A Protocol for KG Generation Tasks Involving Users

Ademar Crotti Junior, Christophe Debruyne

Montefiore Institute, University of Liège, Belgium

Abstract

Knowledge graph construction (KGC) from (semi-)structured data is challenging, and facilitating user involvement is an issue frequently brought up within this community. We cannot deny the progress we have made with respect to (declarative) knowledge generation languages and tools to help build such mappings. However, it is surprising that no two studies report on similar protocols. This heterogeneity does not allow for comparing KGC languages, techniques, and tools. This paper first analyses the various studies that report on studies involving users to identify the points of comparison. These gaps include a lack of systematic consistency in task design, participant selection, and evaluation metrics. Moreover, there needs to be a systematic way of analyzing the data and reporting the findings, which is also lacking. We thus propose and introduce a user protocol for KGC designed to address this challenge. Where possible, we draw and take elements from the literature we deem fit for such a protocol. The protocol, as such, allows for the comparison of languages and techniques for the RDF Mapping Languages core functionality, which is covered by most of the other state-of-the-art techniques and tools. We also propose how the protocol can be amended to compare extensions (of RML). This protocol provides an important step towards a more comparable evaluation of KGC user studies.

URL: https://github.com/chrdebru/kgc-user-study-protocol

Keywords

KG Construction, User studies, Research Methods

1. Introduction

We preach to the choir that knowledge graphs are essential for meaningfully organizing and representing information in various domains. However, as knowledge graphs grow in complexity, efficient methods for their generation are crucial. When dealing with the challenges of (semi-)structured data sources, such as the lack of explicit semantics, which need to be aligned with ontologies or vocabularies, creating such mappings becomes a *knowledge engineering task* where *user involvement is crucial*. Users bring the necessary domain expertise to ensure the mappings are appropriate.

Scholars have systematically analyzed the functionalities of Knowledge Graph Construction (KGC) tools and proposed benchmarks to analyze their behavior in different settings and their memory and CPU usage. It is thus surprising that user involvement and the perception of the user using the languages, tools, etc., have yet to be studied in such detail. Conducting such a study for all languages and tools is an infeasible undertaking for one group, but what is feasible is putting forward a protocol that scholars in the domain should adopt to report on user studies. This paper aims to achieve this goal by proposing a user study protocol for KGC.

This will enable researchers to compare different knowledge graph construction (KGC) languages and techniques, leading to a better understanding of their strengths and weaknesses and, ultimately, to more effective tools for KGC.

The contributions of this paper are twofold. First, we have a review of user studies in the KGC domain, which indicates the challenges mentioned above. And the protocol we present in this paper.

Section 2 provides an overview of the related work on (declarative) approaches to KGC by mapping data sources onto RDF datasets, focusing on those that explicitly report on user studies. The goal is to show that no two papers adopt the same protocol, which makes comparing studies impossible.

Section 3 presents the protocol we have made available with CC-BY-SA 4.0 license. The protocol provides detailed guidelines for recruiting participants and disclosing potential biases. The process

KGCW'25: 6th International Workshop on Knowledge Graph Construction, June 1 or 2, 2025, Portorož, Slovenia

C.Debruyne@uliege.be (C. Debruyne)

0000-0003-1025-9262 (A. C. Junior); 0000-0003-4734-3847 (C. Debruyne)

guidelines for informed consent, pre-questionnaires, familiarization activities, task execution, and post-questionnaires. The tasks consist of five sub-tasks, of which, when comparing two groups, the last two can be changed to ensure a common base for comparison. The focus of the protocol is to facilitate the discovery of problems (i.e., formative testing), not task-level measurements. [1] describes the difference between the two. The protocol does propose task-level measurements and techniques to analyze the data. When sample sizes are small, they can merely give insights.

The related work will show that reporting is often limited to simple metrics and averages. Still, we deem it important to analyze the relationships between task execution, perceived usefulness, and perceived cognitive load. To this end, Section 4 proposes the statistical means to use when adopting this protocol.

In Section 5, we discuss the resources from various aspects, such as the scientific and technical, to elaborate on the soundness of our approach. This section also discusses some of the limitations. Section then concludes the paper and proposes some future directions.

2. Related Work

In [2], the authors presented an excellent survey on declarative KGC tools to help the community and practitioners choose which languages, tools, or techniques fit their needs. However, the article looks at those from a technical perspective. They look at the functionalities offered by the different options. In [3], the authors proposed a benchmark to compare KGC tools and applied it to some well-known implementations such as RMLMapper [4], Morph-KGC [5], and SDM-RDFizer [6]. It is surprising to see that the perceptions of users and practitioners have yet to be examined in a systematic manner.

From a broader perspective, [7] describes three ”personas” that engage with KGs: KG builders, KG analysts, and KG consumers, which were distilled from interviews with practitioners. As the name intuitively implies, the KG builder persona would be responsible for generating the KG from heterogeneous sources, but the persona is also in charge of ontology engineering. The authors state that builders could benefit from tools that help them ensure that the schema is respected (what the authors call an ”enforcer”) as well as adequate visualization tools. While the paper does not explicitly mention KG generation and mappings as tasks, they fall under the ”data integration” umbrella. Key is here is that the interviews indicate that there are challenges impeding uptake.

