From PDFs to Structured and FAIR Knowledge: Organizing NLP4RE ID Cards in the Open Research Knowledge Graph

A Pipeline and Dashboard Integration for Improved Access and Reuse of NLP4RE Knowledge

Amirreza Alasti[2,*], Lena John[1] and Oliver Karras[1]

[1]TIB - Leibniz Information Centre for Science and Technology, Germany
[2]Leibniz University Hannover, Germany

Abstract

[Background.] The natural language processing for requirements engineering (NLP4RE) community introduced NLP4RE ID Cards to describe empirical research practice from scientific publications and foster replicability. However, these artifacts are currently stored as static PDF documents, which limits accessibility, impedes machine actionability, and prevents efficient reuse, e.g., for joint analyses. **[Aims.]** We aim to transform these static documents into a findable, accessible, interoperable, and reusable (FAIR) representation format using a research knowledge graph (RKG). In particular, we organize the NLP4RE ID Cards in the Open Research Knowledge Graph (ORKG) to improve access to the knowledge and its reuse. **[Method.]** We develop a pipeline to extract, normalize, and structure data from NLP4RE ID Cards. This data is ingested into the ORKG using a tailored semantic graph schema. Furthermore, we integrate the data in the neuro-symbolic dashboard EmpiRE-Compass that combines symbolic SPARQL querying with neural large language models (LLMs) to facilitate knowledge exploration, synthesis, and reuse. **[Results.]** With the pipeline, we successfully migrated 50 NLP4RE ID Cards into a persistent, retrievable dataset in the ORKG. The dashboard integration allows researchers to answer complex competency questions — such as identifying common evaluation metrics or dataset properties across all NLP4RE ID Cards — dynamically. The extraction process also reveals significant limitations in the original PDF-based workflow regarding data consistency. **[Conclusions.]** Transitioning to an open science infrastructure, such as the ORKG, significantly enhances the utility of the scientific knowledge that was encapsulated in the PDF documents. To ensure long-term sustainability, we propose extending the static NLP4RE ID Cards in PDF format with direct, schema-driven web interface and LLM-assisted data entry via EmpiRE-Compass and the ORKG.

Keywords

Natural language processing, requirements engineering, research knowledge graph, neuro-symbolic dashboard

1. Introduction

The natural language processing for requirements engineering (NLP4RE) community introduced NLP4RE ID Cards to systematically report empirical research practice from scientific publications and foster replicability [1]. Currently, these artifacts are stored as static PDF documents. While this format ensures layout consistency and human readability, it encapsulates knowledge in isolated files, making it poorly suited for machine actionability and unmanageable for joint analysis.

In this paper, we present a pipeline to transform these static PDF documents into a machine-actionable representation format stored in the open science infrastructure, called *Open Research Knowledge Graph* (ORKG) [2]. By migrating from a document-centric to a data-centric approach, we enhance the findability, accessibility, interoperability, and reusability (FAIR) of NLP4RE knowledge. This transformation empowers the community to move beyond manual data aggregation, enabling dynamic, real-time queries about the state of the discipline. Ultimately, we demonstrate how open science infrastructures can turn static reporting artifacts into a structured and FAIR knowledge representation that actively supports future research and replication efforts. Below, we present the pipeline, the data integration into the neuro-symbolic dashboard *EmpiRE-Compass* [3, 4], and our plans for the tool demonstration.

Joint Proceedings of REFSQ-2026 Workshops, Doctoral Symposium, Posters & Tools Track, and Education and Training Track. Co-located with REFSQ 2026. Poznan, Poland, March 23-26, 2026

*Corresponding author.
amirreza.alasti@stud.uni-hannover.de (A. Alasti); lena.john@tib.eu (L. John); oliver.karras@tib.eu (O. Karras)
0009-0002-1165-773X (A. Alasti); 0009-0007-2097-9761 (L. John); 0000-0001-5336-6899 (O. Karras)

2. Background & Related Work

The increasing volume of scientific publications has made traditional, document-centric representations of scientific knowledge, e.g., as PDF documents, difficult to maintain, update, and reuse [2, 5]. To address these issues, several projects have explored using research knowledge graphs (RKGs) to represent scientific knowledge in a structured, machine-actionable form [6, 7, 8].

