

SciGraphChat: Retrieval-Augmented Generation over Scientific Knowledge Graphs

Angelika Kaplan¹, Jan Keim¹, Xiyu Zhong¹ and Anne Koziolk¹

¹Karlsruhe Institute of Technology (KIT), Germany

Abstract

Effective retrieval and information synthesis of scholarly knowledge remain challenging because of fragmented publication practices, LLM hallucinations, and the limitations of keyword-based search. We present *SciGraphChat*, an operational system deployed as an *AI Assistant* within the NFDIxCS consortium, enabling discovery and exploration of scholarly knowledge represented in research knowledge graphs (KGs) through an integrated conversational and graph interface. The system builds upon a schema-agnostic, training-free retrieval approach for KG question answering, supporting source-aware and semantically informed access to structured scholarly data. By coupling KG-based Retrieval-Augmented Generation with a dialogue-oriented interface, *SciGraphChat* accompanies responses with retrieved RDF subgraphs and links to underlying resources. Unlike conventional chat-based assistants, the system discovers the retrieved graph evidence and enables users to interactively inspect and navigate the underlying knowledge structure through a dedicated graph panel. This paper presents the design and deployment of such a scholarly knowledge discovery system, along with practical design recommendations for integrating Semantic Web technologies and generative AI within research infrastructures. We evaluate the system with NFDIxCS stakeholders and external users via the System Usability Scale (SUS) and the User Experience Questionnaire–Short (UEQ-S), complemented by qualitative feedback. Results show a mean SUS score of 69.4, positive mean pragmatic quality (UEQ-S mean 0.93), and neutral hedonic quality (UEQ-S mean 0.70). Qualitative feedback indicates that users value the combination of conversational access and graph-based inspection, while also requesting tighter integration with scholarly workflows such as reference export, corpus scoping, and adaptive result presentation. Our findings suggest that combining conversational retrieval with transparent graph-based evidence inspection can improve trust, explainability, and usability in knowledge discovery.

Keywords

Retrieval-Augmented Generation, Knowledge Graphs, View-based Interface, FAIR Data Access, Semantic Web in Practice, Research Infrastructures, Knowledge Graph Usability

1. Introduction

Knowledge Graphs (KGs) have become a central infrastructure component for representing and inter-linking scholarly information on the Semantic Web [1]. In research data infrastructures, KGs integrate heterogeneous information sources and enable structured access to research artifacts, including publications, datasets, and software, across disciplinary boundaries [2]. Furthermore, KGs support the implementation of the FAIR principles [3] by enabling machine-actionable, interoperable, and richly interconnected research metadata. Despite these advantages, effective use of and interaction with KG-based systems remains a persistent practical barrier, as access mechanisms typically require familiarity with formal query languages, graph structures, or domain-specific ontologies [4], significantly constraining the broader adoption of Semantic Web technologies.

Interactive interfaces grounded in KGs via Retrieval-Augmented Generation (RAG) offer a promising solution: they combine the accessibility of natural language interaction with explicitly retrieved context from structured Semantic Web data, helping users to inspect the basis of generated answers and address concerns about purely generative systems [5]. Such approaches are particularly relevant for scholarly knowledge discovery, where users seek to explore complex semantic relationships among research artifacts without prior knowledge of the underlying schemas or ontologies.

6th International Workshop on Scientific Knowledge: Representation, Discovery, and Assessment, Oct 2026, Bari, Italy

✉ angelika.kaplan@kit.edu (A. Kaplan); keim@kit.edu (J. Keim); xiyu.zhong@student.kit.edu (X. Zhong); koziolk@kit.edu (A. Koziolk)

ORCID 0009-0009-9101-5833 (A. Kaplan); 0000-0002-8899-7081 (J. Keim); 0000-0002-1593-3394 (A. Koziolk)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

In this paper, we present *SciGraphChat*¹, a KG-driven AI assistant for scholarly knowledge discovery, deployed as a publicly accessible service within the NFDIXCS consortium (National Research Data Infrastructure for and with Computer Science) in Germany [6]. The system builds on HubLink [7], a schema-agnostic, training-free RAG approach for RDF KGs, and provides natural-language access to a curated scholarly KG built upon the Open Research Knowledge Graph (ORKG) [8], currently interlinking software architecture research artifacts [7, 9]. By combining conversational question answering with interactive graph exploration, SciGraphChat supports both the discovery and validation of scholarly knowledge represented in Semantic Web infrastructures. Thus, retrieved RDF subgraphs are presented alongside generated responses, while a tightly integrated visual graph panel enables users to interactively navigate the underlying KG in the context of the retrieved and generated answer.

Beyond presenting a deployed system, this work investigates how scholarly KGs and Semantic Web technologies can support practical knowledge discovery in operational research infrastructures. While usability is a widely recognized bottleneck for KG adoption, empirical evidence from operational deployments targeting non-experts remains scarce, especially for interfaces that deliberately combine conversational answers with explicit graph exploration. To address this gap, we conducted a user study employing the System Usability Scale (SUS) [10, 11] and the User Experience Questionnaire Short (UEQ-S) [12], complemented by qualitative open-ended feedback. Both questionnaires are standardized, widely validated instruments for summative evaluation of interactive systems.

The contributions of this paper are threefold:

- (i) the design and operational deployment of a knowledge graph-based assistant that links conversational answers with interactive inspection of retrieved RDF subgraphs;
- (ii) evidence from a multi-stage evaluation of perceived usability and user experience with domain experts, infrastructure stakeholders, and external researchers, combining SUS, UEQ-S, and qualitative feedback; and
- (iii) available study materials² and design implications for deploying natural language interfaces with inspectable KG context in Semantic Web infrastructures.