It seems that practitioners' or users' roles are sometimes neglected. This is certainly the case for KG Generation, as we will now demonstrate via our literature review. Our review looked at the following papers reporting on users, their experiences, and/or perceived usability: [8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18]. Most of these studies looked at the creation of mappings. Exceptions are [15] reporting on studies on mapping understanding and [18] reporting on a complex data flow that included mapping creation.[1] We compare the various aspects of these user studies in Tables 1, 2, and 3. From these tables, we can observe a couple of important points:

- Some report on comparing mapping languages and/and tools (e.g., [13] and [14]), and others report on comparing mapping languages (e.g., [16] and [17]). Quite a few papers merely report on the perceived usability of their tool without any comparison. We argue that reporting on user studies only makes sense if there is a basis for comparison.
- Looking at the procedure, we see many recurring elements (some (training) resources are being shared, pre-assessment surveys, introduction of tasks, surveys, etc.). No two procedures are the same, which hampers our ability to compare the studies. Some studies reported asking about information such as gender and age but did not report on those in the data analysis.
- Most studies involved participants with expertise in it, databases and/or semantic web technologies. Many studies also report inviting MSc students in computer science or related fields. Self-

[1]Excluded from this survey are publications that do not report on users. For example, in [19], the authors reported involving participants, but no report on the participants' experience was made. Other examples of papers mentioning users, participants, etc. without any detailed reporting include [20], [21], and [22].

reported prior knowledge and competencies are a recurring theme, but no two studies tackle this aspect comparably.
- The same heterogeneity can be observed for the tasks and datasets, where we do notice that most studies adopt datasets that do not require specific domain expertise (people, movies, places, etc.).
- Recurring themes in data being analyzed are time (efficiency), accuracy, and perceived usability. Most rely on System Usability Metrics [23] (SUS) for perceived usability. A few studies rely on Post-Study System Usability Questionnaire [24] (PSSUQ) to obtain information on perceived information quality, interface quality, and system usefulness. Such studies do allow a means to compare results. Few studies have reported on qualitative feedback from users and the mental workload of tools and mapping languages.
- Most studies merely report on averages, which can arguably make sense when authors only report on one group and tools or languages are not compared. Few studies employed techniques to analyze whether groups are (significantly) different or whether certain aspects had a (statistically significant) impact on efficiency, accuracy, etc.

From this survey, we can conclude that there is a critical need for homogeneous protocols, including tasks, for comparing advances in KGC approaches (mapping languages and tools alike). In the next section, we propose a protocol to address these issues and how they can be used.

3. The KGC User Study Protocol

This section presents the protocol for comparing KGC tools and languages. The protocol's[2] structure foresees placeholders for text to be easily adapted for research ethics applications.

The protocol can be used to analyze the perceived usability and cognitive load of a mapping language, tool, ... as well as the accuracy achieved by users and the task execution time. When users use this protocol on only one group, the hypotheses are limited to comparing participants with different demographics, or by comparing the results with other experiments using this protocol. Scholars adopting this protocol can easily extend the protocol to compare two tools, techniques, or languages. This will be explained in Section 3.5.

The focus of the protocol is to facilitate the discovery of problems (i.e., formative testing), not on task-level measurements. [1] describes the difference between the two. The protocol proposes task-level measurements and techniques for analyzing the data. When sample sizes are small, they can merely give insights.

3.1. Participant Selection

Adopters of the protocol should indicate how participants are recruited and from where. Adopters should disclose potential biases by providing details about factors influencing the study or participant behavior. Examples include the hierarchical relationship between research group leaders and researchers, as well as students recruited from classes. There is also a difference between voluntary participation and mandatory representation (e.g., in the context of a teaching activity).

Practitioners involved with UI and UX often state that five users are sufficient to discover most of the usability problems. That is more likely the case for problem discovery than task-level measurements, which requires larger sample sizes [1]. As [25] observed in an experiment, *"increasing the number from 5 to 10 can result in a dramatic improvement in data confidence."* They also found that increasing the number to 20 practically guarantees all problems to be seen, but we recognize that recruiting participants is difficult. As such, we require 10 participants per group.

[2] https://github.com/chrdebru/kgc-user-study-protocol

Table 1
Comparison of KGC languages, techniques, and tools in which participants were observed. We can observe that procedures vary widely.

