Journal Title
XX(X):2–27

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

AI-Assisted Knowledge Graph Metadata Curation: An Empirical Evaluation

Maryam Mohammadi[1], Anas Elghafari[1], Chang Sun[1] and Michel Dumontier[1]

Abstract

Metadata is essential for making knowledge graphs (KGs) findable and reusable in line with the FAIR principles. However, KG metadata is often incomplete or missing, partly because creating high-quality metadata is a tedious and time-consuming manual task. This paper presents an AI-assisted form for KG metadata curation. The form is structured according to a KG metadata specification and uses a large language model to extract candidate values for metadata fields from KG documentation. The model standardizes the extracted values according to the KG metadata specification and presents them as suggestions that users can review, accept, or reject. We evaluated the AI-assisted form in a within-subjects study by comparing it with two baselines: a version of the same form without AI assistance and manual curation using RDF/Turtle. We assessed curator experience using the System Usability Scale, the Technology Acceptance Model, and open-ended feedback. To complement these subjective measures, we measured curation performance in terms of productivity and metadata quality in terms of coverage and precision. The results indicate that the AI-assisted form enables a more efficient curation workflow, leads to more complete metadata, and is preferred by participants over both baselines.

Keywords

Knowledge Graph, Metadata, AI-assisted form, within-subjects study, FAIR

1 Introduction

Knowledge Graphs (KGs) are graph-based structures for representing real-world knowledge, with entities of interest modeled as nodes and relations between them modeled as edges [29]. They are increasingly used in academia and industry to integrate heterogeneous data, support semantic search, and enable querying and analysis [7,29]. However, due to their complex and heterogeneous structures, KGs can be difficult to reuse without metadata that describe their scope, schema, provenance, access conditions, and licensing [16].

Metadata is essential for making KGs findable, accessible, interoperable, and reusable in accordance with the FAIR principles [10]. However, Linked Data and RDF datasets often lack complete and accurate metadata [8,9]. In the Linked Open Data cloud, for instance, 64% of general metadata and 80% of provenance metadata have been reported as missing or erroneous [9]. Such gaps hinder semantic interpretation, provenance tracking, querying, and downstream reuse [10].

To support metadata quality and interoperability, the Semantic Web and Linked Data communities have proposed standardized metadata vocabularies, profiles, and specifications. Vocabularies such as VoID [34], DCAT [31], DCTerms [33], PROV-O [32], and Schema.org [35] provide semantic terms for describing key dataset attributes, including provenance, access information, licenses, and conditions of use. Building on these vocabularies, the HCLS Community defines a profile for dataset descriptions in the health care and life sciences domain [19]. This line of work has recently been extended with a metadata specification tailored specifically to KGs [16]. However, producing metadata that conforms to such vocabularies and specifications is often still a manual process [16,18]. Curators must identify suitable metadata values, understand the expected structures and value formats, and encode the metadata correctly. As a result, metadata provision can require substantial time, effort, and expertise, making it burdensome for data providers [17,18].

Previous work has explored metadata authoring and capture tools that support metadata curation, such as CEDAR and REDCap [25,26]. These systems mainly target scientific experiment metadata and biomedical data capture rather than KG metadata curation. In the context of RDF datasets and KGs, previous studies have investigated RDF dataset profiling, KG summarization, and semi-automatic methods and tools for metadata generation [1–3,5,6,8]. However, practical KG metadata curation still often depends

[1]Institute of Data Science, Department of Advanced Computing Sciences, Maastricht University, The Netherlands

Corresponding author:
Maryam Mohammadi, Institute of Data Science, Department of Advanced Computing Sciences, Paul-Henri Spaaklaan 1, 6229 GS Maastricht, The Netherlands
Email: m.mohammadi@maastrichtuniversity.nl

on manual structured input, including form-based interfaces or direct authoring in RDF/Turtle syntax. These approaches provide limited support for exploiting information that is already available in the KG documentation. Although large language models (LLMs) have shown promise for structured information extraction and metadata curation in related settings[18,21,23,24], their use for specification-compliant KG metadata curation from KG documentation remains underexplored.

To address this gap, we propose an *AI-assisted form* for KG metadata curation. The form is structured according to the metadata elements defined in a recently proposed KG metadata specification[16]. For each metadata element, the tool generates suggestions using LLM-based retrieval-augmented generation over user-provided documentation, while constraining the output to the specification's required structures and value formats. Users can inspect, accept, or reject these suggestions before finalizing the metadata, thereby combining automation support with user control. This paper investigates the following research question (RQ): To what extent does AI-assisted form improve KG metadata curation, in terms of metadata quality and curator experience, compared with the form without AI assistance and manual specification-based curation in RDF/Turtle syntax?

To answer this question, we conducted a within-subjects study[20] comparing the proposed *AI-assisted form* with two baselines: a *standard form* and a *Turtle editor*. This design allows us to examine the two main components of the proposed approach: AI-generated suggestions and a specification-guided form interface. The standard form uses the same specification-guided interface as the AI-assisted form but without AI-generated suggestions, allowing us to isolate the effect of AI assistance. The Turtle editor represents the manual specification-based condition, in which participants curate metadata directly in RDF/Turtle while consulting the KG metadata specification. We assess both the quality of the curated metadata and participants' experience during the curation process. The contributions of this paper are as follows.

- We propose an AI-assisted approach for KG metadata curation that uses LLMs to generate specification-compliant metadata suggestions from KG documentation.
- We conduct a within-subjects study that provides empirical evidence on the effects of AI-assisted approach on KG metadata quality and curator experience.

The remainder of this paper is organized as follows. Section 2 reviews existing approaches to metadata capture, extraction, and curation. Section 3 presents the proposed approach, study design, and evaluation criteria. Section 4 reports the results. Section 5 discusses the findings, limitations, and directions for future research, and concludes the paper.

2 Related Work

2.1 Metadata authoring and capture tools

Several approaches have been proposed to support metadata creation, capture, and curation, including structured interfaces and template-based methods. The CEDAR

Workbench provides an ontology-assisted platform for authoring metadata through form-based interfaces and designing metadata templates. It also supports metadata validation and the export of metadata records in machine-readable formats such as JSON, JSON-LD, and RDF[25]. Martínez-Romero et al. proposed a recommendation framework that uses existing metadata records together with ontology-based metadata specifications to recommend values for metadata fields during authoring[30]. REDCap is a metadata-driven platform that supports clinical and translational research teams in managing project-specific research data[26]. It provides electronic case report forms with features for data collection, storage, validation, sharing, import, and export.

These approaches demonstrate how templates and interfaces can support structured metadata authoring through validation mechanisms and value-recommendation features. However, they mainly target scientific experiment metadata and biomedical data capture rather than KG metadata. Moreover, their assistance relies on ontologies, previously entered metadata values, or study-specific data dictionaries, rather than LLM-generated suggestions derived from KG documentation.

2.2 Metadata extraction from unstructured input

Another body of work focuses on extracting structured metadata from unstructured and semi-structured sources. CERMINE extracts bibliographic metadata, and document structure from scholarly PDFs using a modular workflow based largely on supervised and unsupervised machine-learning techniques[22]. Nayak et al. predict experimental metadata key–value pairs from scientific publications using neural models such as convolutional neural networks (CNNs) and bidirectional long short-term memory networks (BiLSTMs)[27].

More recently, LLMs have been explored for structured information extraction. Dagdelen et al. fine-tune GPT-3 and Llama-2 for named-entity and relation extraction from materials-science text, producing structured outputs such as JSON objects[23]. Busch et al. evaluate zero-shot, few-shot, and fine-tuning strategies for generating DCAT-compatible metadata from text-based data[18]. Sundaram et al. show that GPT-4 improves metadata standardization when prompted with domain information from textual descriptions of CEDAR templates[24]. Building on this direction, our work uses LLMs to extract structured metadata from KG documentation and evaluates how the resulting suggestions support KG metadata curation in practice.

2.3 KG profiling, summarization, and metadata generation

Prior work on RDF datasets and KGs has investigated methods for generating metadata, dataset profiles, and graph summaries. RDF dataset profiling aims to describe datasets by extracting information such as vocabularies, quality indicators, and statistical properties. Ben Ellefi et al.[8] surveyed features, methods, tools, vocabularies, and applications for RDF dataset profiling. Roomba automatically extracts, validates, corrects, and generates dataset profiles for linked datasets[1]. ABSTAT-HD supports scalable profiling of KGs by generating informative and concise dataset summaries and statistics[2].