While generic RKGs such as *OpenAlex* or *Semantic Scholar Academic Graph* focus on bibliographic metadata (e.g., authors, venues), specific RKGs target scientific knowledge (e.g., methods, results) within certain topics (e.g., *SoftwareKG*) or domains (e.g., *Computer Science Knowledge Graph*) [9]. All mentioned RKGs are automatically populated and do not support direct knowledge contribution or curation by third parties. In contrast, the ORKG is a special case: It is a cross-domain, cross-topic, and community-driven RKG designed to organize scientific knowledge [9, 2]. The flexible architecture of the ORKG allows us to model the deep, specialized information of NLP4RE ID Cards while ensuring interoperability within a much larger, multidisciplinary infrastructure. In this way, the ORKG enables comparisons, searches, and joint analyses that move beyond simple metadata or narrow domain silos.

In requirements engineering (RE), there is growing evidence that the ORKG and, consequently, RKGs facilitate sustainable literature reviews [3, 5, 10, 9]. For example, EmpiRE-Compass utilizes KG-EmpiRE, a sub-graph in the ORKG comprising 776 publications, and maintains it to make the data openly accessible and retrievable [9]. However, an investigation of RKG-based applications reveals not only the potential of RKGs for organizing knowledge from scientific publications, but also the challenges associated with schema design, accuracy of information extraction, and quality assurance [6].

Although prior research has introduced structured reporting artifacts in NLP and RE [11, 12], as well as initiatives to improve transparency and replicability [13], these efforts have not yet addressed how replication-relevant information from NLP4RE studies can be systematically represented and reused at scale. Existing work focuses either on proposing reporting artifacts or on advancing FAIR open science infrastructures, but no approach has connected these strands to enable the organization of consistent and machine-actionable NLP4RE knowledge. This gap motivates our work on a structured and FAIR representation of NLP4RE knowledge for exploration, synthesis, and reuse.

3. Use Case and Pipeline

We developed a pipeline to transform static NLP4RE ID Cards into a FAIR, machine-readable knowledge representation that can be used in the EmpiRE-Compass dashboard. The design and development of this pipeline were primarily driven by the specific use case presented in Section 3.1, focusing on the large-scale ingestion of existing NLP4RE ID Cards [14] into the ORKG. The pipeline consists of two main phases: The digitization and ingestion of data into the ORKG, and its integration into the neuro-symbolic dashboard EmpiRE-Compass for knowledge exploration, synthesis, and reuse. The source code is available in our *repository*.

3.1. Use Case

Below, we outline the use case with its inputs and outputs. Primarily, we developed the pipeline to process the complete set of NLP4RE ID Cards [14] from the original NLP4RE ID Card publication [1]. In this primary use case, the authors of this paper executed the script to ingest the NLP4RE ID Cards into the ORKG. However, the pipeline is also designed for decentralized use by individual researchers. Any author who has created or is currently creating an NLP4RE ID Card can execute the pipeline to easily add their NLP4RE ID Card to the ORKG, making it also accessible within the EmpiRE-Compass dashboard. The data flow of the pipeline is defined by the following inputs and outputs:

- **Input**: One or a batch of multiple NLP4RE ID Cards in static PDF format.
- **Intermediate Output**: The first phase of the pipeline generates one or more JSON files (depending on the input) containing the cleaned, structured data mapped to the semantic graph schema.

- **Final Output**: The JSON files are processed in the first phase for ingestion, resulting in persistent semantic entries in the ORKG. In this way, the structured data is accessible to EmpiRE-Compass via the ORKG SPARQL endpoint for exploration, synthesis, and reuse in the second phase.

3.2. Pipeline - Phase 1: From Static PDFs to Structured and FAIR Knowledge

The first phase focuses on extracting data from the existing NLP4RE ID Cards in PDF format and populating the ORKG. This process begins with *Schema Development*, in which we manually developed a semantic *graph schema* in an iterative manner, following our approach that has already been successfully applied in requirements engineering [9], engineering sciences [15], and medicine [16]. This schema maps the unstructured fields of the NLP4RE ID Cards to typed, semantic entities and properties.

Afterward, the *Extraction and Normalization* stage employs the *PDF Form Extractor* to parse the NLP4RE ID Cards. The tool uses spatial analysis to map form field labels to values and handles various input types such as checkboxes and free text. A critical component is the *Data Cleaning Module*. As the NLP4RE ID Card allows free-text entries, we encountered inconsistencies such as mixed numeric/text values or duplicate entries. The module normalizes these inputs, resolves author/title ambiguities using DOIs, and merges duplicated NLP4RE ID Cards (e.g., from internal and external validation [1, 14]). This process resulted in a clean dataset of *50 NLP4RE ID Cards*. Finally, the *ORKG Ingestion* stage converts the cleaned data into JSON and automatically imports it via the ORKG REST API, transforming the static PDF documents into structured and FAIR knowledge as a public, persistent, and retrievable dataset.