Reference	Tool	Language	Comparison	Procedure
Pinkel et al., 2014	Unnamed tool	R2RML	No comparison to other tools.	- self-assessment on DB and SW concepts - briefing into R2RML and the domain - task
Heyvaert et al., 2016	RMLEditor	RML	No comparison to other tools.	- brief introduction to SW, and UI tool - self-assessment (on what unclear) - pre-assessment on expected usability - task - usability
Sicilia et al., 2017	Map-On	R2RML	No comparison to other tools.	- 50 minutes study - brief introduction to SW concepts, RDF, R2RML and tool - self-assessment on DB and SW expertise - task - usability (SUS)
Bak et al., 2017	SQuaRE	R2RML	No comparison to other tools.	- 1 hour study - briefing about the tool including tutorial - task with SPARQL queries to evaluate provided mappings) - open questions defined by authors
Crotti Junior et al., 2017	Juma R2RML	R2RML	No comparison to other tools.	- self-assessment on SW and R2RML - briefing about RDF, R2RML and the tool - task - usability
Heyvaert et al., 2018	MapVOWL in RMLEditor	RML	Comparison between visual representation and RML.	- self-assessment on SW and demographics - tasks - questionnaire (user preference, confidence)
Ibid.	RMLEditor, RMLx Visual editor	RML	Comparison between RMLEditor (graphical) vs RMLx Visual editor (form based)	- self-assessment on SW and demographics - tasks - additional information - another task - usability - questionnaire (user preference, confidence)
Crotti Junior et al., 2018	Juma Uplift	R2RML	Comparison between tool and R2RML.	- 1 hour study - self-assessment on SW and R2RML - briefing about RDF, R2RML and the tool - task - usability and mental workload
Crotti Junior, 2019	Juma Uplift	R2RML	Comparison between tool and R2RML.	- self-assessment on SW and R2RML - briefing about RDF, R2RML and the tool - task - usability and mental workload
Garcia-Gonzalez et al., 2020	n/a	SheXML, YARRRML, SPARQL-Generate	Comparison of SheXML, YARRRML, SPARQL-Generate.	- printed manual of tool is given to participants for 20min - start mouse and keyboard software - task - nonstandard evaluation questionnaire
Warren et al., 2023	n/a	YARRRML, SPARQL Anything	Comparison YARRRML and SPARQL Anything.	- 1 hour study - tutorial and material before the experiment (days before) - self-assessment on SW and mapping languages used, and demographics - tasks - nonstandard evaluation questionnaire
Brouwer et al., 2024	RMLEditor, Matey	RML, YARRRML	Developers use Matey, non-experts use RMLEditor.	- self-assessment on SW - briefing about technologies and tools - task - usability plus some specific questions

3.2. Process

Participants begin by reviewing informed consent materials and completing a pre-questionnaire assessing their demographics, prior knowledge, and expectations. Next, they attend a presentation intro-

Table 2
Comparison of KGC languages, techniques, and tools in which participants were observed. We can observe that no two papers share tasks and data analysis techniques.

Reference	Dataset	Tasks	Analysis	Statistical measures
Pinkel et al., 2014	Music Brainz and Music ontology	3 mapping tasks divided in steps	time (actual and perceived) accuracy	Averages
Heyvaert et al., 2016	Exployees and projects. Movies (DBpedia) and directors.	2 use cases, unclear how many mappings. One schema based (start from ontology) and one data based (start from data).	accuracy usability (SUS) qualitative feedback	Averages
Sicilia et al., 2017	Derived from RODI benchmark	3 tasks of low, medium and high complexity as described by the authors	time accuracy	Averages
Bak et al., 2017	Movie Ontology and IMDB Movie Ontology	10 mappings, 2 classes/3 object properties/5 data type properties. (Maybe this means 1 mapping task in several steps but not clear).	completion rate (# people who completed tasks) qualitative feedback	Number of prople who completed tasks.
Crotti Junior et al., 2017	Northwind database	1 mapping in 3 parts (2 classes and linking them)	time accuracy usability (PSSUQ)	Averages on accuracy and time Welch Two Sample t-test and Friedman non-parametric test for PSSUQ
Heyvaert et al., 2018	Cars and people/twitter/papers	understanding the mapping i.e. what is being mapped	accuracy (11 questions on what is being mapped) users' preference and confidence questions	Averages
Ibid.	Employees and projects. Directors and movies.	2 use cases, one per dataset.	accuracy usability (SUS) qualitative feedback	Averages
Crotti Junior et al., 2018	Northwind database	1 mapping in 3 parts (2 classes and linking them)	accuracy usability (PSSUQ) mental workload (MWL and NASA-TLX)	Averages on accuracy ANOVA, Anderson darling, Welch and Wilcoxon tests Normality tests Correlations with Pearson and Spearman Reliability
Crotti Junior, 2019	People and places	1 mapping on each representation relating people and places	time accuracy usability (PSSUQ) mental workload	Averages on accuracy ANOVA, Anderson darling, Welch and Wilcoxon tests Normality tests Correlations with Pearson and Spearman Reliability
Garcia-Gonzalez et al., 2020	Movies	2 tasks	mouse and keyboard activity a modified usability questionnaire accuracy time	Average, standard deviation, min and max values One way ANOVA Kruskal-Wallis test
Warren et al., 2023	Art (Tate modern)	fill gaps and partial solutions were provided	accuracy (manually)	Subjective evaluation when analyzing users' results
Brouwer et al., 2024	heath care	task was on the overall workflow and not directly on mapping	accuracy usability (SUS)	Averages

ducing the technology, review relevant documentation, and engage in a familiarization activity. The core of the study involves participants executing a defined task using the technology. Finally, participants complete a post-questionnaire evaluating their experience, including usability, efficiency, and perceived cognitive load. This structured process aims to gather comprehensive data on user interac-

Table 3
Comparison of KGC languages, techniques, and tools in which participants were observed. We can observe that self-reported prior knowledge and skills, and participant selection varies widely. Also, the number of participants varies.