The remainder of this paper is structured as follows. Section 2 situates our work within foundational and related literature. Section 3 presents SciGraphChat’s system design. Section 4 details the evaluation design. Results are reported in Section 5. Section 6 discusses the implications of the findings, lessons learned from the deployment, and opportunities for future improvements. Section 7 addresses threats to validity. Section 8 concludes with an outlook on future work.

2. Background and Related Work

This section positions SciGraphChat within prior work, covering information overload in academia (Section 2.1), KG-based user interfaces (Section 2.3), conversational interfaces for scholarly knowledge (Section 2.3), and usability evaluation methods (Section 2.4).

2.1. Information Overload in Academia

The continuously growing volume of scholarly publications, datasets, and software artifacts has intensified information overload in academia [8]. Conventional keyword-based search engines, while widely used, capture neither semantic relationships nor complex information needs, leaving researchers with large, weakly structured result sets that require substantial manual effort.

While LLMs offer natural language access to scholarly information, they raise well-documented concerns about hallucination and transparency [5], which conflict with scholarly requirements for traceability and reproducibility. KGs address these limitations by providing structured, provenance-aware access to scholarly information [1]. Furthermore, RAG approaches combine the linguistic capabilities of LLMs with KG retrieval, grounding responses in verified data sources [13, 7].

¹<https://nfdixcs.org/services/nfdixcs-chat-assistant> [Last accessed on 2026-08-05]

²Materials available at <https://doi.org/10.6084/m9.figshare.32216784>

2.2. Knowledge Graph–Based User Interfaces

KGs, serialized using W3C RDF standards³, enable the representation and interlinking of heterogeneous scholarly data using shared vocabularies and formal semantics [1, 2]. The vision of interconnected, navigable scholarly knowledge traces back to Vannevar Bush’s “Memex” [14]. Modern KG-based systems realize many aspects of this vision, yet still require users to understand graph structures, ontologies, or query languages such as SPARQL, creating a persistent usability barrier.

Existing interfaces such as faceted browsers (e.g., Wikidata, OpenAIRE) and graph visualization tools lower this barrier incrementally, but presuppose familiarity with the underlying data model and do not support expressive, open-ended natural language queries [5]. Consequently, the combination of LLMs and KGs is attracting increasing interest as a path toward interfaces that are both intuitive and semantically grounded.

2.3. Conversational Interfaces for Scholarly Knowledge

Researchers have studied natural language interfaces for KGs as Knowledge Graph Question Answering (KGQA), aiming to map user questions to structured queries or KG triples [5]. Early systems relied on template-based or semantic-parsing approaches, effective for well-defined question types but limited in handling lexical diversity and open-ended queries. Recent RAG-based architectures address these shortcomings. Systems such as ASK-DBLP [4] demonstrate natural language access to bibliographic RDF KGs. SciGraphChat employs HubLink [7] as its KGQA backend, a schema-agnostic, training-free retrieval approach that requires neither SPARQL generation nor schema-specific fine-tuning, enabling transferability across heterogeneous scholarly KGs.

Despite these advances, conversational interfaces for scholarly KGs remain largely prototypical, with few production deployments in real research infrastructures and most evaluations conducted under controlled laboratory conditions.

2.4. Usability Evaluation of Semantic Web Systems

Usability is a well-recognized bottleneck hindering the adoption of Semantic Web technologies beyond expert users, yet empirical evaluations of deployed systems with non-expert populations remain scarce [15, 16]. Most published work focuses on functional benchmarks rather than subjective user experience.

Standardized instruments address this gap. The System Usability Scale (SUS) [10, 11] yields a single composite score that facilitates cross-system comparison, while the User Experience Questionnaire Short (UEQ-S) [12, 17] separately captures pragmatic and hedonic quality. Both are instrument-independent, require no domain expertise, and provide established norm data for contextualized interpretation, making them well-suited for evaluating operational Semantic Web applications, as in this work. To support transparent interpretation and comparability across studies, results are typically summarized using descriptive statistics, confidence measures, and reliability indicators. The following metrics are commonly reported and interpreted. The mean is the primary summary measure of overall usability or user experience, complemented by the median to account for skewed distributions. The standard deviation reflects response variability, while minimum and maximum values indicate the observed score range and help identify extreme ratings.

3. SciGraphChat: System Overview

SciGraphChat is a conversational interface for scholarly knowledge graphs, deployed as an operational service within the NFDIxCS research infrastructure for the computer science community. It integrates RAG [13] with an extended version of the HubLink retrieval approach [7] for schema-agnostic, training-free question answering over RDF knowledge graphs. Each response includes an explicitly retrieved RDF subgraph and links to underlying resources, supporting inspection of the information used to

³<https://www.w3.org/RDF/> [Last accessed on 2026-08-05]

generate the answer [3]. The interface combines a *Chat Panel* for natural language interaction with a *Graph Panel* that enables interactive exploration of retrieved knowledge graph structures.

3.1. Use Case and Target Users

SciGraphChat targets researchers, research software engineers, and practitioners seeking intuitive access to scholarly metadata without requiring prior expertise in SPARQL or knowledge graph technologies. Typical use cases include discovery of research artifacts in the available corpus and exploration of their relationships, topics, methods, and authorship networks. The system is designed for iterative and exploratory information seeking rather than single-shot fact retrieval. The system returns a natural language answer accompanied by an interactive visualization of the retrieved knowledge graph neighborhood and links to the underlying resources.

3.2. Technical Architecture

SciGraphChat connects a natural language web frontend to an RDF knowledge graph backend through a RAG pipeline. The underlying scholarly knowledge graph encodes publications, datasets, software, authors, venues, and research topics as RDF triples. The HubLink retrieval engine [7] processes user queries and extracts relevant subgraphs in a schema-agnostic, training-free manner, eliminating the need for SPARQL query generation and schema-specific fine-tuning. The retrieved subgraphs are serialized as grounding context for a large language model, which synthesizes a natural language response anchored in the structured data. The subgraphs are also embedded in the answer and rendered in an interactive graph visualization, allowing users to explore the retrieved information in both textual and visual modalities. Links to the underlying knowledge graph resources are surfaced to users as textual annotations and clickable links, enabling users to inspect the retrieved context for each answer and supporting FAIR information access [3, 2]. The system is deployed as a continuously available web service within the NFDIxCs infrastructure, with backend error handling and safety guardrails to support reliable operation in real-world usage scenarios.