3.3. Pipeline - Phase 2: Dashboard Integration for Improved Access and Reuse

We integrated the NLP4RE knowledge organized in the ORKG into the neuro-symbolic dashboard EmpiRE-Compass to improve its accessibility and reusability for researchers. The dashboard combines symbolic SPARQL querying with neural large language models (LLMs) to answer curated competency questions[1] and custom competency questions[2] (see Figure 1)[3].

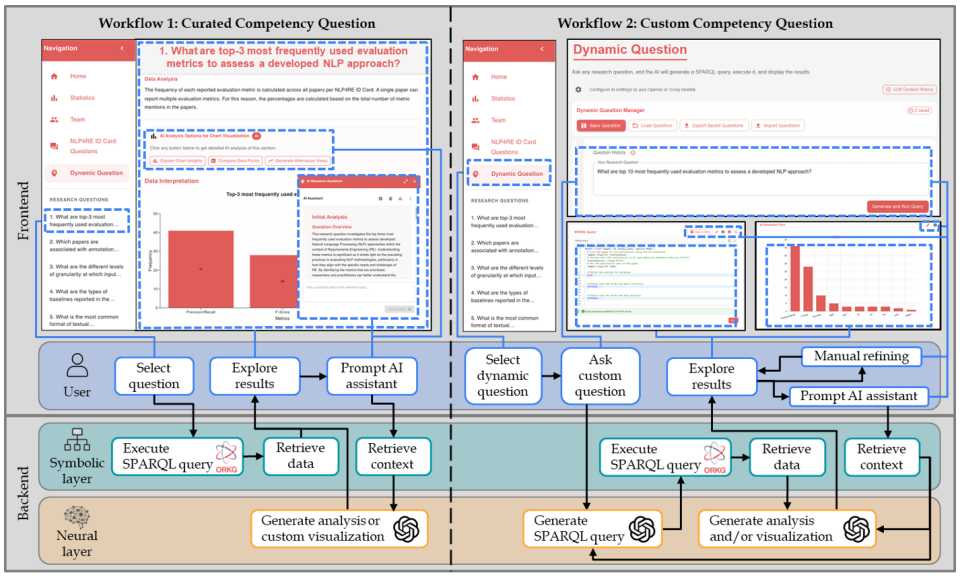

Figure 1: System design of EmpiRE-Compass with its two main workflows for knowledge exploration, synthesis, and reuse based on curated and custom competency questions[4].

Workflow 1 - Curated Competency Question: The user selects a curated competency question. The symbolic layer then executes a curated, pre-validated SPARQL query against the ORKG to retrieve precise data. This workflow presents predefined visualizations and answers to ensure immediate

[1]We collected the 10 curated competency questions from domain experts, namely the authors of the NLP4RE ID Card [1].
[2]A competency question is a natural language query expressing an information need that an RKG must be able to answer.
[3]The figure is also *online* available.
[4]The custom competency question used is provided online as example question "*Top-10 evaluation metrics for NLP approaches*".

correctness and reliability. The neural layer supports the further exploration, synthesis, and reuse of these results, allowing the user to interactively prompt the LLM for deeper insights, alternative summaries, or contextual explanations based on the retrieved context.

Workflow 2- Custom Competency Question: This workflow supports dynamic knowledge exploration, synthesis, and reuse. The user types a natural language question into the interface. The neural layer translates this question into a SPARQL query, which the symbolic layer executes. Based on the retrieved data, the neural layer generates a custom visualization with interpretations and explanations. The system includes a feedback loop for refining all results (see Figure 1). This refinement can be done either manually or through iterative natural language prompting, giving users complete control over the results. Similar to Workflow 1, the neural layer supports again the further exploration, synthesis, and reuse of the results based on the retrieved context.

The two layers are based on a clear separation of concerns. The symbolic layer interfaces directly with the ORKG, acting as the data source by executing SPARQL queries to ensure factual accuracy. Conversely, the neural layer bridges the gap between user intent and RKG logic, handling dynamic query generation and enabling the flexible interpretation and natural language synthesis of the structured data. Crucially, the system persists these interactions, allowing the complete analysis history, including queries, results, and reasoning, to be exported for transparent replication and sustainable reuse.

4. Demonstration Plan

During the NLP4RE workshop, we present an interactive walkthrough of the entire lifecycle of an NLP4RE ID Card, from a static PDF document to structured and FAIR knowledge in the ORKG, which can be accessed and reused in EmpiRE-Compass. We designed the demonstration plan around two key learning outcomes: (1) Participants will understand how to transform their own PDF-based NLP4RE ID Cards into semantic representations using our pipeline, and (2) Participants will learn how to conduct complex analyses across NLP4RE ID Cards using EmpiRE-Compass. The demonstration is structured into three parts, which are presented below.