Reference	Participant selection	Participant background	Skill assessment	# participants
Pinkel et al., 2014	Invitation	DB experts, SW experts, and non experts. (47 participants but 16 quit during experiment). 13 experts in SW and 11 in DB	self-assessed using Likert scale of 1 to 5. 4 and 5 considered experts.	31
Heyvaert et al., 2016	Invitation (?) from within university and research group	10 SW experts and 5 non experts.	Likely self-assessed. Authors mentions a pre-questionnaire which is unavailable.	15
Sicilia et al., 2017	Selected to have similar profile.	DB experts but not SW experts. 2 professors, 2 from industry and 1 from research	Self-assessed pre-questionnaire on DB and SW topics.	5
Bak et al., 2017	Invitation (Students)	5 MsC students in SW. 1 PhD student with SW general knowledge	Reports that participants were enrolled in MsC or PhD in SW area.	6
Crotti Junior et al., 2017	Invitation	5 Web developer, 5 SW and 5 R2RML experts	self-assessed using Likert scale of 1 to 7. 1 to 4 is considered familiar with technology.	15
Heyvaert et al., 2018	RML experienced users were contacted	SW practitioners	self-assessed.	9
Ibid.	RML experienced users were contacted	SW practitioners	self-assessed.	10
Crotti Junior et al., 2018	Invitation (Students)	MsC students in SW. 12 using tool and 14 using text editor.	Students of SW.	26
Crotti Junior, 2019	Online (email and mailing lists)	SW experts, R2RML experts, and non experts	self-assesed using Likert scale of 1 to 7. 1 to 4 is considered familar with technology.	95
Garcia-Gonzalez et al., 2020	Invitation (?) from within university and research group	20 students of MsC in Web Engineering. 3 left experiment in task one, 13 left in task 2	Students of SW.	20
Warren et al., 2023	Invitation from W3C group KGC and SPARQL 1.2	SW practitioners	SW practitioners	18
Brouwer et al., 2024	Invitation within university.	SW practitioners	SW practitioners	8

tion and perception of the technology.

Next to presenting and demonstrating the tool or mapping language, we also request participants to handle the environment. We deem this familiarization activity novel compared to the related work. This activity ensures participants are comfortable executing mappings within the provided (tool's) environment. We guide participants in demonstrating the practical aspects of using the tool's interface, such as utilizing the command line in the terminal or identifying the correct buttons to click. This focused familiarization will prevent the environment from becoming an obstacle, allowing us to assess the tool or language's usability and gather unbiased feedback on its functionalities.

Furthermore, we ask authors to report on how responses were submitted (e.g., email, paper, form) and the anticipated duration of the experiment. While in-class experiments often have time constraints, other environments may be more flexible. Our protocol foresees 1 hour for the five tasks. If, for example, all steps are conducted in a classroom setting, the experiment would require 2:30. Finally, clarify whether participants will be allowed to ask questions during the experiment. Help should be limited to aspects not core to the KG generation process and experiment. For example, helping

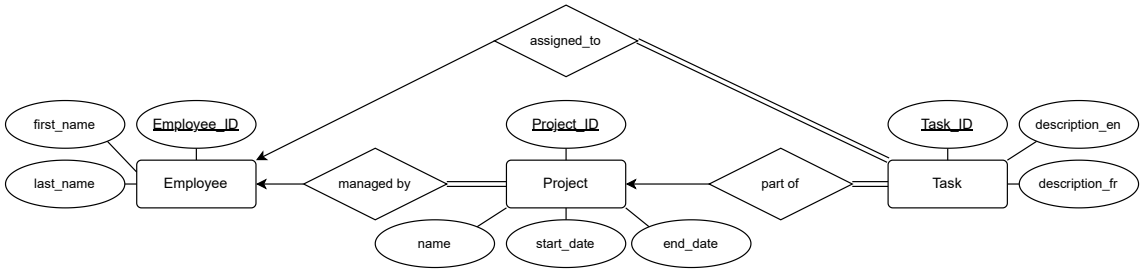

Figure 1: Enter Caption

participants navigate to the correct directory in a terminal or assessing whether a network issue is permitted, but providing help to execute a mapping is not. Ideally, studies would report on those (and their number of occurrences).

3.3. Pre-questionnaire

Studies often inquire about participants and group them based on self-reported information on their background and proficiency with specific techniques. We aim to homogenize this by proposing an exhaustive list of current roles, formal training, and self-perceived competency levels in certain Semantic Web technologies. We also included three questions related to intrinsic motivation (enjoyment, curiosity, and value). It is important to analyze the impact of self-selection bias in voluntary participation.

3.4. Mapping tasks

Participants will be requested to complete five mapping tasks, some of which are interdependent. In our state-of-the-art surveys, many different types (e.g., mapping understanding vs. mapping creation) are applied in many domains. We propose choosing a general domain for participants to understand: projects, project tasks, and employees who manage projects and are assigned tasks. The rationale is that there are three distinct types, and users should, at this point, not struggle with ontological questions such as roles vs. types.

Figure 1 depicts the UoD of the data to be transformed into RDF. We use an ERD, but JSON and XML files can easily represent the data. To ensure attribute names do not mislead participants, we ensured all attributes are unambiguous. In this simple UoD, all relations are many-to-one, though this can be easily extended to many-to-many when transforming documents.

The tasks can be summarized as follows:

1. Generate instances of `ex:Employee` with their first and last-names. The IRIs of employees are based on the name.
2. Generate instances of `ex:Project` with their name, start- and end-date. Both dates are of the type `xsd:date`, allowing us to assess the creation of typed literals. The IRIs of projects are based on the project's ID.
3. Generate `ex:managedBy` properties from projects to employees.
4. Generate instances of `ex:Task` with their descriptions (in two languages). The IRIs of tasks are based on the task's ID. The descriptions allow us to assess the creation of language tags.
5. Generate `ex:of` and `ex:assignedTo` properties from tasks to, respectively, projects and employees.