Graph summarization approaches generate compact representations of KGs that help users understand large and complex graphs[4]. These summaries constitute a form of descriptive metadata, as they capture the content and structure of a KG. ABSTAT uses ontology-driven abstraction and pattern minimalization to summarize RDF datasets through schema-level patterns[3]. PCSG generates compact pattern-coverage snippets that exemplify representative patterns of entity descriptions and links in RDF datasets[5]. GLIMPSE addresses personalized KG summarization by creating summaries that contain the facts most relevant to a user's inferred interests, under a user- and device-specific size constraint[6].

While existing works focus on KG profiling, summarization, metadata generation from graph data, and user interests, our work instead exploits KG documentation to facilitate specification-compliant KG metadata curation.

2.4 User interaction with metadata curation systems

Previous studies have explored user interactions with digital tools during structured data entry and metadata curation. Evaluations of such systems often consider both task outcomes and user-perceived measures, such as usability and perceived usefulness[12,13]. Recent work has applied similar evaluation approaches in the context of AI-assisted curation workflows. For example, SciDaSynth is an interactive LLM-based system that supports structured data extraction from scientific literature. It uses a retrieval-augmented generation approach to generate structured data tables from users' questions by integrating information from text, tables, and figures[21]. SciDaSynth was evaluated through a within-subjects study that examined both the efficiency of the system and users' perceptions of the generated outputs.

Another relevant system is SMECS, which addresses research software metadata curation. It harvests existing metadata from software repositories, such as GitHub, and presents the collected information in an interactive curation interface. SMECS was evaluated using the System Usability Scale and semi-structured interviews[28]. These studies evaluated not only the quality of metadata, but also users' perceptions of metadata curation systems and their interactions with them. Our work follows this user-centered perspective in the context of KG metadata curation. We conduct a within-subjects study to evaluate both the user experience during the curation process and the quality of the resulting metadata.

In summary, prior work has shown the importance of metadata templates and form-based tools, explored automatic and semi-automatic methods for metadata extraction and prediction, and evaluated their usability and usefulness. However, existing work has not specifically examined the use of AI for specification-compliant KG metadata curation, nor has it empirically evaluated the impact of such approaches on metadata quality and curator experience. This study addresses this gap by implementing an AI-assisted form for KG metadata curation and evaluating its impact through a within-subjects study.

3 Methodology

In this paper, we investigate the extent to which AI-assisted form supports KG metadata curation in terms of curator experience and metadata quality. To this end, we designed a within-subjects study in which each participant curated metadata using three study conditions: the AI-assisted form, the standard form, and the Turtle editor.

Figure 1 illustrates the workflow of our methodology, which comprises four main stages: study design and preparation, implementation of the study conditions, user study session, and evaluation. We first established the study protocol, prepared the study materials, and defined the recruitment process. We then implemented the three tools corresponding to the study conditions and prepared the data collection instruments, including questionnaires for gathering participants' ratings and feedback. During the user study sessions, participants completed the metadata curation tasks and questionnaires. Finally, in the evaluation stage, we analyzed the collected metadata, task durations, and questionnaire responses to assess metadata quality, curation performance, and curator experience. The following subsections first introduce the KG metadata specification used throughout the study, and then describe each stage of the methodology in detail.

3.1 KG Metadata Specification

The KG metadata specification[16] defines a community-developed framework for structured, FAIR-aligned descriptions of KGs. It comprises 33 metadata elements for describing KG characteristics such as access conditions, provenance, licensing, distributions, endpoints, and links to related resources. In this specification, each element is associated with a set of constraints, such as expected cardinality and value type. For example, the title element has a cardinality of exactly one, meaning that each KG should have only one main title, and its value should be provided as a string or a language-tagged string. Another example is the element specifying the language(s) of the KG, which may have multiple values in the form of language tags.

The specification includes elements with varying expected data types, such as IRIs, timestamps, string literals, and language-tagged strings. In addition, some elements have a more complex structure. Specifically, five elements are composite: distributions, SPARQL endpoints, roles, example resources, and linked resources. Each composite element is described by a set of sub-elements that specify its associated properties. For example, a KG distribution is described through sub-elements specifying its title, description, access URL, media type, and download URL.

In this study, the KG metadata specification was used as a FAIR-aligned basis for KG metadata curation across all conditions. It defined the metadata elements represented in the form-based interfaces and informed the prompts used to generate LLM suggestions. In the Turtle editor condition, the specification was provided to participants in spreadsheet format to guide manual metadata curation.

3.2 Study Design and Preparation

We first defined the study protocol and designed the metadata curation tasks based on the RQ. We then submitted the protocol and data management plan to the ethics committee

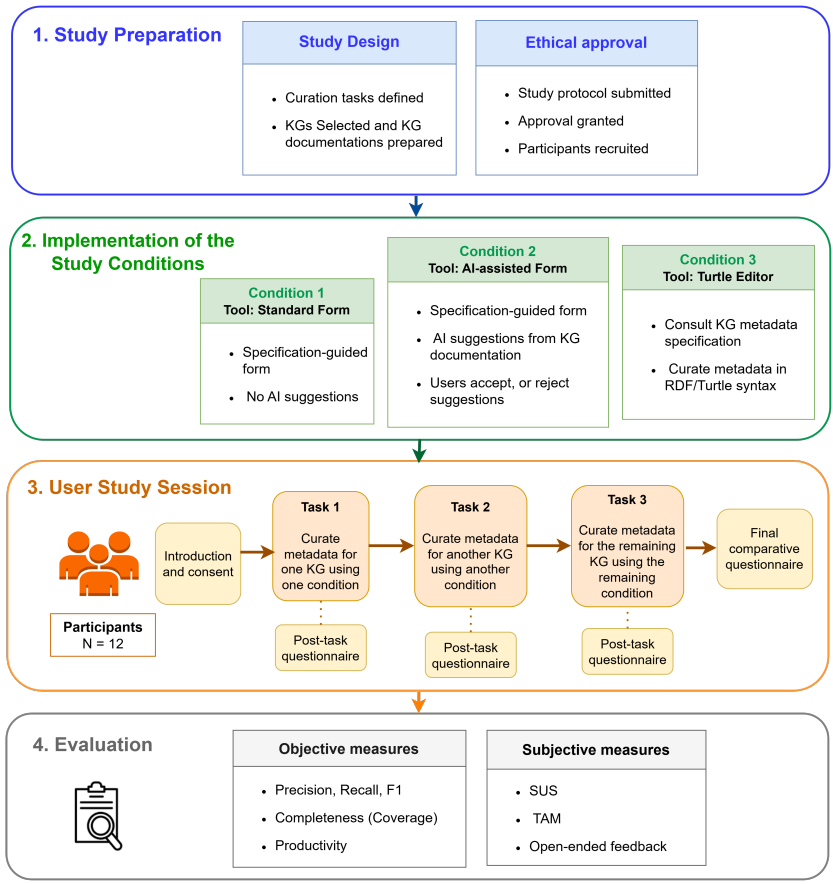

Figure 1. Overview of the methodology.

and obtained approval before recruiting participants for the user study. Finally, we selected the KGs used in the experiment and prepared the corresponding documentation and reference metadata for the curation tasks.

Our study design required each participant to complete three metadata curation tasks, each using a different tool. Using the same KG across all tools could have introduced learning effects, as participants might have become familiar with the KG and reused metadata values encountered in earlier tasks. To address this concern, we used a different KG for each task. In addition, we selected KGs from distinct domains to reduce potential bias from participants' prior familiarity with any particular domain. Specifically, we used YAGO, a general-purpose KG; QuoteKG, a KG focused on quotations; and Bio2RDF, a bioinformatics KG. For each KG, we prepared two resources: a textual KG documentation file and a reference metadata set.

The KG documentation was provided to participants together with the assigned curation tool and served as the input for extracting metadata about the KG. The documentation was designed to reflect a metadata curation scenario in which metadata creators have access to existing descriptive files about a KG, but the relevant information is not necessarily presented in a structured or specification-compliant format. To create this documentation, we collected the information required for the metadata elements defined in the KG metadata specification from online resources. We then used ChatGPT to generate Wikipedia-style descriptive documentation containing this information. For selected metadata elements, namely KG language, created date, modified date, and issued date, we intentionally expressed the values in forms that differed from the exact formats required by the KG metadata specification. For example, language values were written as full language names, such as English, rather than language tags, such as `en`; dates were expressed in natural language rather than in the required date format.