I. The Pipeline. We present the entire pipeline, from loading an NLP4RE ID Card in PDF format to executing the *PDF Form Extractor* and *Data Cleaning Module* to generating the output (see Section 3.2).

1. *Selection of an NLP4RE ID Card:* Loading the PDF document into the pipeline.
2. *Extraction and Normalization:* Showing how the *PDF Form Extractor* and *Data Cleaning Module* automatically extract the values and resolve common inconsistencies.
3. *Output Generation:* Generating the JSON file that enforces the semantic graph schema.

II. The ORKG. Next, we present the ORKG. We show the result of the ingestion by navigating to the newly created paper (corresponding NLP4RE ID Card) within the ORKG. This step highlights the difference between the original unstructured NLP4RE ID Card as a PDF document and the structured and FAIR knowledge representation in the ORKG, emphasizing typed, semantic entities and properties.

III. The EmpiRE-Compass. Finally, we show the two main workflows of the neuro-symbolic dashboard EmpiRE-Compass where we integrated the NLP4RE knowledge for improved access and reuse.

1. *Workflow 1 - Curated Competency Question*: We select one or more curated competency questions to show real-time retrieval, exploration, synthesis, and reuse of NLP4RE knowledge.
2. *Workflow 2 - Custom Competency Question*: We invite the audience to propose ad-hoc questions to demonstrate the *Dynamic Question* feature to translate the natural language questions into SPARQL queries on the fly, demonstrating the neuro-symbolic capabilities (cf. Section 3.3).

5. Discussion

The transformation of the 50 NLP4RE ID Cards from static PDF documents into structured and FAIR knowledge in the ORKG creates an added value for the NLP4RE community and underscores the need

for machine-actionable data in addition to document-centric reporting. This shift highlights both the transformative potential of FAIR data and the technical rigor required to move beyond legacy formats.

Added Value. Transforming the 50 NLP4RE ID Cards into a structured and FAIR knowledge representation in the ORKG significantly lowers the barrier to accessing and reusing the NLP4RE knowledge. It enables the NLP4RE community to move from manually aggregating individual PDFs to easily analyzing the state of the discipline. Using the ORKG as one of the leading open science infrastructures fosters transparency and enables sustainable, up-to-date *overviews of NLP4RE research* [17], transforming the static reporting artifacts into a dynamic *community asset*. For this reason, we consider the semantically structured representation to be an additional layer alongside the PDF format, which enables the computational "heavy lifting" while preserving the human-readable document for archival purposes.

Lessons Learned. The development of the *PDF Form Extractor* revealed critical insights regarding the transition from a document-centric to a data-centric approach. The primary challenge lies in the limitations of PDF-based data entry. While the PDF format ensures layout consistency and human readability, it poses a significant hurdle for machine actionability. We observed that the NLP4RE ID Cards as PDF documents often "trapped" the data, requiring complex extraction logic to map spatial fields to typed, semantic entities and properties. Furthermore, the lack of strict data types and cardinalities in the original PDF form allowed respondents to utilize free-text fields with high flexibility. We found several inconsistent data entries, such as numbers represented as text (e.g., "one" versus "1") or mixing URLs with explanatory text. As a result, we needed a custom *Data Cleaning Module* to enforce data consistency before ingesting the data into the ORKG. These observations substantiate the need for rigorous development of a graph schema. To mitigate ambiguity in future iterations, we propose strict typing (e.g., using controlled vocabularies, defined data types with cardinalities, and persistent identifiers (DOIs)) to replace free-text fields, ensuring semantic data consistency.

Relevance of the Pipeline. Although we plan to develop a schema-driven web interface for data entry (see Section 6), this pipeline remains essential for the NLP4RE community. It offers a practical way to ensure that no existing or future PDF-based NLP4RE ID Card is "left behind" when the community transitions to more structured representation formats. In this way, it serves as a bridge for researchers who prefer the document-centric approach but still want to integrate their work into the ORKG.

6. Conclusion & Future Work

In this paper, we presented a pipeline and dashboard integration for organizing NLP4RE ID Cards in the ORKG, transforming them from static PDF documents to structured and FAIR knowledge for improved access and reuse of NLP4RE knowledge. Our demonstration plan illustrates the entire process from PDF extraction and data normalization to ORKG ingestion and finally to the retrieval, exploration, synthesis, and reuse of the knowledge via the neuro-symbolic dashboard EmpiRE-Compass.