3.3. User Interaction Model

The interface comprises four main components (see Figure 1): a *Header* for language and theme settings; a *Sidebar Panel* for managing and revisiting persistent conversation sessions; a *Chat Panel* for natural language interaction; and a *Graph Panel* for interactive knowledge graph exploration.

The *Chat Panel* (text view of research artifacts), shown on the left side of Figure 2, returns, for each query in natural language, a response grounded in retrieved scholarly metadata, with direct access to the underlying subgraph, related research artifacts, and provenance links to original resources. Sessions can be shared via persistent links and exported in Markdown, HTML, or PDF. During answer generation, the user is informed about the retrieval and generation process through a progress indicator. Request–response handling, i.e., from the initial user query to the final generated answer, is supported by robust backend error-handling mechanisms of the retrieval strategies that address retrieval, reasoning, and generation failures. In addition, guardrails such as filtering for toxic language or out-of-scope queries are continuously integrated to ensure responsible and safe system behavior.

The *Graph Panel* (graph view of metadata about research artifacts), shown on the right side of Figure 2, provides an interactive environment for navigating entities, relationships, and graph-based search, linked directly to conversational responses for seamless transitions between textual and graph-structured representations. Moreover, users can expand the retrieved subgraph with double-click interactions, which trigger additional retrievals to incrementally explore the knowledge graph neighborhood around the initial query results, enabling deeper exploration of the semantic network.

The interface adheres to established web *accessibility* guidelines [18], supports keyboard navigation, multi-language use, configurable themes (e.g., light & dark mode), and *responsive* layouts across devices from smartphones to large screens. These design features promote accessibility and inclusiveness while reducing barriers to adoption for diverse international research communities.

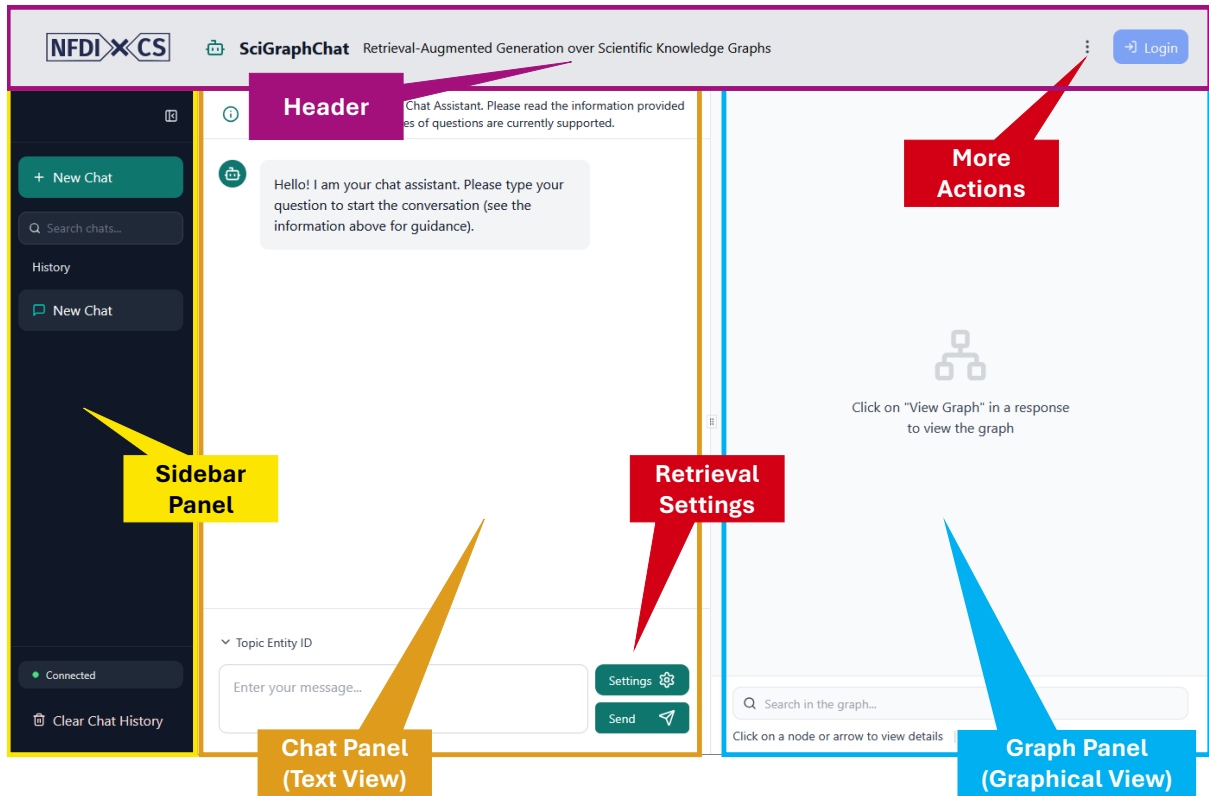


Figure 1: SciGraphChat UI: Header (top), Sidebar Panel for session management (left), Chat Panel with natural language responses grounded in retrieved RDF subgraphs (center), and Graph Panel for knowledge graph exploration (right).

