procedure and vi) analysis and presentation of data. Based on these six dimensions, we proposed a protocol representing the minimum information required to report on a user study to ensure that it can be interpreted, compared and at best even replicated.

Our findings support our impression as VOILA organizers that comparatively few user studies are being conducted in SW contexts and that even fewer are reported adequately. However, the SW community seems to increasingly recognize the importance of evaluating interactive SW approaches, as indicated by the recent release of corresponding benchmarks and data collections. We hope that our discussion of the design space, and the framework composed by the guidelines, recommendations and reporting protocol we presented provides guidance and can foster the realization of more user studies for SW approaches with higher quality both in experimental design and in reporting. We aim as future work to validate the protocol by promoting its adoption within the VOILA community, a natural step in furthering the consolidation of user studies in SW contexts.

8 Acknowledgements

Catia Pesquita is funded by the Portuguese FCT through the LASIGE Strategic Project (UID/CEC/00408/2013), and also by FCT grant PTDC/EEI-ESS/4633/2014. Patrick Lambrix is funded by the Swedish e-Science Society (SeRC). Steffen Lohmann is partly funded by the Fraunhofer Cluster of Excellence Cognitive Internet Technologies (CIT).

References

1. McGrath, J.E.: Human-computer interaction. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1995) 152–169
2. Thomas, D.R.: A general inductive approach for analyzing qualitative evaluation data. American Journal of Evaluation **27**(2) (2006) 237–246
3. Mitschick, A., Nieschalk, F., Voigt, M., Dachselt, R.: IcicleQuery: A web search interface for fluid semantic query construction. In: 3rd International Workshop on Visualization and Interaction for Ontologies and Linked Data. Volume 1947 of CEUR Workshop Proceedings., CEUR-WS.org (2017) 99–110
4. Marchionini, G.: Exploratory search: From finding to understanding. Communications of the ACM **49**(4) (2006) 41–46
5. Vega-Gorgojo, G., Slaughter, L., Giese, M., Heggestøyl, S., Soylu, A., Waaler, A.: Visual query interfaces for semantic datasets: An evaluation study. Journal of Web Semantics **39** (2016) 81–96
6. Nuzzolese, A.G., Presutti, V., Gangemi, A., Peroni, S., Ciancarini, P.: Aemoo: Linked data exploration based on knowledge patterns. Semantic Web **8**(1) (2017) 87–112
7. Lazar, J., Feng, J.H., Hochheiser, H.: Research methods in human-computer interaction. Morgan Kaufmann (2017)
8. Hwang, W., Salvendy, G.: Number of people required for usability evaluation: the 10±2 rule. Communications of the ACM **53**(5) (2010) 130–133
9. Kittur, A., Chi, E.H., Suh, B.: Crowdsourcing user studies with mechanical turk. In: SIGCHI conference on human factors in computing systems, ACM (2008) 453–456
10. Marcolin, B.L., Compeau, D.R., Munro, M.C., Huff, S.L.: Assessing user competence: Conceptualization and measurement. Information Systems Research **11**(1) (2000) 37–60
11. Ziemkiewicz, C., Ottley, A., Crouser, R.J., Chauncey, K., Su, S.L., Chang, R.: Understanding visualization by understanding individual users. IEEE Computer Graphics and Applications **32**(6) (2012) 88–94

12. Dragisic, Z., Ivanova, V., Lambrix, P., Faria, D., Jiménez-Ruiz, E., Pesquita, C.: User validation in ontology alignment. In: 15th International Semantic Web Conference, Springer (2016) 200–217
13. Dadzie, A.S., Pietriga, E.: Visualisation of linked data – reprise. Semantic Web **8**(1) (2017) 1–21
14. Sarasua, C., Simperl, E., Noy, N., Bernstein, A., Leimeister, J.M.: Crowdsourcing and the semantic web: A research manifesto. Human Computation (HCOMP) **2**(1) (2015) 3–17
15. White, R.W., Roth, R.A.: Exploratory search: Beyond the query-response paradigm. Synthesis lectures on information concepts, retrieval, and services **1**(1) (2009) 1–98
16. Frøkjær, E., Hertzum, M., Hornbæk, K.: Measuring usability: are effectiveness, efficiency, and satisfaction really correlated? In: SIGCHI conference on Human Factors in Computing Systems, ACM (2000) 345–352
17. Huang, W., Eades, P., Hong, S.H.: Beyond time and error: A cognitive approach to the evaluation of graph drawings. In: 2008 Workshop on BEyond Time and Errors: Novel evaLuation Methods for Information Visualization - BELIV, ACM (2008) 3:1–3:8
18. Saraiya, P., North, C., Duca, K.: An insight-based methodology for evaluating bioinformatics visualizations. IEEE Trans. on Visualization and Computer Graphics **11**(4) (2005) 443–456
19. White, R.W., Drucker, S.M., Marchionini, G., Hearst, M., schraefel, m.c.: Exploratory search and HCI: Designing and evaluating interfaces to support exploratory search interaction. In: CHI '07 Extended Abstracts on Human Factors in Computing Sys., ACM (2007) 2877–2880
20. Fu, B., Noy, N.F., Storey, M.A.: Eye tracking the user experience–an evaluation of ontology visualization techniques. Semantic Web **8**(1) (2017) 23–41
21. Miles, M.B., Huberman, A.M., Saldana, J.: Qualitative data analysis. Sage (2013)
22. Rubin, J., Chisnell, D.: Handbook of usability testing: how to plan, design, and conduct effective tests. John Wiley & Sons (2008)
23. Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C.A., Causton, H.C., Gaasterland, T., Glenisson, P., Holstege, F.C., Kim, I.F., Markowitz, V., Matese, J.C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J., Vingron, M.: Minimum information about a microarray experiment (miame) – toward standards for microarray data. Nature genetics **29**(4) (2001) 365–371
24. Field, A., Hole, G.: How to Design and Report Experiments. Sage (2003)
25. Plaisant, C.: The challenge of information visualization evaluation. In: Working conference on Advanced visual interfaces, ACM (2004) 109–116
26. Kamdar, M.R., Walk, S., Tudorache, T., Musen, M.A.: Bionic: A catalog of user interactions with biomedical ontologies. In: 16th International Semantic Web Conference, Springer (2017) 130–138
27. Ivanova, V., Lambrix, P., Åberg, J.: Requirements for and evaluation of user support for large-scale ontology alignment. In: 12th European Semantic Web Conference, Springer (2015) 3–20
28. González Sánchez, J.L., García González, R., Brunetti Fernández, J.M., Gil Iranzo, R.M., Gimeno Illa, J.M.: Using swet-qum to compare the quality in use of semantic web exploration tools. Journal of Universal Computer Science **19** (2013) 1025–1045
29. Nunes, T., Schwabe, D.: Frameworks for information exploration-a case study. In: 4th International Workshop on Intelligent Exploration of Semantic Data - IESD. (2015)
30. Nunes, T., Schwabe, D.: Frameworks of information exploration-towards the evaluation of exploration systems. In: 5th International Workshop on Intelligent Exploration of Semantic Data - IESD. (2016)
31. García, R., Gil, R., Gimeno, J.M., Bakke, E., Karger, D.R.: Besdui: A benchmark for end-user structured data user interfaces. In: 15th International Semantic Web Conference, Springer (2016) 65–79