In practice, KG documentation may be distributed across multiple artifacts, such as project pages, schema or ontology files, README files, and descriptive notes. In this study, however, we represented the available documentation for each KG as a single narrative document. This design choice made the tasks manageable within the time-bounded study setting by reducing the need to search across heterogeneous documentation sources.

The reference metadata sets were used to assess both the participant-submitted metadata and the AI suggestions. They were represented in a key–value format aligned with the KG metadata specification and covered all metadata elements defined in the specification, including the sub-elements of composite elements. Counting elements with a single value, elements with multiple values, and sub-elements of composite elements as separate entries, the reference metadata sets contained 66, 72, and 76 entries for YAGO, QuoteKG, and Bio2RDF, respectively.

Before the study session, the study protocol and data management plan were reviewed by Maastricht University's Ethics Review Committee Inner City Faculties (ERCIC). The committee concluded that there were no ethical objections to the execution of the research project (reference: ERCIC_753_08_09_2025_Mohammadi). The protocol covered the study design, tasks, procedure, and data protection procedures. All data collected during the study were pseudonymized and linked to participant IDs rather than directly identifying information. The mapping between participant IDs and participant identities was stored separately by the principal investigator, enabling participants to request withdrawal of their data in accordance with the study's data protection procedures.

Participants were recruited through email announcements and posters distributed at Maastricht University. Recruitment targeted a broad audience with varying levels of familiarity with KGs and metadata curation. Twelve participants volunteered and completed the study. Each participant received an information letter describing the study goals, procedure, and data protection measures.

3.3 Implementation of Study Conditions

3.3.1 Standard Form. The standard form is an interface structured according to the KG metadata specification. The form displays a definition and the expected datatype for each element of the specification, and shows validation messages when an input value does not match the required format. For the license field, the form provides a drop-down list of possible values to support metadata standardization. In this condition, participants used the KG documentation to find metadata values, convert them to the expected format, and enter them into the form. The tool supports the export of curated metadata in JSON format.

3.3.2 AI-assisted Form. The AI-assisted form extends the standard form by using the GPT-4o-mini large language model to generate suggestions for all metadata elements. The user uploads a free-text narrative description of the KG (as a .docx file), and the tool issues a single prompt that instructs the model to both extract candidate values and return them in the format expected by each metadata element (e.g., IRIs, ISO 8601 dates, BCP-47 language tags, and literals). The prompt encodes field-specific guidance, semantic matching, and inference heuristics that direct the model to reason beyond surface-level keyword matching. In addition, it specifies validation constraints for each value type. OpenAI's structured output feature is used to enforce a predefined JSON schema on the model's response.

The tool processes the KG documentation and generates suggestions for all metadata elements in a single step, typically within 60–75 seconds. As shown in Figure 2, suggestions appear in a dedicated panel alongside the form, where users can accept or reject each suggestion before finalizing the metadata, maintaining a human-in-the-loop workflow.

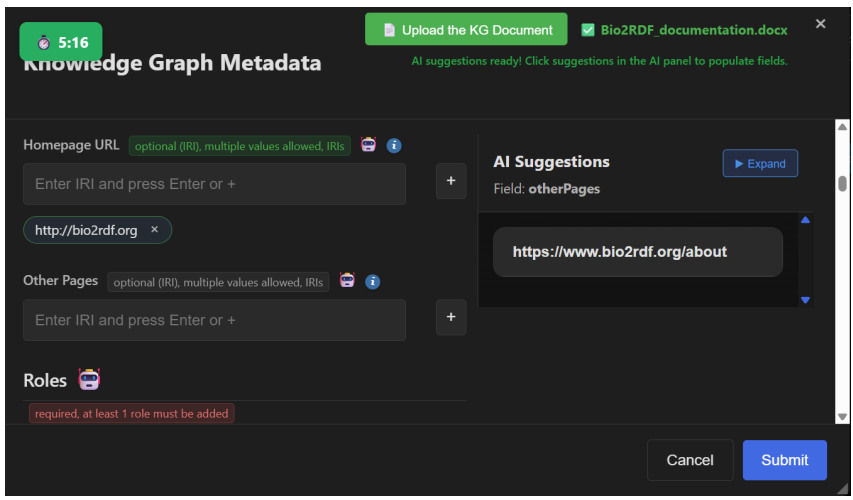

Figure 2. Screenshot of the AI-assisted form.

3.3.3 Turtle Editor. The Turtle editor served as a manual, specification-based baseline representing a syntax-based workflow. It provided a text-based environment for creating KG metadata in Turtle syntax, with syntax validation and highlighting of syntactic errors. In this condition, participants used the KG documentation to identify metadata values and consulted the KG metadata specification provided to them as a spreadsheet to produce metadata in Turtle syntax.

3.4 User Study Session

The study was conducted on site. Following a brief introduction, participants were given access to a Microsoft Form questionnaire sent to them by email. This form guided them through each step of the study.

At the beginning of the session, participants submitted an informed consent form and completed a background questionnaire. They then performed three metadata curation tasks, each followed by a post-task questionnaire. For each task, the questionnaire provided links to the relevant KG documentation and the assigned metadata curation tool. Participants were instructed to use the provided materials to curate metadata, export the resulting metadata, and submit their output via a shared folder. After completing all three tasks, participants answered a final comparative questionnaire assessing their overall preferences and impressions of the tools. To mitigate potential order effects and to avoid consistently pairing the same tools with the same KGs, we designed six study forms (A–F). These forms varied in the order in which participants used the tools and in the assignment of KGs to tools.

3.5 Evaluation

The evaluation consisted of a pre-study background questionnaire, subjective measures, and objective measures. The pre-study questionnaire captured participants' prior familiarity with the concepts relevant to the study. Subjective measures captured participants' perceptions of each tool through post-task and final questionnaires. The objective measures assessed AI suggestions, submitted metadata, and task completion time. The following subsections describe these measures.

3.5.1 Pre-study background questionnaire. The pre-study background questionnaire shown in Table 1 assessed participants' familiarity with concepts such as KGs, metadata standards, data modeling, Turtle syntax, and related Semantic Web technologies. Questionnaire responses were converted to numerical values and summarized into a single expertise score. Likert-scale familiarity items were mapped to a five-point scale, and open-ended items were mapped to a four-point scale. For each participant, we computed the mean across the six questionnaire items to obtain one score representing their overall level of relevant expertise.

3.5.2 Post-task Questionnaires. After each task, participants completed a post-task questionnaire about the tool they had just used. The questionnaire consisted of three parts: the System Usability Scale (SUS)[13], constructs from the Technology Acceptance Model (TAM)[12], and open-ended questions for additional feedback.

Table 1. Pre-study background questionnaire.

Questionnaire items
1. Semantic Web (e.g., RDF, SPARQL, FAIR) 2. Metadata standards (e.g., Dublin Core, schema.org) 3. Turtle syntax 4. Data modeling (e.g., designing database schemas, XML, or ontologies) 5. Working with structured data (e.g., tabular data, or converting unstructured data to structured data) *Options:* 5-point Likert scale (1 = Not at all familiar, 5 = Extremely familiar).
6. How many years of experience do you have in Semantic Web technologies? *Options:* None; Less than one year; 1–3 years; More than 3 years

Table 2. The Standard System Usability Scale. Item 8 is shown with "awkward" in place of the original "cumbersome"[11]

Q	Item
1	I think that I would like to use this system frequently.
2	I found the system unnecessarily complex.
3	I thought the system was easy to use.
4	I think that I would need the support of a technical person to be able to use this system.
5	I found the various functions in this system were well integrated.
6	I thought there was too much inconsistency in this system.
7	I would imagine that most people would learn to use this system very quickly.
8	I found the system very awkward to use.
9	I felt very confident using the system.
10	I needed to learn a lot of things before I could get going with this system.

The SUS was used to evaluate the perceived usability of each tool, which consists of ten items shown in Table 2. It produces a standardized usability score on a scale from 0 to 100, facilitating direct comparison across the three tools[13]. While SUS provides a general measure of usability, TAM was used to capture participants' task-specific perceptions in the context of metadata curation.

Table 3 presents the items associated with the TAM constructs examined in this study: perceived usefulness (PU), perceived ease of use (PEOU), and behavioral intention (BI)[12]. In the context of metadata curation, PU refers to the extent to which participants perceive the tool as helpful in improving their metadata curation task performance. PEOU denotes the extent to which participants perceive the tool as understandable and believe that it requires minimal effort to learn and use. BI captures participants' intention to use the tool for future KG metadata curation tasks or recommend it to others, reflecting the likelihood of tool adoption. To measure these constructs, responses were recorded on a five-point Likert scale. For each construct, one item was reverse-worded to reduce

acquiescence bias; these items were reverse-coded before calculating the mean construct scores.