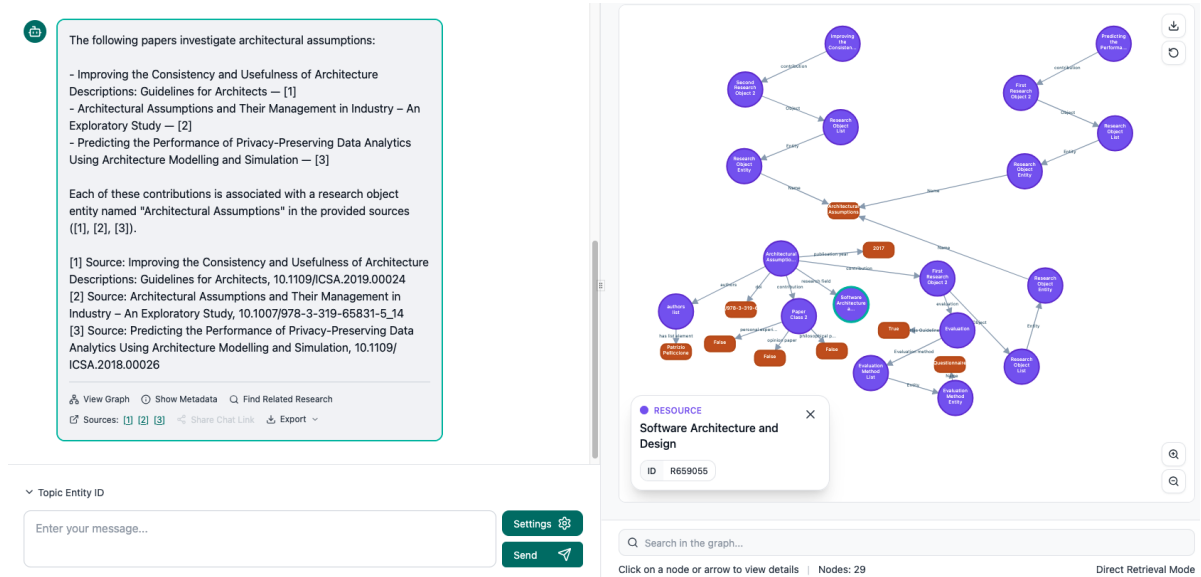


Figure 2: SciGraphChat with an example answer along with the corresponding graph.

In summary, users can alternate between asking questions in the Chat Panel and exploring retrieved subgraphs in the Graph Panel, combining conversational and visual modalities for transparent, interpretable access to semantic scholarly data. The production deployment within NFDIxCs provides the empirical foundation for the usability evaluation reported in the following sections.

4. Evaluation Method

We evaluate the perceived usability and user experience of SciGraphChat across multiple user groups and realistic research contexts. Our evaluation adopts a mixed-methods strategy, combining standardized questionnaires with qualitative feedback to both quantify usability and surface contextual insights into interaction challenges and adoption barriers.

4.1. Evaluation Goals and Research Questions

We structure the evaluation using the Goal–Question–Metric (GQM) approach [19, 20]. The overall evaluation goal is to analyze SciGraphChat, integrating a *chat interface* with *graph exploration* UI components, for the purpose of assessing perceived *usability* with respect to *overall usability*, *pragmatic quality*, and *hedonic quality*, from the viewpoint of end users (researchers and practitioners), in the context of a *guided system walkthrough* of the SciGraphChat service within research data infrastructures.

This goal is operationalized through the following research questions:

RQ1: How usable is SciGraphChat for KG-based information exploration using a combined chat and graph interface?

RQ2: What are the perceived benefits for information-seeking tasks?

RQ3: Which usability aspects present challenges for users?

4.2. Survey Design

We combined two complementary standardized questionnaires with open-ended questions. Standardized scales provide scores comparable across systems and studies, while open-ended questions capture domain-specific usability issues and insights that quantitative measures alone cannot reflect.

The System Usability Scale (SUS) [10, 11] is a ten-item instrument yielding a single composite usability score. Its well-established benchmark of ≥ 68 for above-average usability enables direct comparison of SciGraphChat against a broad reference population of interactive systems (RQ1).

The User Experience Questionnaire – Short Version (UEQ-S) [21] measures user experience along the two dimensions *pragmatic quality* (usefulness, efficiency, clarity) and *hedonic quality* (stimulation, interest, novelty), which is particularly suited here, as the chat and graph exploration modalities may exhibit divergent UX profiles (RQ1). Values ≥ 0.8 are interpreted as clearly positive.

The Open-Ended Questions were included as optional free-text fields in all studies to elicit qualitative feedback (RQ2 and RQ3). Participants were asked to reflect on: 1. aspects of the chat interface they found helpful or problematic; 2. aspects of the graph-based exploration they found helpful or problematic; and 3. any further comments or suggestions.

4.3. Study Design

We conducted two studies. The first study served as a formative evaluation with domain experts attending a scientific workshop, whereas the second study involved stakeholders of the deployed system as well as academic researchers and practitioners beyond the consortium. Both studies employed the same survey instruments (see Section 4.2) within a guided system walkthrough to ensure comparability.

Pre-Study at SE2026 As a formative first step, we conducted a pre-study at the software engineering conference SE2026⁴ during a co-located workshop on research data management. Participants were active researchers in areas relevant to the system’s domain. The objective was to gather early expert feedback and identify interaction issues before wider deployment.

⁴<https://se2026.inf.unibe.ch/de/> [Last accessed on 2026-08-05]

Survey with NFDIxCS Members and External Researchers The second study targeted members of the NFDIxCS consortium, i.e., the direct institutional stakeholders of the deployed SciGraphChat service, and academic researchers and practitioners outside the consortium without prior exposure to the system. This combined cohort allowed us to collect feedback from intended service stakeholders and users with diverse scholarly information-seeking backgrounds.