We point out that the mappings mainly focus on RML-core [22] functionality. Part of RML-core are multi-valued expression maps, which are irrelevant for CSV files. People can easily adapt the JSON files to provide one or more task descriptions in different languages to test more complex language maps, for instance.

We draw attention to the fact that employees' IRIs are based on their names, whereas the project data sources refer to employees via their IDs. This requires users to join the data in the two sources. In the case of RML, this requires users to "join" the two sources either at the level of the logical source or by using a referencing object map.

3.5. Variants

We stated that the tasks focus on RML-core functionality, which is also covered by other KGC languages and techniques such as ShExML [16] and SPARQL Anything [26]. While one can argue that the set of desired functionalities is limited, [22] covered the requirements of all RML extensions. It is not feasible to formulate tasks that cover practically all use cases, not only in time but also in data source complexity.

In this section, we describe how this protocol can be used for comparing different aspects.

- When comparing languages, techniques, or tools. One can provide the same tasks to two different groups. One can compare different mapping languages (e.g., ShExML vs. RML), compare editors vs. "bare bones" languages (e.g., RMLEditor vs. RML), compare languages and abstractions of languages (e.g., RML vs. YARRRML), and even different editors and languages.
- When the aim is to compare support for an advanced KGC requirement such as named graphs, collections and containers, or RDF* (among others), then one can take this protocol as is for one group, and only change the last two tasks for the second in which those requirements are covered, with the first three tasks giving a comparable baseline, unless the extension covered another aspect such as storing the resulting trip.
- When the aim is to such compare KG construction across languages, techniques, or tools, then the protocol must be amended for both. Again, the first three tasks must remain unchanged as to ensure some comparative baseline with literature. This approach can be used to compare the languages and tools for more advanced KGC requirements such as named graphs, RDF collections and containers, and function calls, among others.

Participants are assumed to have access to prepared "resources" or "environments" to focus on the tasks. In the case of RML, for instance, the logical sources would be provided in the tool or for them to copy and paste. This allows researchers to assess the languages and tools with respect to these aspects by giving one group the prepared artifacts and requesting the other to formulate the logical sources themselves. We deem this a special case of comparing a baseline with an extension as described above, but where the five tasks remain unchanged.

3.6. Post-questionnaire

Both SUS and PSSUQ are used to measure usability. SUS is adequate for a rapid and general measure of a system's usability. Still, the latter offers more advantages because it assesses three aspects of a system: system usefulness, information quality, and interface quality. Furthermore, there is a question about the system as a whole, which allows one to damped the perception of individual aspects. The original PSSUQ survey uses 19 questions (as adopted by [12], for instance), but recent iterations have removed three redundant questions.

As for the perceived (mental) workload, we adopt both the Workload Profile (WP) [27] and the NASA Task Load Index (NASA-TLX) [28].

- WP adopts a theory in which participants have different capacities (dimensions) related to the stage, mode, input, and output of information processing. The eight dimensions are each quantified through subjective rates, and participants must rate the proportion of attentional resources used for performing a given task with a value from 0 to 100 after task completion. A rating of 0 means that the task placed no demand, while 100 indicates that it required maximum attention. The WP of a participant is calculated as $\frac{1}{8}\sum_{i=1}^{8} d_i$.

- NASA-TLX has been validated in several domains [28] and combines six factors believed to influence the mental workload. Each factor is quantified with a subjective judgment and a weight computed via a paired comparison procedure. For each possible pair of the six factors, participants must decide which factor contributed the most to the mental workload during the task. The weights w are the number of times each dimension was selected. The possible weights range from 0 (irrelevant) to 5 (most important). The final score is computed as a weighted average, considering the subjective rating of each attribute d_i and the correspondent weights w_i: $\frac{1}{15}\sum_{i=1}^{6} d_i * w_i$. It is possible to calculate the scores by eliminating the weighted procedure, which yields the so-called *Raw TLX*.

Both instruments are used in industry and research. You may notice that both instruments use different rating systems, which may confuse participants in a paper survey. Erroneous inputs can be prevented by adopting electronic forms. We choose not to harmonize the scales, which has been done in [12], for instance, to obtain results that can be compared to other studies faithfully adopting those instruments.

Studies should explicitly report the method used to calculate performance measures, such as accuracy. For instance:

- Accuracy should be determined by *graph isomorphism* (did they generate the expected graph, which is true or false), and, for more nuanced numbers, *precision* (the proportion of triples that are generated that are in the expected graph), *recall* (the proportion of expected triples that are in the generated graph), and *F-measure* combining precision and recall.

This approach accounts for situations where a participant generates additional triples, for instance. Figures should be reported both on task and global level.

- Efficiency is often erroneously conflated with *task execution time*. The former is a broad concept that measures how well resources are used, which is not only time. Most studies measured the time it took for tasks to be completed. We require studies to report on task execution time, and the method to measure it. One can manually record time or use software to record user interactions to time the tasks. Another approach is to request users to report on the time or use electronic forms that keep track of time.

While we strongly encourage placing a time limit on the tasks for the experiment to obtain comparable results, there are two important cases to track: *did not finish* and *did not start*. The former may indicate insufficient time left to finish a task or that the task was too difficult. The latter merely indicates that the user never started the task.