Table 3. TAM constructs. Participants rated each statement on a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree).

Construct	Statement
PU	This tool enables me to accomplish metadata curation tasks.
	This tool supports critical aspects of my metadata curation work.
	Using this tool improves the quality of the metadata I curate.
	Using this tool increases my productivity on curating metadata.
	This tool does not support critical aspects of my metadata curation work. (reverse coded)
PEOUS	This tool provides helpful guidance in performing tasks.
	Interacting with this tool does not require a lot of my mental effort.
	I find it easy to get this tool to do what I want it to do.
	Interacting with this tool is frustrating. (reverse coded)
BI	I plan to use this tool again for future Knowledge Graph metadata curation.
	I would recommend this tool to others who work with metadata.
	I do not intend to use this tool again for future Knowledge Graph metadata curation. (reverse coded)

All Likert-scale responses were converted into numerical values for analysis. SUS scores were calculated according to the standard procedure, in which odd-numbered items were scored as response minus 1 and even-numbered items as 5 minus response. The sum of the ten item scores was then multiplied by 2.5 to obtain the final usability score. For TAM, responses were aggregated into construct scores, and mean scores for each construct were then calculated across participants for each tool.

Table 4. Post-task open-ended questionnaire items.

Open-ended Question
What did you like most about this tool?
What difficulties did you experience while using this tool?
Please share any feedback about your experience using this tool (positive or negative).
How could this tool be improved?

Finally, the post-task questionnaire included open-ended questions about participants' experiences with the tool. As shown in Table 4, these questions focused on what participants liked most, the difficulties they encountered, and their suggestions for improvement. To analyze the responses to these questions, we employed an inductive thematic analysis following Braun and Clarke [15]. This approach is a bottom-up qualitative method in which codes and themes are derived from the data itself rather than being based

on predefined frameworks. In line with this approach, the responses to the open-ended questions were first read in full to ensure familiarity with the dataset. They were then coded iteratively to identify recurring patterns across participants' answers, after which related codes were grouped into broader themes. In the Results section, we report the number of coded statements associated with each theme and illustrate the themes with representative participant quotations.

3.5.3 Final comparative questionnaire. After completing all tasks, participants filled in a final comparative questionnaire capturing their overall impressions of the tools. As shown in Table 5, the questionnaire included three questions: (i) which tool(s) they would prefer to use for future metadata curation (participants could select up to two tools), (ii) which tool enabled the most efficient task completion (single-choice), and (iii) an optional open-ended prompt for additional comments, or suggestions.

Table 5. Final comparative questionnaire.

Question	Response Type
Which tool or tools would you prefer to use for metadata curation in the future?	select two of the three tools
Which tool do you think helped you complete the task most efficiently?	select one of the three tools
Please share any additional comments, suggestions, or feedback about the tools.	open-ended answer

3.5.4 Evaluation of Submitted Metadata. This section presents the objective metrics used to evaluate the metadata curated by participants in each task. Metadata quality was assessed in terms of completeness and correctness, while curation performance was assessed using task completion time. The following sections describe the preprocessing of curated metadata and define each metric.

Preprocessing of submitted metadata. As described in Section 3.1, the specification includes single-valued, multi-valued, and composite metadata elements. Before analysis, both the submitted metadata and the reference metadata were transformed into a flat representation. In this representation, each individual element–value entry was treated as a separate unit for evaluation. For example, a multi-valued element such as [language: en, fr, ar] was split into separate entries: [language: en], [language: fr], and [language: ar]. Similarly, a composite element such as distribution: [title: QuoteKG dataset files, accessURL: https://zenodo.org/record/4702544] was decomposed into sub-elements, such as [distribution.title: QuoteKG dataset files] and [distribution.accessURL: https://zenodo.org/record/4702544]. The resulting flat element–value entries were used to calculate the evaluation metrics.

Correctness. To assess correctness, we measured the similarity between each submitted value and its corresponding reference value. We used a type-aware similarity framework because the specification includes different value types, such as IRIs, timestamps, descriptions, categorical terms, names, and string literals. This framework applies different similarity functions depending on the expected datatype of each

metadata element. Table 6 presents the KG metadata specification elements categorized according to the similarity measure used in this study.

Metadata elements requiring standardized or format-constrained values, such as IRIs, language tags, dates, and URLs, were evaluated using exact match. For example, a language value was considered correct only when entered using the required language tag, such as `en`, rather than a label such as `English`. Exact match is defined as:

$$\mathrm{Exact}(r, s) = \begin{cases} 1, & \text{if } r = s, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

where r is the reference value and s is the submitted value.

Description fields were evaluated using ROUGE-1 F-measure, computed with the `rouge-score` library. ROUGE-1 measures word-level overlap between the reference value and the submitted description. Unlike exact match, this metric assigns partial credit when the submitted description contains overlapping words with the reference, even if the two descriptions are not identical.

Term-based fields, such as keywords and themes, were evaluated using cosine similarity over TF-IDF vectors. In this setting, cosine similarity was used as a lexical overlap measure and computed as:

$$\mathrm{Cosine}(r, s) = \frac{\mathbf{v}_r \cdot \mathbf{v}_s}{\|\mathbf{v}_r\| \, \|\mathbf{v}_s\|}, \tag{2}$$

where $\mathbf{v}_r$ and $\mathbf{v}_s$ are the TF-IDF vectors of the reference and submitted value, respectively.

Metadata elements such as creators, and publishers, for which the expected values are names, were evaluated using fuzzy string similarity as implemented in the `fuzzywuzzy` library. This metric accounts for minor spelling, punctuation, and formatting differences between strings. For example, `Max-Planck-Institut` and `Max Planck Institut` receive a high similarity score despite differences in punctuation and spacing.

For metadata elements with multiple values and for composite elements with sub-elements, correctness was assessed using an order-invariant greedy one-to-one matching procedure based on similarity scores. A matched pair was counted as correct if it satisfied the corresponding metric criterion: exact-match fields required identical normalized values, while fuzzy, ROUGE, and cosine-based fields required a similarity score of at least 0.7. The resulting counts of submitted, reference, and correctly matched values were used to compute precision, recall, and F1 score.

Completeness (Coverage). Completeness measures the proportion of reference element–value entries for which participants submitted a value, irrespective of whether the submitted value was correct:

$$\mathrm{Completeness} = \frac{N_{\mathrm{s}}}{N_{\mathrm{r}}}, \tag{3}$$

where N_{s} is the number of reference entries for which a participant submitted a value, and N_{r} is the total number of reference entries for the corresponding KG. We first computed

Table 6. Mapping of metadata elements and sub-elements to similarity metrics. Arrows indicate the sub-elements of composite elements.

Metric	Elements and sub-elements
Exact match	`KG title`, `alternative title`, `created date`, `modified date`, `issued date`, `version`, `acronym`, `language`, `type`, `access rights`, `number of triples`, `license`, `conforms to`, `name space`, `IRI template`, `references`, `home page`, `other pages`, `REST API`, `primary reference`, `vocabulary`, `meta graph`, `primary source`, `sample query`. *Sub-elements:* `publisher/creator` → (`role`, `email`). `distribution` → (`title`, `media type`, `download URL`, `access URL`). `SPARQL endpoint` → (`endpoint URL`, `identifier`, `title`, `status`). `example resource` → (`title`, `access URL`). `linked resources` → (`target`, `number of triples`).
ROUGE-1	`KG description` *Sub-elements:* `distribution` → (`description`). `SPARQL endpoint` → (`description`). `example resource` → (`description`).
Cosine	`KG keyword`, `KG theme`.
Fuzzy	*Sub-elements:* `publisher/creator` → (`name`).

completeness for each participant–tool combination and then aggregated the scores at the tool level for the comparative analysis.

Productivity. Productivity captures the rate at which participants submitted metadata values, irrespective of correctness. We measured it as the number of submitted element–value entries per minute:

$$\text{Productivity} = \frac{N_{\text{s}}}{T}, \tag{4}$$

where N_{s} is the number of submitted element–value entries and T is the task duration in minutes.