4.4. Participants and Evaluation Context

Across both studies, we recruited participants whose work relates to scholarly information seeking, research data management, software engineering, or Semantic Web infrastructures. The pre-study included seven active software engineering researchers who attended a research data management workshop. The main study comprised 32 completed responses from four expert- and domain-oriented groups: NFDIxCS consortium stakeholders familiar with the target infrastructure, academic researchers, research software or infrastructure practitioners, and external users without prior exposure to SciGraphChat. Thus, the sample combines intended service stakeholders with users who can assess the system from a broader scholarly search perspective. We did not collect fine-grained demographic attributes or formal technical proficiency scores; consequently, subgroup comparisons are treated as contextual interpretation rather than inferential evidence.

4.5. Data Analysis

We scored quantitative data from the SUS and UEQ-S following their respective standard procedures. We interpreted SUS scores against established usability benchmarks. We analyzed UEQ-S scores separately for pragmatic and hedonic quality dimensions.

Qualitative free-text responses from both studies were analyzed using an inductive thematic grouping approach to identify recurring benefits, challenges, and improvement opportunities. These themes contextualize the quantitative findings and address RQ2 and RQ3.

5. Evaluation Results

This section presents the results of the two studies introduced in Section 4. We first report the findings of the pre-study conducted at the SE2026 workshop (Section 5.1), followed by the results of the main survey involving NFDIxCS consortium members, external researchers, and practitioners (Section 5.2).

5.1. Results from the Pre-study

The pre-study assessed perceived usability (RQ1) and surfaced initial insights into perceived benefits (RQ2) and interaction challenges (RQ3) via the SUS and open-ended free-text questions. Due to the workshop setting, the sample was small ($n = 7$). The results should therefore be interpreted as formative rather than confirmatory evidence.

Quantitative Results (RQ1) The mean SUS score was 77.1 (median: 77.5, SD: 7.8), which exceeds the established above-average threshold of 68 [10, 11] and places SciGraphChat firmly in the *good* usability band. Individual scores ranged from 67.5 to 90.0. Applying the standard adjective-rating scale [11], 42.9% of participants (3/7) rated the system as *excellent* (≥ 80), 42.9% (3/7) as *good* (68–79), and 14.3% (1/7) as *acceptable* (50–67). No participant scored the system below 50, indicating no fundamental usability barriers even in a first-contact walkthrough scenario.

SUS Item-Level Analysis Item-level SUS analysis revealed strong acceptance, ease of use, and learnability, with the highest ratings for frequent intended use and system simplicity. The lowest-scoring item concerned initial learning effort, suggesting opportunities to improve onboarding and user guidance rather than address core usability issues.

Qualitative Feedback Free-text responses were predominantly positive. Participants characterized SciGraphChat in general as a promising and valuable tool for scholarly search, with one respondent noting it would be “very valuable and appreciated by users.” The thematic analysis identified a cross-cutting concern regarding knowledge graph data quality: participants observed that manual metadata annotation is frequently incomplete in practice and does not scale, drawing an explicit parallel to the well-documented limitations of collaborative tagging in earlier social Web platforms [22]. This finding is particularly relevant for the semantic web community: even when a system’s query and retrieval pipeline performs well, the quality of the underlying knowledge graph remains a critical determinant of answer quality and user trust. Participants proposed three categories of mitigation: (i) automated metadata extraction to reduce manual curation burden; (ii) pre-populated annotation fields to shift the user role from authoring to reviewing; and (iii) streamlined contribution workflows to lower participation barriers overall.

Summary and Implications The pre-study establishes SciGraphChat’s baseline usability within the *good* range ($\bar{x}_{\text{SUS}} = 77.1$) and confirms its perceived relevance for scholarly information-seeking among domain experts (RQ1, RQ2). The principal actionable finding, i.e., reduced user confidence despite otherwise positive usability perceptions, is attributable to insufficient capability transparency rather than core interaction design failures (RQ3). This informed two concrete iterations before the main study: introducing example queries and a guided onboarding flow, and running a requirements elicitation session to derive user stories and competency questions for subsequent feature development.

5.2. Results from the Main Study

Building on the strong positive results of the pre-study, the main study employed an extended survey instrument comprising the SUS, the UEQ-S, and a set of more detailed qualitative questions. A total of 32 participants completed the questionnaire. We calculated internal consistency metrics and descriptive statistics for all measures.

Table 1

Descriptive statistics and reference-based interpretation for SUS and UEQ-S (N = 32).

SUS scores range from 0 to 100, with 68 commonly used as an above-average reference benchmark. UEQ-S pragmatic and hedonic quality scores range from −3 (most negative) to +3 (most positive); mean values above 0.8 are commonly interpreted as positive.

Statistic	SUS	UEQ-S Pragmatic Qual.	UEQ-S Hedonic Qual.
Mean	69.4	0.9297	0.7031
Median	72.5	1.25	0.75
Standard Deviation	15.1	0.857	1.031
Minimum	25.0	-1.5	-2
Maximum	95.0	2.25	3
Interpretation	Mean above benchmark	Positive mean	Neutral mean

Quantitative Results (RQ1) Table 1 summarizes the quantitative results. Across all 32 participants, SciGraphChat achieved a mean SUS score of 69.4 (median: 72.5, SD: 15.1). This mean is slightly above the commonly cited above-average reference benchmark of 68 [10, 11], although the 95% confidence interval includes that benchmark (see Table 1). UEQ-S pragmatic quality averaged 0.93 (median: 1.25, SD: 0.86), above the 0.8 reference value commonly used for a positive mean rating. Its confidence interval includes 0.8, so interpret this finding cautiously. Hedonic quality scored a mean of 0.70 (median: 0.75, SD: 1.03), falling just below the threshold and thus rated as neutral. The wide spread (min: −2, max: 3) reflects considerable variation in how stimulating and novel participants perceived the system. For professional, task-oriented tools, this is a common and expected pattern: pragmatic quality tends to be the primary driver of adoption and satisfaction, while hedonic quality matures with the system

over time [23]. The results indicate a promising functional foundation, with enriching the interaction experience as a direction for future development.