We propose to limit the protocol to these five metrics (four on accuracy and one related to task execution time). Studies are free to include other metrics, such as the number of times a mapping was executed (i.e., trial and error), but that would indirectly impact the task execution time.

As the experiment should not be too time-consuming, we avoided interviews to obtain qualitative feedback. We also avoid ”think aloud” experiments as they can impact the cognitive load. Whether a study reports on it or not, we require studies to report on any additional instruments they used. There is, however, a qualitative dimension to our protocol, as the PSSUQ does leave room for comments on each of the 16 questions.

4. Results and Analysis

As part of our protocol, we recommend a structured approach for reporting collected data. As mentioned, while many user studies focus primarily on presenting averages and standard deviations, we emphasize the importance of extending these reports to include statistical tests. This ensures robust comparisons between groups and tools, enhancing the general reliability and interpretability of the experiments. The following describes the recommended aspects and tests to be considered when reporting results and analysis.

Reliability and Internal Consistency Reliability refers to the degree to which the items within a test or survey consistently measure the same construct. High internal consistency strengthens the statistical reliability of metrics, thereby enhancing the validity of group comparisons. We recommend using *Cronbach's Alpha* to evaluate internal consistency. Higher alpha coefficients indicate greater shared covariance among items, suggesting they assess the same underlying concept. A Cronbach's Alpha value of ≥ 0.7 is generally considered acceptable. [29]

Data Normality Data normality refers to how much data distribution aligns with a normal curve. While normality is not always required, t-tests and ANOVA that assume data follows such a distribution. ANOVA is relatively robust when data is not normally distributed, when sample sizes are large, but that is difficult when dealing with user studies. We, therefore, require studies to test for normality and report on normality. Participants may, if they wish, use other statistical measures. As the sample sizes of a group will likely not exceed 50, we propose the *Shapiro-Wilk test* to assess normality. In this test, the sample is compared to a theoretical normal distribution.

Homogeneity Some statistical tests, such as ANOVA, assume that the variances across groups are equal. This is known as the *homogeneity of variances*. Again, this assumption should be verified as part of the analysis. *Levene's Test* is a standard method for evaluating this assumption.

Group Comparisons To determine whether differences between groups are statistically significant, researchers should employ well-established statistical tests. The choice of test depends on the assumptions about the data. The recommended tests when the data is normally distributed (also known as parametric tests) are *Welch's t-test* when comparing two groups and *ANOVA* when comparing more than two groups simultaneously. Both tests assume normality and homogeneity of variance. The recommended tests when the data is not normally distributed (also known as non-parametric tests) are the *Wilcoxon test* for comparing two groups and the *Kruskal-Wallis test* for multiple.

Correlation Analysis Correlation methods assess the strength and direction of the relationship between variables, which can provide deeper insights into study outcomes. For instance, examining correlations between usability, accuracy, and mental workload can reveal relationships in user behavior. When data is normally distributed, we recommend the *Pearson's Correlation* to measure the strength of a relationship, for example, between accuracy and usability. Otherwise, one should use the *Spearman's Correlation*. Researchers must report on correlations between all relevant variables (e.g., usability and mental workload, usability and accuracy, etc.) to provide a comprehensive analysis.

Transparency and Accessibility To promote transparency and reproducibility, all collected data and the statistical tests performed should be publicly available online. Providing access to raw data, analysis scripts, and detailed methodology facilitates validation and enhances the study's credibility. Moreover, sharing such data would allow one to compare results across studies more easily, provided the conditions are similar.

5. Discussion

We presented a comparison of user studies in the KGC domain, and we noticed that all studies were different. This makes it impossible to compare KGC languages, tools, and software. To this end, we analyzed the related work, distilled elements we appreciated, and proposed others to establish a common protocol. As such, we proposed a resource, a user study protocol, that provides the KG generation community with a better way to present, compare, and scrutinize contributions.

When designing the protocol, we selected and refined elements from the state-of-the-art that we appreciated. Examples include the use of accuracy and task execution time as simple measures, the use of PSSUQ over SUS to obtain more fine-grained information on usability, usefulness, and information

quality, and measuring the mental workload right after the task. Reporting on the correlations between the perceived usability, task execution, and mental workload could shed interesting insights. [14] We furthermore provided guidelines on what statistical techniques to use when reporting on user studies with this protocol, as most merely reported on averages.

Somewhat new compared to the related work is our informed decision to adopt an accessible domain, focus on RML-core functionality as a basis, and formulate five tasks. In Section 3, we provided a rationale that, when comparing two groups, the last two tasks can be replaced for one group so that extensions or variants within the same language or tool can be compared. As such, the resource is sufficiently general for use in the KGC community, and the approach in designing this protocol may inspire others within the Semantic Web community.

The resources have not only been made available with a DOI on a long-term preservation platform, but they are also available on a GitHub repository. The latter allows peers to contribute to the project. The directory structure we use allows for variants of the protocol to be made available.[3]

The protocol has not yet been used at this stage, but its first uses are planned in the spring of 2025. However, many of the protocol's separate elements are drawn from existing studies. As such, those parts have already been validated in the community. We also aim to engage with the wider KGC community via the W3C working group on adopting this protocol across different institutions.