3.5.5 Accuracy of AI Suggestions. We assessed the correctness and completeness of the initial AI-generated suggestions before participant modification by comparing them with the reference metadata. These suggestions were stored in the task outputs of the AI-assisted tool. Values that were identical to the reference were counted as exact matches. Suggested values that were present but did not match the reference were counted as unmatched. Reference values that were absent from the AI suggestions were counted as missing.

3.5.6 Statistical analysis. To assess whether differences across tools were statistically significant, we conducted repeated-measures ANOVAs followed by Holm-corrected post-hoc pairwise comparisons for all evaluated measures.

4 Results

4.1 Participant background

Based on the scoring procedure described in Section 3.5.1, participants' responses to the six background questions were converted into numerical values. These values were then aggregated into a single composite expertise score for each participant by calculating their mean. Figure 3 presents the distribution of composite expertise scores, with each bar indicating the number of participants in each score range. Expertise scores ranged from 1.33 to 4.83 ($M = 2.71$, $SD = 1.25$), indicating that the sample included participants with different levels of familiarity with Semantic Web technologies, metadata standards, Turtle syntax, and data modeling. The distribution was slightly positively skewed (skew $= 0.56$), with most participants falling in the low-to-moderate expertise range and a smaller number demonstrating higher expertise.

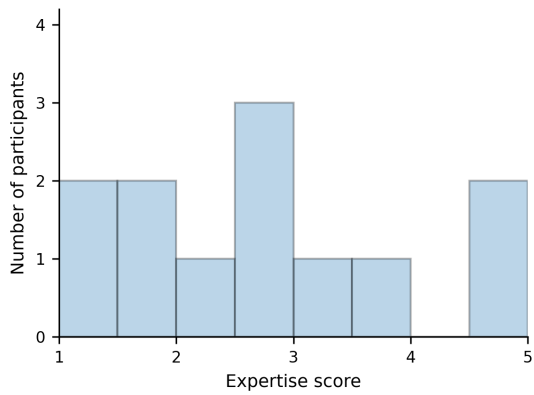

Figure 3. Distribution of participants' expertise scores.

4.2 Post-task Questionnaires

4.2.1 SUS and TAM Results. The SUS and TAM scores were calculated for each participant following the procedure described in Section 3.5.2. These scores were then averaged across participants to obtain tool-level mean scores. As shown in Figure 4, the AI-assisted form achieved the highest SUS score ($M = 80.6$), which is commonly interpreted in the literature as excellent usability [14]. The standard form obtained a mean SUS score of 64.0, corresponding to marginal usability. In contrast, the Turtle editor received a substantially lower score of 32.7, indicating poor usability. Statistical analysis

showed that these differences were significant. The AI-assisted form outperformed both the standard form ($p = 0.004$) and the Turtle editor ($p < 0.001$). The standard form also scored significantly higher than the Turtle editor ($p < 0.001$).

Figure 5 shows a consistent pattern for the TAM constructs. The AI-assisted form received the highest mean scores for PU, PEOU, and BI, followed by the standard form, while the Turtle editor obtained the lowest scores. Pairwise comparisons indicated that the AI-assisted form was scored significantly higher than the Turtle editor on all three constructs ($p < 0.001$). Compared to the standard form, it also received significantly higher scores for PEOU ($p = 0.002$) and PI ($p = 0.009$). Although the AI-assisted form showed a higher mean PU score than the standard form, this difference was not statistically significant ($p = 0.133$). The standard form, in turn, scored significantly higher than the Turtle editor among all three constructs ($p < 0.001$).

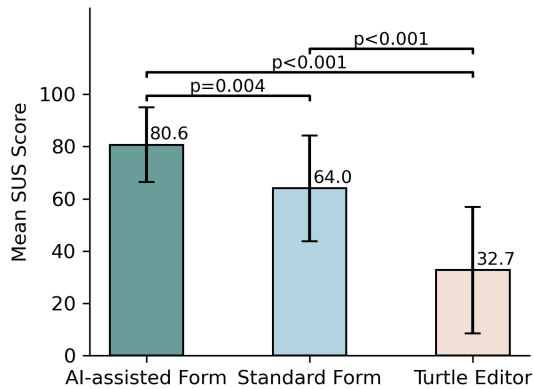

Figure 4. SUS scores for the three evaluated tools.

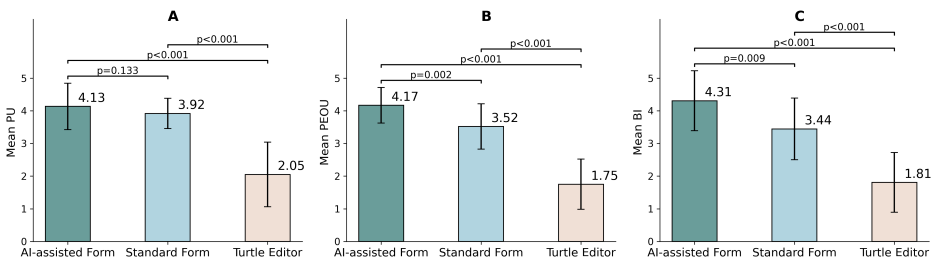

Figure 5. Mean TAM scores across tools: (A) perceived usefulness, (B) perceived ease of use, and (C) intention to use.

4.2.2 Qualitative Feedback Analysis. Using the inductive thematic analysis procedure described in Section 3.5.2, we analyzed participants' responses to the open-ended

questions and identified recurring themes. The findings for each tool are presented below and illustrated with representative quotations from participants, labeled by participant IDs (P1–P12). Ellipses (. . .) within quotations denote omitted text that was not directly relevant to the reported theme. For example, P4: "... the interface was easy to follow ..." indicates an excerpt from participant P4's complete response. The full set of quotations per tool is provided in the *Supplementary Excel file Qualitative Feedback Analysis*, where the themes are distinguished using color coding of codes derived from quotations.

Turtle Editor. A total of 45 codes were grouped into four broader themes: (1) Technical Features (15 codes), (2) Workflow and Usability (15 codes), (3) Background Knowledge (Turtle/RDF) (8 codes), and (4) Overall Impression and Emotional Reactions (7 codes). These themes are described below, together with representative participant quotations.

Technical Features. A total of 15 codes were about technical features. The participants appreciated the syntax validation feature (P12: "It has syntax check ...", P3: "It also provides feedback on the grammatical correctness ..."). They also noted the limitations of existing features, especially in error handling and guidance (P3: "There are no predefined prefixes or autocomplete ...", P2: "It provides basic error detection of syntax ..."). The desired features that were noted are mainly about integration with AI assistance and adding autocompletion (P1: "... add suggestions or hints or AI tab completion", P12: "... real-time syntax correction and suggestion like copilot").

Workflow and Usability. Fifteen codes were related to workflow and usability. Participants highlighted that using the tool requires strong background knowledge of semantic web concepts (P12: "I need to learn a lot", P7: "I need to understand the RDF data model, Turtle syntax, URI specifications, and vocabulary (such as dct:, dcat:, void:)"). Workflow-related difficulties were also reported (P6: "Following schema also takes time to check", P8: "it was difficult to switch between documentation (schema) and the editor.", P5:"I could not understand how to use it.", P4: "very hard to use ..."). A few participants provided positive comments regarding the flexibility of this approach (P3: "I had the flexibility to create KG metadata", P6: "I can control it").

Background Knowledge (Turtle/RDF). Seven coded statements indicated participants' limited prior knowledge of Turtle/RDF (P4: "... not familiar with the format of Turtle and my knowledge of namespace is insufficient.", P6: "If I do not know turtle background, it is difficult").

Overall Impression and Emotional Reactions. Seven codes reflected the overall impressions of the participants and the emotional responses to their experience using the tool, all negative and pointing to frustration and dissatisfaction (P3: "terrible and frustrated", P10: "I did not complete a valid input, very disappointing", P7: "It was not great."). In addition, some codes showed their unwillingness to use the tool again in the future (P4: "I wouldn't like to use it ever again.").

Standard Form. Fifty codes were identified and grouped into three themes: (1) Technical Features (29 codes), (2) Workflow and Usability (15 codes), and (3) Overall Impression (6 codes).

Technical Features. Almost half of the codes (29/50) were related to technical features. A few participants mentioned the validation feature and the dark theme as positive aspects of the tool (P3: ”it could capture and validate multiple answers, where warranted”, P5: ”... the fact that it had night theme.”). Some participants reported input-handling difficulties (P2: ”... does not indicate that we have to press enter after adding date.”, P8: ”I wanted to add a comma-separated list of languages but instead I had to add them one by one.”).