Qualitative Feedback (RQ2 + RQ3) Three open-ended questions elicited qualitative feedback on perceived benefits (RQ2) and interaction challenges (RQ3), covering the two main UI components, i.e., the chat interface and the graph panel, as well as general impressions. We analyze each component in turn before synthesizing an overall interpretation.

Overall, participants found the *chat interface* intuitive and easy to use, largely because it resembles familiar AI chatbots and GenAI interaction patterns. This familiarity enabled quick onboarding and supported flexible use across devices, which participants saw as particularly beneficial in mobile or constrained contexts. Features such as inline graph previews attached to each chat response (see Figure 1), progress indicators, and metadata display were highlighted as helpful for understanding system behavior and tracking retrieval processes. These inline previews were specifically valued for enabling exploration of relationships between research artifacts such as papers and supplementary materials directly within the conversational flow. Reported usability challenges centered on missing academic workflow features, notably reference export options (e.g., BibTeX or RIS) and clearer, more extensive explanations of retrieval configuration settings within the chat panel.

Regarding the *graph panel*, participants generally viewed graph-based exploration as a valuable complement to chat-based interaction, particularly for contextualizing responses and revealing entity relationships that may not be apparent in text alone. Visual cues such as distinct node colors and shapes were perceived as helpful, and several users appreciated response-specific subgraphs that surface only the most relevant nodes and relations, supporting focused sense-making without visual overload. Suggested improvements included layering or clustering of similar nodes, clearer labeling of graph views, and alternative visual encodings for research artifact attributes.

Some responses to the first two questions could not be attributed exclusively to either the chat panel or the graph panel. They instead reflected the integrated interaction experience as a whole. Participants perceived this tight coupling as complementary: conversational answers supported initial information seeking, while graph exploration supported inspection and follow-up. The study does not isolate the effects of either modality. Participants also welcomed the ability to extend the initial response-specific subgraph for exploratory follow-up, underscoring the value of progressive graph disclosure as an interaction strategy.

General feedback also highlighted the importance of interoperability with established scholarly infrastructure: participants called for integration with literature databases such as ACM DL, IEEE Xplore, and GI DL, as well as corpus scoping by venue or source. From a Linked Data perspective, these requests highlight the need for federated knowledge graph access and cross-repository entity alignment as concrete next steps. Additionally, participants noted that certain result types (e.g., ranked lists or comparative overviews) would benefit from tabular rendering, suggesting that adaptive, content-aware result presentation is desirable.

Summary and Implications Overall, the main study provides encouraging evidence that combining a conversational interface with knowledge graph exploration resonates with researchers and practitioners as a way to access scholarly information (RQ2). The perceived complementarity between the chat and graph panels stands out as a key result: users experienced the two views as a coherent whole. Controlled comparisons are needed to establish each component’s separate contribution. On the challenge side (RQ3), participant feedback converged on two practical gaps. The first is interoperability: participants frequently noted the absence of reference export formats such as BibTeX and the inability to scope queries to specific venues or databases, pointing to a need for tighter coupling with existing scholarly infrastructure and Linked Data repositories. The second is presentation flexibility: participants found graph visualization effective for relational exploration but noted that other result types, such as ranked lists or comparative tables, call for different representations. Addressing both gaps constitutes a natural agenda for the next iteration of the system.

6. Discussion

The evaluation positions SciGraphChat less as a new KGQA algorithm and more as an operational Semantic Web system that makes LLM and existing KG retrieval capabilities usable in a research infrastructure setting. This distinction is central to the contribution: the technical novelty lies not in replacing SPARQL, RDF, or RAG components, but in integrating schema-agnostic retrieval, provenance-aware generation, and interactive graph inspection into a deployed service and evaluating how representative end users perceive this combination.

The qualitative results suggest that the chat and graph panels contribute different forms of value and support complementary aspects of scholarly information seeking. The chat panel lowers the entry barrier by allowing users to articulate information needs in natural language and receive concise, synthesized responses. The graph panel supports inspection, contextualization, and exploratory sense-making by exposing the retrieved RDF relationships underlying those answers. Participants' comments on inline graph previews, response-specific subgraphs, and progressive graph expansion suggest that the two interaction modalities support complementary phases of scholarly information seeking, from initial articulation to verification and exploration. The present study does not establish their separate causal contributions.

The neutral hedonic UEQ-S score and the wide SUS range point to a more nuanced interpretation than a uniformly positive usability result. Users who expected polished scholarly workflow support missed features such as BibTeX export, corpus scoping, and alternative result presentations for ranked or comparative outputs. These lower-scoring perceptions are informative because they separate core interaction viability from adoption requirements in daily academic work. For Semantic Web systems in practice, the finding suggests that provenance-aware answers and graph transparency are necessary but not sufficient. Sustained use requires integration with established literature databases, reference management workflows, and task-specific views.

Finally, our evaluation deliberately focused on perceived usability and user experience rather than answer correctness, latency, throughput, or ablation of individual components. These technical dimensions remain important follow-up work. The present study therefore provides deployment-grounded usability evidence and concrete design implications, while future evaluations should complement it with controlled comparisons against chat-only, graph-only, and SPARQL-based baselines, as well as capability metrics for factual correctness, completeness, provenance quality, and response time.

7. Threats to Validity

For our evaluation, we used SUS, UEQ-S, and optional qualitative feedback collected through three open-ended questions. The analysis was based on 32 completed responses obtained from four expert- and domain-oriented participant groups. While the results indicate overall positive usability and user experience, several threats to validity must be considered.

Construct validity is limited because SUS and UEQ-S capture subjective perceptions rather than objective aspects. Qualitative feedback suggests that some participants incorporated perceived functional properties or envisioned system capabilities into their ratings, potentially conflating usability judgments with assessments of system functionality and value.