6. Conclusions

Prior KGC user studies used different protocols, making comparison impossible. This paper thus highlights the lack of standardized protocols in user studies related to KGC, making it difficult to compare different studies. We present a new protocol to address these inconsistencies, focusing on participant selection, task design, and evaluation metrics. The protocol suggests detailed guidelines for recruiting participants and disclosing potential biases. The process guidelines for informed consent, pre-questionnaires, familiarization activities, task execution, and post-questionnaires. Five specific mapping sub-tasks are proposed, which can be solved with the equivalent of the RML-Core specification. The protocol recommends using the Post-Study System Usability Questionnaire (PSSUQ) for usability, and the NASA Task Load Index (NASA-TLX) for mental workload, among others. We designed the protocol in such a way that one can analyze one group, or compare groups. To this end, we provided guidelines on which statistical instruments to use.

This protocol aims to provide a more comparable evaluation of KGC user studies, ultimately leading to more effective tools for knowledge graph construction. As such, the protocol is an essential artifact for future longitudinal and comparative studies. Future work involved encouraging the adoption of this protocol by various KGC scholars. While ambitious, it is hoped that this protocol will form the basis of a new, open repository of KGC user studies.

Declaration on Generative AI

During the preparation of this work, the author(s) used Grammarly to improve grammar, check spelling, and reword. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] J. R. Lewis, Sample sizes for usability tests: mostly math, not magic, Interactions 13 (2006) 29–33. URL: https://doi.org/10.1145/1167948.1167973. doi:10.1145/1167948.1167973.

[3] For the reviewers: at the time of writing, we are currently discussing with the W3C KGC Working Group whether the initiative could be moved to https://github.com/kg-construct/.

[2] D. V. Assche, T. Delva, G. Haesendonck, P. Heyvaert, B. D. Meester, A. Dimou, Declarative RDF graph generation from heterogeneous (semi-)structured data: A systematic literature review, J. Web Semant. 75 (2023) 100753. URL: https://doi.org/10.1016/j.websem.2022.100753. doi:10.1016/J.WEBSEM.2022.100753.

[3] D. V. Assche, D. Chaves-Fraga, A. Dimou, KROWN: A benchmark for RDF graph materialisation, in: G. Demartini, K. Hose, M. Acosta, M. Palmonari, G. Cheng, H. Skaf-Molli, N. Ferranti, D. Hernández, A. Hogan (Eds.), The Semantic Web - ISWC 2024 - 23rd International Semantic Web Conference, Baltimore, MD, USA, November 11-15, 2024, Proceedings, Part III, volume 15233 of *Lecture Notes in Computer Science*, Springer, 2024, pp. 20–39. URL: https://doi.org/10.1007/978-3-031-77847-6_2. doi:10.1007/978-3-031-77847-6_2.

[4] A. Dimou, M. V. Sande, P. Colpaert, R. Verborgh, E. Mannens, R. V. de Walle, RML: A generic language for integrated RDF mappings of heterogeneous data, in: C. Bizer, T. Heath, S. Auer, T. Berners-Lee (Eds.), Proceedings of the Workshop on Linked Data on the Web co-located with the 23rd International World Wide Web Conference (WWW 2014), Seoul, Korea, April 8, 2014, volume 1184 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2014. URL: https://ceur-ws.org/Vol-1184/ldow2014_paper_01.pdf.

[5] J. Arenas-Guerrero, D. Chaves-Fraga, J. Toledo, M. S. Pérez, Ó. Corcho, Morph-kgc: Scalable knowledge graph materialization with mapping partitions, Semantic Web 15 (2024) 1–20. URL: https://doi.org/10.3233/SW-223135. doi:10.3233/SW-223135.

[6] E. Iglesias, S. Jozashoori, D. Chaves-Fraga, D. Collarana, M. Vidal, Sdm-rdfizer: An RML interpreter for the efficient creation of RDF knowledge graphs, in: M. d'Aquin, S. Dietze, C. Hauff, E. Curry, P. Cudré-Mauroux (Eds.), CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020, ACM, 2020, pp. 3039–3046. URL: https://doi.org/10.1145/3340531.3412881. doi:10.1145/3340531.3412881.

[7] H. Li, G. Appleby, C. D. Brumar, R. Chang, A. Suh, Knowledge graphs in practice: Characterizing their users, challenges, and visualization opportunities, IEEE Transactions on Visualization and Computer Graphics 30 (2023) 584–594. URL: https://doi.org/10.1109/TVCG.2023.3326904. doi:10.1109/TVCG.2023.3326904.

[8] C. Pinkel, C. Binnig, P. Haase, C. Martin, K. Sengupta, J. Trame, How to best find a partner? an evaluation of editing approaches to construct R2RML mappings, in: The Semantic Web: Trends and Challenges - 11th International Conference, ESWC 2014, Anissaras, Crete, Greece, May 25-29, 2014. Proceedings, volume 8465 of *Lecture Notes in Computer Science*, Springer, 2014, pp. 675–690.

[9] P. Heyvaert, A. Dimou, A. Herregodts, R. Verborgh, D. Schuurman, E. Mannens, R. Van de Walle, Rmleditor: A graph-based mapping editor for linked data mappings, in: The Semantic Web. Latest Advances and New Domains - 13th International Conference, ESWC 2016, Heraklion, Crete, Greece, May 29 - June 2, 2016, Proceedings, volume 9678 of *Lecture Notes in Computer Science*, Springer, 2016, pp. 709–723.