The suggested improvements included integration with AI assistance, autocompletion, and more examples and guidance for the user interface (UI) (P5: ”By adding more examples, better descriptions and maybe a list of possible choices, like it was in a field with choosing license.”, P4: ”integrate with LLM”, P3: ”suggest well known values (e.g. vocabularies used, languages in dataset, etc)”). Some participants preferred completing the form step by step rather than viewing it in full (P3: ”focus on required fields first, hide completed fields, ...”, P9: ”... give the user only a few entries to start with, then a few more along the way ...”). A save option was also suggested; however, this feature already exists but was disabled during the study to record metadata curation time (P6: ”The tool may need a save function to continue working”, P12: ”In my experience on the most websites, the text after inputing would be kept in that text box and the add icon could add a new input box. So it would be better to keep this manner.”).

Workflow and Usability. Fifteen codes were related to workflow and usability. Several codes indicate that participants found the tool easy to follow and well structured (P8: ”I had a list of fields I need to fill in and I didn’t have to switch between the editor and the schema”, P10: ”does not need recalling mundane details”, P12: ”Good structure to input metadata”). Some also noted that the tool does not require domain background knowledge (P7: ”Intuitive and easy to use, no professional knowledge required”, P5: ”do not need too much technology domain knowledge”), with one exception (P3: ”quite hard to understand for someone without experience.”). Some participants described difficulties in finding required information in the document and entering values manually (P12: ”Sometimes I had trouble finding the right information, for example, sometimes stuff is required but I cant find the right information for it. And It is also hard to copy paste stuff, e.g. urls are long and to type it by hand is not nice”).

Overall Impression. Five out of six codes reflected positive impressions of the tool, and one neutral opinion (P1: ”Definitely helpful to create KG metadata, if you know all the values beforehand.”, P5: ”It is easy to use for general researchers”, P6: ”The tool was quite nice to use”, P3: ”Rather neutral”).

AI-assisted Form. Fifty codes were identified and grouped into four themes: (1) Technical Features (15 codes), (2) Workflow and Ease of Use (19 codes), (3) AI suggestion Reliability (8 codes), and (4) Overall Impression and Emotional Reactions (8 codes).

Technical Features. Fifteen codes addressed technical features. Among these, some codes highlighted the tool’s ability to handle elements with sub-elements (P3: ”... very nice that it could complete multi-field blocks (e.g. distributions, contributors)”. Some participants reported limitations, such as not being able to edit suggestions (P8: ”I wanted

to select one option but then modify it (removing something I thought was unnecessary) and I couldn't do that."), and noted that the UI was sometimes unclear (P12: "The input box is a little bit confusing. At first, I did not know that the input text would appear as labels below the input box, It was challenging to figure out how to make sure that I input successfully"). Several suggestions were made for improving the UI, including adding autocomplete feature, showing a preview of curated metadata, enabling editing of suggestions, and reducing document load time (P2: "The tool can include a final overview that presents the completed metadata preview and illustrates how it appears in the knowledge graph format.", P8: "Adding an option to edit the suggested or accepted response would be useful.", P11: "... faster load times (maybe on the fly)").

Workflow and ease of use. Nineteen codes focused on workflow. Many highlighted that the process is helpful, easy to use, time-saving, and reduces concern about syntax errors while providing guidance and removing mundane work (P2:"It provides assistance to easily curate metadata without worrying about syntax of the graph.", P6: "it extracts accurately the information, helping me save time to complete", P8: "It pointed me (as a novice) to the right areas in the document.", P7: "It is very easy to use", P11: "I like that it suggests me about the stuff to fill in, giving me guidance on what specific thing I need to fill in and it being just 1 press of a button is nice").

There was some concern that suggestions might lead users to accept answers without sufficient checking (P10: "there is a high propensity for me to just pick the suggested answers. this can become an issue. because the majority of the answers are probably correct, the attention of the user would drop with time and some error will go through."). One participant suggested adding a disclaimer (P8: "you need a disclaimer or notice about the model you are using for the sake of transparency").

AI Suggestion Reliability. Eight codes were about AI suggestions. Some participants expressed confidence (P9: "The AI suggestions are great", P12: "After checking the results the AI suggested, they were quite precise"), while others raised concerns about correctness (P5: "I am not sure if suggested input is correct.", P6: "I still need to check if the filled answer is reliable or correct"). Some participants suggested improving the reliability of AI suggestions by providing the surrounding context and citing sources (P10: "I feel it could be better for the LLM to also show the text around the answer which can then be read by the user in a dynamic way.", P6: "maybe add the source where LLM came up with the answer").

Overall Impression. A total of 8 codes reflected participants' overall perceptions of the tool, all of which were highly positive (P1: "very cool idea", P7: "It was very useful to a novice.", P4: "comfortable and confident").

4.2.3 Final comparative questionnaire. As described in Section 3.5.3 and presented in Table 5, the final comparative questionnaire collected participants' overall impressions of the tools. For the first question, participants were asked which tool or tools they would prefer to use for metadata curation in the future, with the possibility to select up to two tools. Fifty percent of the participants (6/12) selected only the AI-assisted form, 8.3% (1/12) selected only the standard form, and 41.7% (5/12) selected both the AI-assisted form and the standard form. No participant selected the Turtle editor. For the

second question, participants were asked which tool helped them complete the task most efficiently, with only one tool allowed to be selected. All participants (100%; 12/12) selected the AI-assisted form.

Providing additional comments was optional, and six out of twelve participants shared their feedback. While these comments were similar to the themes identified in the qualitative analysis above, one notable suggestion was to integrate the three tools into a unified workflow that could better support both novice and expert users (P7: "The three tools represent three distinct user personas, a unified platform that combines all three modes would be far more powerful. For example, integrate the tools into a single hybrid interface, this would support both novices and experts in the same workflow"). This functionality had already been implemented. However, participants were not given access to the integrated interface during the study, as each task was designed to evaluate only one tool at a time and followed a predefined tool order.

4.3 Evaluation of Submitted Metadata

This section presents the metadata evaluation results based on the metrics defined in Section 3.5.4. Although 12 participants completed the study, this quantitative analysis includes only the nine participants who submitted valid metadata for all three tools. The remaining three participants were excluded because they had invalid submissions; for example, some submissions did not correspond to the KG specified in the task. This exclusion allowed for a direct comparison of participants' performance across the three conditions.

Figure 6 presents the precision, recall, and F1 scores across the three tools. The AI-assisted form achieved the highest precision, while the standard form and the Turtle editor showed lower and relatively similar precision scores. Precision was relatively consistent across participants for both forms, whereas the Turtle editor exhibited greater variability. The AI-assisted form significantly outperformed the standard form ($p = 0.005$). No statistically significant differences were observed between the AI-assisted form and the Turtle editor ($p = 0.083$), or between the standard form and the Turtle editor ($p = 0.675$).

More substantial differences were observed for recall and F1 score across the tools. The AI-assisted form achieved the highest scores and significantly outperformed the two other tools ($p < 0.001$ for all comparisons). The standard form also performed significantly better than the Turtle editor ($p = 0.002$ for recall; $p = 0.005$ for F1). These findings suggest that AI assistance helped participants capture a larger proportion of the required metadata.

Figure 7 shows the completeness and productivity results. The AI-assisted form achieved the highest completeness scores and significantly outperformed both the standard form ($p = 0.009$) and the Turtle editor ($p < 0.001$). The standard form also achieved significantly higher completeness than the Turtle editor ($p = 0.002$).

A similar pattern was observed for productivity. Participants using the AI-assisted form submitted more metadata values per minute than participants using the standard form ($p = 0.017$) and the Turtle editor ($p < 0.001$). The standard form also outperformed the Turtle editor ($p = 0.001$). Productivity was relatively consistent for the standard form

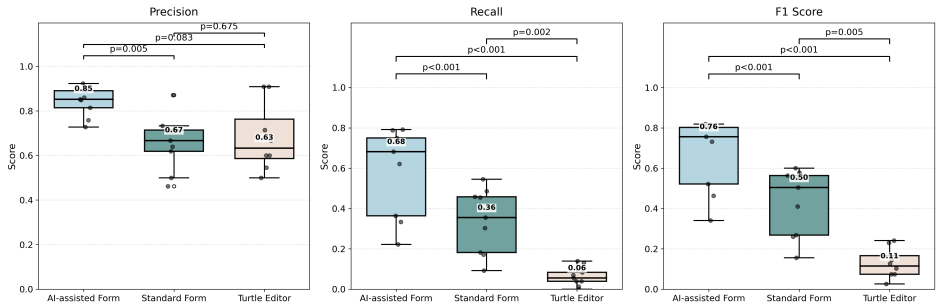

Figure 6. Precision, recall, and F1 scores across the three evaluated tools.

and the Turtle editor, whereas performance with the AI-assisted form varied more widely across participants.