Internal validity may be affected by presentation and explanation effects inherent in walkthrough-based evaluations. Although all sessions followed a common structure and standardized presentation materials, the walkthroughs were conducted separately for each group, and allowing questions may have influenced participants' understanding and perceptions. The SUS results show moderate dispersion ($SD = 15.1$) and a wide observed range (25–95), indicating heterogeneous perceptions, likely related to differences in professional background, expectations, and evaluation focus.

Reliability of the SUS instrument is well established in the literature [10, 11]. All participants completed the scale's ten items without missing responses. UEQ-S results show positive mean values for both pragmatic ($M = 0.93$) and hedonic quality ($M = 0.70$). However, reliability is constrained by the

overall sample size and small subgroup sizes, as well as notable variability in UEQ-S responses, including negative minimum values. For context, in-use evaluations of Semantic Web systems typically operate with expert samples in the range of ten to thirty participants [24], where recruitment is constrained by the specialized nature of the target audience; our sample of $N = 32$ is well within this range. Optional qualitative feedback further limits consistency, as not all participants provided comments.

External and conclusion validity are limited by the expert-oriented sample and the short-term, mediated evaluation setting. Consequently, quantitative results and benchmark interpretations should be viewed as indicative rather than definitive, supported but not fully substantiated by qualitative insights.

8. Conclusion and Future Work

In this paper, we presented SciGraphChat, an operational conversational interface that supports natural language access to an ORKG-based scholarly knowledge graph (KG) within the NFDIxCS research infrastructure. SciGraphChat combines Semantic Web technologies with retrieval-augmented generation to support transparent, grounded, and user-friendly interaction with structured scholarly metadata. By integrating conversational querying with visual KG exploration, the system reduces access barriers while preserving explainability and trustworthiness for scholarly knowledge discovery.

To assess the system in practice under realistic usage conditions, we conducted a multi-stage empirical evaluation across different contexts and stakeholder groups using standardized questionnaires (SUS and UEQ-S), complemented by qualitative feedback. The mean SUS score in the main study was slightly above a commonly cited reference benchmark, while mean pragmatic quality was positive and hedonic quality neutral. Participants reported that they value the conversational access paradigm and the inspectable graph context for scholarly information seeking. At the same time, the evaluation highlighted adoption requirements beyond core usability, particularly reference export, corpus scoping, adaptive result presentation, and clearer guidance during interaction.

From a broader perspective, this work contributes empirical evidence and practical insights on the usability and adoption of Semantic Web systems in real-world research infrastructures. The findings underline the importance of combining natural language interfaces with explicit semantic grounding and transparent interaction mechanisms when deploying KG-based systems for non-expert users. Moreover, the reported design implications can inform similar applications that aim to make Semantic Web technologies more accessible in practice and to enable AI-assisted access to scholarly knowledge.

Several directions for future work follow from these findings. First, while SciGraphChat currently operates on a single scholarly KG, a natural extension is the exploratory integration of additional RDF-based KGs, which would broaden the scope of queryable scholarly artifacts and stress-test the schema-agnostic design of the underlying HubLink retrieval approach across heterogeneous graph structures. Such an extension would further support cross-graph scholarly knowledge discovery and interoperability between research KGs. Second, the present usability evaluation was constrained by a relatively small participant sample; future studies should involve larger and more diverse user cohorts to obtain more statistically robust SUS and UEQ-S scores and to capture a wider range of usage patterns. Third, beyond usability, future evaluations should further assess the *answer quality* of the system, covering dimensions such as factual correctness, completeness, and provenance transparency, ideally through a dedicated benchmark over the deployed KG. With researchers using the system in a real-world context, this may also allow capturing the different (types of) questions to create a realistic benchmark dataset for future RAG-based KGQA systems. Finally, personalization features, such as query suggestions based on conversation history, the incorporation of chat memory, and user-specific graph and chart views, could further reduce interaction barriers and support more targeted and efficient scholarly discovery workflows.

Overall, the results of our usability study suggest that KG-driven AI assistants can lower access barriers to scholarly KGs by coupling conversational interaction with transparent and inspectable retrieved context of research artifacts in operational research infrastructures.

Acknowledgments

This work is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under the National Research Data Infrastructure – NFDI 52/1 – project number 501930651, NFDIxCs, supported by funding from the pilot program Core Informatics at KIT (KiKIT) of the Helmholtz Association (HGF) and supported by KASTEL Security Research Labs.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-5.5, Claude Sonnet 4.6, and Grammarly only for language editing, including spelling and grammar correction. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