[10] Á. Sicilia, G. Nemirovski, A. Nolle, Map-on: A web-based editor for visual ontology mapping, Semantic Web 8 (2017) 969–980.

[11] J. Bak, M. Blinkiewicz, A. Lawrynowicz, User-friendly visual creation of R2RML mappings in square, in: Proceedings of the Third International Workshop on Visualization and Interaction for Ontologies and Linked Data co-located with the 16th International Semantic Web Conference (ISWC 2017), Vienna, Austria, October 22, 2017, volume 1947 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2017, pp. 139–150.

[12] A. Crotti Junior, C. Debruyne, D. O'Sullivan, Juma: An editor that uses a block metaphor to facilitate the creation and editing of R2RML mappings, in: The Semantic Web: ESWC 2017 Satellite Events - ESWC 2017 Satellite Events, Portorož, Slovenia, May 28 - June 1, 2017, Revised Selected Papers, volume 10577 of *Lecture Notes in Computer Science*, Springer, 2017, pp. 87–92.

[13] P. Heyvaert, A. Dimou, B. De Meester, T. Seymoens, A. Herregodts, R. Verborgh, D. Schuurman, E. Mannens, Specification and implementation of mapping rule visualization and editing: Mapvowl and the rmleditor, J. Web Semant. 49 (2018) 31–50.

[14] A. Crotti Junior, C. Debruyne, L. Longo, D. O'Sullivan, On the mental workload assessment of

uplift mapping representations in linked data, in: Human Mental Workload: Models and Applications - Second International Symposium, H-WORKLOAD 2018, Amsterdam, The Netherlands, September 20-21, 2018, Revised Selected Papers, volume 1012 of *Communications in Computer and Information Science*, Springer, 2018, pp. 160–179.

[15] A. Crotti Junior, A Jigsaw Puzzle Metaphor for Representing Linked Data Mappings, Ph.D. thesis, Trinity College Dublin, 2019.

[16] H. García-González, I. Boneva, S. Staworko, J. E. L. Gayo, J. M. C. Lovelle, Shexml: improving the usability of heterogeneous data mapping languages for first-time users, PeerJ Comput. Sci. 6 (2020) e318.

[17] P. Warren, P. Mulholland, E. Daga, L. Asprino, Path-based and triplification approaches to mapping data into rdf: User behaviours and recommendations, Semantic Web (2023) 1–27.

[18] M. D. Brouwer, P. Bonte, D. Arndt, M. V. Sande, A. Dimou, R. Verborgh, F. D. Turck, F. Ongenae, Optimized continuous homecare provisioning through distributed data-driven semantic services and cross-organizational workflows, J. Biomed. Semant. 15 (2024) 9.

[19] P. Heyvaert, A. Dimou, R. Verborgh, E. Mannens, R. Van de Walle, Semantically annotating CEUR-WS workshop proceedings with RML, in: Semantic Web Evaluation Challenges - Second SemWebEval Challenge at ESWC 2015, Portorož, Slovenia, May 31 - June 4, 2015, Revised Selected Papers, volume 548 of *Communications in Computer and Information Science*, Springer, 2015, pp. 165–176.

[20] I. A. Ibrahim, T. Choudhury, J. Sargeant, R. Shah, M. J. Hossain, S. M. Islam, CEREI: an open-source tool for cost-effective renewable energy investments, SoftwareX 26 (2024) 101708.

[21] P. Heyvaert, B. De Meester, A. Dimou, R. Verborgh, Declarative rules for linked data generation at your fingertips!, in: The Semantic Web: ESWC 2018 Satellite Events - ESWC 2018 Satellite Events, Heraklion, Crete, Greece, June 3-7, 2018, Revised Selected Papers, volume 11155 of *Lecture Notes in Computer Science*, Springer, 2018, pp. 213–217.

[22] A. Iglesias-Molina, D. Chaves-Fraga, I. Dasoulas, A. Dimou, Human-friendly RDF graph construction: Which one do you chose?, in: Web Engineering - 23rd International Conference, ICWE 2023, Alicante, Spain, June 6-9, 2023, Proceedings, volume 13893 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 262–277.

[23] J. Brooke, Sus: a retrospective, J. Usability Studies 8 (2013) 29–40.

[24] J. R. Lewis, Psychometric evaluation of the pssuq using data from five years of usability studies, International Journal of Human-Computer Interaction 14 (2002) 463–488.

[25] L. Faulkner, Beyond the five-user assumption: Benefits of increased sample sizes in usability testing, Behavior Research Methods, Instruments, & Computers 35 (2003) 379–383.

[26] E. Daga, L. Asprino, P. Mulholland, A. Gangemi, Facade-x: an opinionated approach to SPARQL anything, CoRR abs/2106.02361 (2021). URL: https://arxiv.org/abs/2106.02361. arXiv:2106.02361.

[27] P. S. Tsang, V. L. Velazquez, Diagnosticity and multidimensional subjective workload ratings, Ergonomics 39 (1996) 358–381.

[28] S. G. Hart, Nasa-task load index (nasa-tlx); 20 years later, in: Proceedings of the human factors and ergonomics society annual meeting, volume 50, 2006, pp. 904–908.

[29] R. A. Peterson, A meta-analysis of cronbach's coefficient alpha, Journal of consumer research 21 (1994) 381–391.