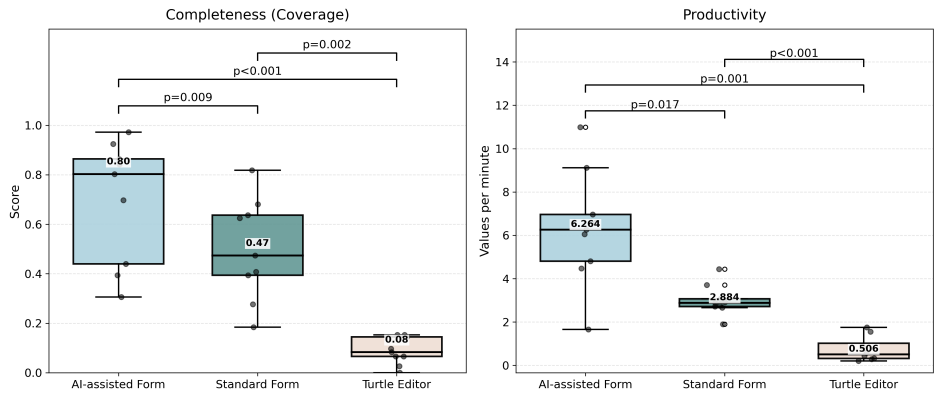

Figure 7. Completeness and productivity across the three evaluated tool

4.4 *Accuracy of AI Suggestions*

As described in Section 3.5.5, we analyzed the initial AI-generated suggestions to evaluate the accuracy of the model's outputs. Eleven of the twelve participants submitted valid task outputs for the AI-assisted tool, resulting in 11 suggestion sets. Table 7 reports the average counts of exact matches, unmatched values, and missing values for each KG. Overall, 72.4% of the suggestions exactly matched the reference values for Bio2RDF, 75.0% for YAGO, and 62.5% for QuoteKG. Among the missing values, some metadata elements were missing from the AI suggestions across all three KGs, namely primary source, conforms to, references, and SPARQL endpoint identifiers. Since the values for these elements were available in the KG documentation provided to the model, their absence suggests that the prompt may need further refinement to help the model identify and include these elements.

Table 7. Accuracy of AI suggestions per KG. TR denotes the total number of reference metadata values. EM, UM, and M denote the average counts of exact matches, unmatched values, and missing values in the AI suggestions, respectively. Elements commonly missing from the AI suggestions across all three KGs are highlighted in bold.

KG name (TR)	EM	UM	M	Missing elements and sub-elements
Bio2RDF (76)	55.0	2.0	19.0	**Primary source**; **Conforms to**; **references**. *Sub-elements:* **SPARQL endpoint** → (**identifier**)
YAGO (66)	49.5	7.5	9.0	primary reference; **conforms to**; access rights; **primary source**; other pages; **references**. *Sub-elements:* **SPARQL endpoint** → (**identifier**); publisher/creator → (role type, name, email)
QuoteKG (72)	45.0	6.0	21.0	**Primary source**; sample query; primary reference; **conforms to**; KG title; **references**. *Sub-elements:* publisher/creator → (role type, name, email); distribution → (title, description, media type, download URL, access URL); **SPARQL endpoint** → (**identifier**)

5 Discussion and Conclusion

To answer the RQ, we conducted a within-subjects study comparing the AI-assisted form with a version of the same form without AI assistance and a manual RDF/Turtle-based curation approach. The results show that the AI-assisted form achieved the highest scores across both objective and subjective evaluation measures.

The AI-assisted form received the most positive evaluations among the three conditions. Participants rated it significantly higher than both baselines in terms of SUS, PEOU, and BI. The qualitative analysis further supports these findings, revealing a clear shift in participants' perceptions across the tools, ranging from "frustrating" for the Turtle editor to "a great tool" and "very easy to use" for the AI-assisted form. In addition, comments such as "It was very useful to a novice" suggest that AI-assisted form may reduce the technical expertise required to complete KG metadata curation tasks, thereby making specification-compliant metadata creation more accessible to less experienced curators.

The objective evaluation results show a similar pattern. The AI-assisted form achieved significantly higher completeness, recall, and F1 scores, while maintaining high precision. In addition, it achieved the highest curation productivity. These results suggest that the AI-assisted form not only improved the curator experience but also supported the creation of more complete and specification-aligned metadata.

Taken together, these findings provide evidence that the AI-assisted form can improve KG metadata curation compared with both baselines. However, several limitations should be considered when interpreting these results.

First, the evaluation was conducted using only one LLM and within a controlled KG documentation setting. Further research should examine how different LLMs and prompting strategies affect suggestion correctness and completeness. Moreover, the capability of LLMs to extract KG metadata should be evaluated across multiple documentation sources with more complex structures, in order to better reflect real-world documentation scenarios. Beyond the specific evaluation setting used in this study, the approach is inherently dependent on the quality and coverage of the available documentation. In particular, when the documentation does not contain the information required by the target specification, the generated suggestions may be incomplete.

Second, the evaluation of the submitted metadata was limited by the small sample size, which reduced the statistical power of the analysis. As described in Section 4.3, this analysis was based on the nine participants, out of 12, who provided valid submissions across all three conditions. For example, although the AI-assisted form received a higher mean precision score than the Turtle editor, this difference was not statistically significant. In this comparison, the combination of a small sample size and relatively high variability in participants' precision scores, particularly in the Turtle editor condition, reduced the ability of the analysis to detect statistically significant differences. A larger study would therefore be needed to obtain more robust findings.

Third, the qualitative analysis relied on manual coding of statments, which may introduce a degree of subjectivity. Some participant statements could plausibly be assigned to multiple themes or interpreted in different ways. However, the identified themes were supported by a relatively large number of coded statements, many of which could be assigned to their respective themes with high confidence.

Fourth, the resulting correctness scores are influenced by the choice of similarity metrics within the type-aware similarity framework. Although the metrics used in this study were selected based on established practices, alternative assignments could also be justified and may lead to different quantitative outcomes. These alternatives were not explored in the present study and remain an important direction for future work. In addition, the correctness threshold of 0.7 was chosen as a practical compromise between strict exact matching and a more permissive criterion, such as 0.5. While this threshold provides a reasonable balance between strictness and flexibility, it is not derived from a principled optimization procedure. Future work should therefore investigate more robust methods for threshold selection. Nevertheless, threshold-based similarity was applied only to a small subset of metadata elements, whereas many elements were evaluated through exact matching. Consequently, its influence on the overall results is expected to be limited.

Finally, future work should refine the tool's implementation in line with participants' suggestions, particularly by improving interface clarity, supporting inline autocomplete functionality, and reducing document processing time.

In conclusion, this study shows that AI-assisted curation improves curator experience, increases metadata completeness, and makes specification-compliant KG metadata

creation more accessible. More broadly, this work contributes empirical evidence for the use of AI-assisted curation as a practical approach to supporting FAIR KG metadata.

Funding

This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 860801.

Supplemental material

All resources associated with this study are publicly available:

- Study materials, task outputs, and analysis scripts: https://github.com/MaastrichtU-IDS/AI-Assisted-Knowledge-Graph-Metadata-Curation
- Web-based tools: https://maastrichtu-ids.github.io/knowledge-graph-metadata-form/
- Source code for the tools: https://github.com/MaastrichtU-IDS/knowledge-graph-metadata-form.