- [1] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. de Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, A. N. Ngomo, A. Polleres, S. M. Rashid, A. Rula, L. Schmelzeisen, J. F. Sequeda, S. Staab, A. Zimmermann, Knowledge graphs, *ACM Computing Surveys* 54 (2022) 71:1–71:37. doi:10.1145/3447772.
- [2] L. Rossenova, F. Limani, C. Fortmann-Grote, A. Kaplan, F. Thiery, A. Latif, R. Shigapov, B. Zapilko, M. Schubotz, H. Fliegl, T. Tietz, How are nfdi consortia using knowledge graphs? an overview of common functions and challenges, in: *The 2nd Conference on Research Data Infrastructure (CoRDI), Section (Meta)data, Terminologies, Provenience / Working Group "Knowledge Graphs"*, 2025. URL: <https://zenodo.org/records/16736078>, accessed: 2025-09-27.
- [3] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al., The fair guiding principles for scientific data management and stewardship, *Scientific data* 3 (2016) 1–9. doi:10.1038/sdata.2016.18.
- [4] T. A. Taffa, P. Neises, S. Ollinger, P. Westphal, M. R. Ackermann, D. Banerjee, R. Usbeck, ASK-DBLP: answering questions over DBLP, in: I. Celino, O. Hassanzadeh, A. Bernstein, N. F. Noy, G. Cheng, S. Wang, S. Ferrada, T. Souldard, K. Kozaki, H. Takeda, A. L. Gentile (Eds.), *Joint Proceedings of Industry, Doctoral Consortium, Posters and Demos of the 24th International Semantic Web Conference (ISWC-C 2025)*, Nara, Japan, November 2-6, 2025, *CEUR Workshop Proceedings*, CEUR-WS.org, 2025, pp. 455–461. URL: <https://ceur-ws.org/Vol-4085/paper76.pdf>.
- [5] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, X. Wu, Unifying large language models and knowledge graphs: A roadmap, *IEEE Transactions on Knowledge and Data Engineering* 36 (2024) 3580–3599. doi:10.1109/TKDE.2024.3352100.
- [6] M. Goedicke, U. Lucke, Research data management in computer science - nfdixcs approach, *INFORMATIK 2022*, 2022. doi:10.18420/inf2022_112.
- [7] A. Kaplan, J. Keim, M. Schneider, R. Reussner, Hublink: A novel question answering retrieval approach over knowledge graphs, in: *RAGE-KG 2025: The Second International Workshop on Retrieval-Augmented Generation Enabled by Knowledge Graphs (RAGE-KG 2025) co-located with 24th International Semantic Web Conference (ISWC 2025)*. Ed.: D. Dobrij, 2025, p. 1. URL: <https://ceur-ws.org/Vol-4079/paper1.pdf>, this Paper received the Best Paper Award.
- [8] S. Auer, A. Oelen, M. Haris, M. Stocker, J. D'Souza, K. E. Farfar, L. Vogt, M. Prinz, V. Wiens, M. Y. Jaradeh, Improving access to scientific literature with knowledge graphs, *Bibliothek Forschung und Praxis* 44 (2020) 516–529. doi:10.1515/bfp-2020-2042.
- [9] M. Konersmann, A. Kaplan, T. Kühn, R. Heinrich, A. Koziolk, R. H. Reussner, J. Jürjens, M. al-Doori, N. Boltz, M. Ehl, D. Fuchß, K. Großer, S. Hahner, J. Keim, M. Lohr, T. Saglam, S. Schulz, J. Töberg, Evaluation methods and replicability of software architecture research objects, in: *19th IEEE International Conference on Software Architecture, ICSA 2022, Honolulu, HI, USA, March 12-15, 2022*, IEEE, 2022, pp. 157–168. doi:10.1109/ICSA53651.2022.00023.

- [10] J. Brooke, Sus-a quick and dirty usability scale, *Usability evaluation in industry* 189 (1996) 4–7.
- [11] J. R. Lewis, The system usability scale: past, present, and future, *International Journal of Human–Computer Interaction* 34 (2018) 577–590.
- [12] B. Laugwitz, T. Held, M. Schrepp, Construction and evaluation of a user experience questionnaire, in: *Symposium of the Austrian HCI and usability engineering group*, Springer, 2008, pp. 63–76.
- [13] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive NLP tasks, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6–12, 2020, virtual*, 2020. URL: <https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html>.
- [14] V. Bush, et al., As we may think, *The atlantic monthly* 176 (1945) 101–108.
- [15] A. Hogan, The semantic web: Two decades on, *Semantic Web* 11 (2020) 169–185. doi:10.3233/sw-190387.
- [16] F. J. Estrada-Martínez, J. R. Hilera, S. Otón, J. Aguado-Delgado, Semantic web technologies applied to software accessibility evaluation: a systematic literature review, *Universal Access in the Information Society* 21 (2020) 145–169. doi:10.1007/s10209-020-00759-y.
- [17] A. Hinderks, Design and evaluation of a short version of the user experience questionnaire (ueq-s), *International Journal of Interactive Multimedia and Artificial Intelligence* (2017). doi:10.9781/IJIMAI.2017.09.001.
- [18] World Wide Web Consortium (W3C), Web content accessibility guidelines (wcag) 2.2, 2023. URL: <https://www.w3.org/TR/WCAG22/>, accessed: 2026-05-05.
- [19] V. R. Basili, D. M. Weiss, A methodology for collecting valid software engineering data, *IEEE Trans. Software Eng.* 10 (1984) 728–738. doi:10.1109/TSE.1984.5010301.
- [20] V. R. Basili, G. Caldiera, D. H. Rombach, The goal question metric approach, *Encyclopedia of software engineering* (1994) 528–532.
- [21] B. Laugwitz, T. Held, M. Schrepp, Construction and evaluation of a user experience questionnaire, in: A. Holzinger (Ed.), *HCI and Usability for Education and Work*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008, pp. 63–76.
- [22] M. Guy, E. Tonkin, Folksonomies: Tidying up tags?, *D-Lib Magazine* 12 (2006). doi:10.1045/january2006-guy.
- [23] M. Hassenzahl, M. Burmester, F. Koller, *AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität*, Vieweg+Teubner Verlag, Wiesbaden, 2003, pp. 187–196. doi:10.1007/978-3-322-80058-9_19.
- [24] S. Angioni, A. A. Salatino, F. Osborne, A. Birukou, D. R. Recupero, E. Motta, Leveraging knowledge graph technologies to assess journals and conferences at springer nature, in: U. Sattler, A. Hogan, C. M. Keet, V. Presutti, J. P. A. Almeida, H. Takeda, P. Monnin, G. Pirrò, C. d’Amato (Eds.), *The Semantic Web - ISWC 2022 - 21st International Semantic Web Conference, Virtual Event, October 23–27, 2022, Proceedings, Lecture Notes in Computer Science*, Springer, 2022, pp. 735–752. doi:10.1007/978-3-031-19433-7_42.