References

1. Assaf A, Senart A and Troncy R. Roomba: automatic validation, correction and generation of dataset metadata. In: *Proceedings of the 24th International Conference on World Wide Web*, May 18, 2015, (pp. 159–162). DOI: https://doi.org/10.1145/2740908.2742827.
2. Alva Principe RA, Maurino A, Palmonari M , Ciavotta M, Spahiu B. ABSTAT-HD: a scalable tool for profiling very large knowledge graphs. *The VLDB Journal* 2022 Sep;31(5):851–876. DOI: https://doi.org/10.1007/s00778-021-00704-2.
3. Spahiu B, Porrini R, Palmonari M, Rula A, Maurino A. ABSTAT: ontology-driven linked data summaries with pattern minimalization. In: *European Semantic Web Conference* 2016 May 29 (pp. 381-395). Cham: Springer International Publishing. DOI: https://doi.org/10.1007/978-3-319-47602-5_51.
4. Čebirić Š, Goasdoué F, Kondylakis H, Kotzinos D, Manolescu I, Troullinou G, Zneika M. Summarizing semantic graphs: a survey. *The VLDB Journal* 2019 Jun 1;28(3):295-327. DOI: https://doi.org/10.1007/s00778-018-0528-3
5. Wang X, Cheng G, Lin T, Xu J, Pan JZ, Kharlamov E, Qu Y. PCSG: pattern-coverage snippet generation for RDF datasets. In: *International Semantic Web Conference* 2021 Sep 30 (pp. 3-20). Cham: Springer International Publishing. DOI: https://doi.org/10.1007/978-3-030-88361-4_1.
6. Safavi T, Belth C, Faber L, Mottin D, Müller E, Koutra D. Personalized knowledge graph summarization: from the cloud to your pocket. In: *IEEE International Conference on Data Mining (ICDM)* 2019 Nov 8 (pp. 528-537). IEEE. DOI: https://doi.org/10.1109/ICDM.2019.00063.
7. Schneider P, Schopf T, Vladika J, Galkin M, Simperl E, Matthes F. A decade of knowledge graphs in natural language processing: a survey. In: *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing* (Volume 1: Long Papers) 2022 Nov (pp. 601-614). DOI: https://doi.org/10.18653/v1/2022.aacl-main.46.

8. Ben Ellefi M, Bellahsene Z, Breslin JG, Demidova E, Dietze S, Szymański J, Todorov K. RDF dataset profiling: a survey of features, methods, vocabularies and applications. *Semantic Web* 2018 Aug 27;9(5):677-705. DOI: https://doi.org/10.3233/SW-180294.
9. Assaf A, Troncy R, Senart A. What's up LOD cloud? Observing the state of linked open data cloud metadata. In: *European Semantic Web Conference* 2015 May 31 (pp. 247-254). Cham: Springer International Publishing. DOI: https://doi.org/10.1007/978-3-319-25639-9_40
10. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data* Scientific data. 2016 Mar 15;3(1):1-9. DOI: https://doi.org/10.1038/sdata.2016.18
11. Finstad K. The system usability scale and non-native English speakers. *Journal of Usability Studies* 2006 Aug 1;1(4):185-8. https://uxpajournal.org/wp-content/uploads/sites/7/pdf/JUS_Finstad_Aug2006.pdf
12. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly* 1989 Sep 1;13(3):319-40. DOI: https://doi.org/10.2307/249008
13. Brooke J. SUS: a quick and dirty usability scale. *Usability Evaluation in Industry*. 1996 Jun 11;189(194):4-7.
14. Cheah WH, Jusoh NM, Aung MM, Ab Ghani A, Rebuan HM. Mobile technology in medicine: development and validation of an adapted system usability scale (SUS) questionnaire and modified technology acceptance model (TAM) to evaluate user experience and acceptability of a mobile application in MRI safety screening. *Indian Journal of Radiology and Imaging.* 2023 Jan;33(01):036-45. DOI: https://doi.org/10.1055/s-0042-1758198
15. Braun V and Clarke V. Using thematic analysis in psychology. *Qualitative Research in Psychology* 2006 Jan 1;3(2):77-101.
16. Mohammadi M, Sun C, Brewster C, Dumontier M. Knowledge graph metadata specification and validation. Data Science. 2026 Feb;9(1). DOI: https://doi.org/10.1177/24518492261436215.
17. Nault R, Cave MC, Ludewig G, Moseley HN, Pennell KG, Zacharewski T. A case for accelerating standards to achieve the FAIR principles of environmental health research experimental data. *Environmental Health Perspectives* 2023 Jun 23;131(6):065001. DOI: https://doi.org/10.1289/EHP11484. https://pmc.ncbi.nlm.nih.gov/articles/PMC10289218/
18. Busch L, Tebernum D and Velarde G. Exploring LLM capabilities in extracting DCAT-compatible metadata for data cataloging. *arXiv preprint arXiv:2507.05282* 2025. DOI: https://doi.org/10.48550/arXiv.2507.05282.
19. Dumontier M, Gray AJ, Marshall MS, Alexiev V, Ansell P, Bader G, Baran J, Bolleman JT, Callahan A, Cruz-Toledo J, Gaudet P. The health care and life sciences community profile for dataset descriptions. *PeerJ* 2016 Aug 16;4:e2331. DOI: https://doi.org/10.7717/peerj.2331
20. Lazar J, Feng JH and Hochheiser H. *Research methods in human-computer interaction* Morgan Kaufmann; 2017 Apr 28.
21. Wang X, Huey SL, Sheng R, Mehta S, Wang F. SciDaSynth: interactive structured data extraction from scientific literature with large language model. *Campbell Systematic Reviews* 2025 Dec;21(4):e70073. DOI: https://doi.org/10.1002/cl2.70073.
22. Tkaczyk D, Szostek P, Fedoryszak M, Dendek PJ, Bolikowski Ł. CERMINE: automatic extraction of structured metadata from scientific literature. *International Journal on Document*

Analysis and Recognition 2015 Dec;18(4):317-35. DOI: https://doi.org/10.1007/s10032-015-0249-8.

23. Dagdelen J, Dunn A, Lee S, Walker N, Rosen AS, Ceder G, Persson KA, Jain A. Structured information extraction from scientific text with large language models. *Nature Communications* 2024 Feb 15;15(1):1418. DOI: https://doi.org/10.1038/s41467-024-45563-x.
24. Sundaram SS, Solomon B, Khatri A, Laumas A, Khatri P, Musen MA. Use of a structured knowledge base enhances metadata curation by large language models. *arXiv preprint arXiv:2404.05893.* 2024 Apr 8. DOI: https://doi.org/10.48550/arXiv.2404.05893.
25. Gonçalves RS, O'Connor MJ, Martínez-Romero M, Egyedi AL, Willrett D, Graybeal J, Musen MA. The CEDAR Workbench: an ontology-assisted environment for authoring metadata that describe scientific experiments. *International Semantic Web Conference* 2017 Oct 4 (pp. 103-110). Cham: Springer International Publishing. DOI: https://doi.org/10.1007/978-3-319-68204-4_10.
26. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research Electronic Data Capture (REDCap): a metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of biomedical informatics* 2009 Apr 1;42(2):377-81. DOI: https://doi.org/10.1016/j.jbi.2008.08.010.
27. Nayak S, Zaveri A, Serrano PH, Dumontier M. Experience: automated prediction of experimental metadata from scientific publications. *Journal of Data and Information Quality (JDIQ)* 2021 Aug 12;13(4):1-1. DOI: https://doi.org/10.1145/3451219.
28. Ferenz S, Jafarbigloo A, Werth O, Nieße A. SMECS: a software metadata extraction and curation software. *arXiv preprint arXiv:2507.18159.* 2025 Jul 24. DOI: https://doi.org/10.48550/arXiv.2507.18159.
29. Hogan A, Blomqvist E, Cochez M, d'Amato C, Melo GD, Gutierrez C, Kirrane S, Gayo JE, Navigli R, Neumaier S, Ngomo AC. Knowledge graphs. *ACM Computing Surveys* 2021 Jul 2;54(4):1-37 DOI: https://doi.org/10.1145/3447772.
30. Martínez-Romero M, O'Connor MJ, Shankar RD, Panahiazar M, Willrett D, Egyedi AL, Gevaert O, Graybeal J, Musen MA. Fast and accurate metadata authoring using ontology-based recommendations. *AMIA Annual Symposium Proceedings* 2018 Apr 16 (Vol. 2017, p. 1272). https://pmc.ncbi.nlm.nih.gov/articles/PMC5977712/.
31. World Wide Web Consortium (W3C). Data Catalog Vocabulary (DCAT). https://www.w3.org/TR/vocab-dcat/
32. World Wide Web Consortium (W3C). The PROV Ontology (PROV-O). https://www.w3.org/TR/prov-o/
33. Dublin Core Metadata Initiative. Dublin Core Metadata Terms. https://www.dublincore.org/specifications/dublin-core/dcmi-terms/
34. W3C Interest Group Note. Vocabulary of Interlinked Datasets (VoID). 2011. https://www.w3.org/TR/void/
35. Schema.org. https://schema.org/