For future work, our focus is on evolving the data ingestion process to be more sustainable and user-friendly. In our upcoming research project *SciD-QuESt*, we plan to develop a *Dynamic Data Entry Interface*. This approach will extend the current PDF-based option by allowing users to input data directly into a schema-driven web interface. By enforcing real-time validation and strict data types during data entry, this interface will mitigate the consistency issues identified, leading to higher data quality. In addition, we plan to explore utilizing LLMs and small language models (SLMs) for assisted data entry to analyze a publication and suggest values for the NLP4RE ID Card fields that a user can interactively accept, reject, or refine. This "Human-in-the-Loop" approach aims to reduce the manual effort while fostering rapid growth and adoption within the NLP4RE community.

It is important to mention that this approach does not make the PDF format obsolete. Instead, the data collected via the interface can be used to automatically generate human-readable PDFs or other formats. In this way, we ensure that the NLP4RE ID Card remains available as a portable artifact that can be co-located with the actual research tool. Furthermore, the semantically structured representation in the ORKG provides a persistent, linkable, and citable research object that can be assigned its own DOI, further incentivizing authors to contribute to the *NLP4RE community's knowledge graph*.

Declaration on Generative AI

The author(s) used Microsoft Copilot and Grammarly for grammar and spelling check. The authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] S. Abualhaija, et al., Replication in Requirements Engineering: The NLP for RE Case, ACM Transactions on Software Engineering and Methodology 33 (2024). doi:10.1145/3658669.

[2] S. Auer, et al., Open Research Knowledge Graph: A Large-Scale Neuro-Symbolic Knowledge Organization System, in: Handbook on Neurosymbolic AI and Knowledge Graphs, 2025. doi:10.3233/FAIA250216.

[3] O. Karras, et al., EmpiRE-Compass: A Neuro-Symbolic Dashboard for Sustainable and Dynamic Knowledge Exploration, Synthesis, and Reuse, in: Joint Proceedings of REFSQ-2026 Workshops, Doctoral Symposium, Posters & Tools Track, and Education and Training Track, 2026.

[4] O. Karras, et al., EmpiRE-Compass, 2026. doi:10.5281/zenodo.18170203.

[5] O. Karras, Research Knowledge Graphs for Sustainable Literature Reviews in Software Engineering and Beyond, in: Software Engineering 2025 – Proceedings, Gesellschaft für Informatik, 2025. doi:10.18420/se2025-ws-30.

[6] M. Zloch, et al., Research Knowledge Graphs: The Shifting Paradigm of Scholarly Information Representation, in: The Semantic Web. ESWC 2025., Springer, 2025. doi:10.1007/978-3-031-94578-6_8.

[7] K. Silva, et al., Research Knowledge Graphs in NFDI4DataScience: Key Activities, Achievements, and Future Directions, in: INFORMATIK 2025, Gesellschaft für Informatik, 2025. doi:10.18420/inf2025_102.

[8] M. Stocker, et al., SKG4EOSC - Scholarly Knowledge Graphs for EOSC: Establishing a Backbone of Knowledge Graphs for FAIR Scholarly Information in EOSC, Research Ideas and Outcomes 8 (2022). doi:10.3897/rio.8.e83789.

[9] O. Karras, et al., Divide and Conquer the EmpiRE: A Community-Maintainable Knowledge Graph of Empirical Research in Requirements Engineering, in: 17th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM), IEEE, 2023. doi:10.1109/ESEM56168.2023.10304795.

[10] O. Karras, et al., Researcher or Crowd Member? Why not both! The Open Research Knowledge Graph for Applying and Communicating CrowdRE Research, in: 29th International Requirements Engineering Conference Workshops, IEEE, 2021. doi:10.1109/REW53955.2021.00056.

[11] H. Cheng, et al., Generative AI for Requirements Engineering: A Systematic Literature Review, Software: Practice and Experience (2025). doi:10.1002/spe.70029.

[12] M. Mitchell, et al., Model Cards for Model Reporting, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, ACM, 2019. doi:10.1145/3287560.3287596.

[13] I. Magnusson, et al., Reproducibility in NLP: What Have We Learned from the Checklist?, in: Findings of the Association for Computational Linguistics: ACL 2023, 2023. doi:10.18653/v1/2023.findings-acl.809.

[14] S. Abualhaija, OnlineAnnex-Replication and Verifiability in Requirements Engineering: the NLP for RE Case (2024). doi:10.6084/m9.figshare.21824481.v1.

[15] O. Karras, et al., Organizing Scientific Knowledge from Engineering Sciences Using the Open Research Knowledge Graph: The Tailored Forming Process Chain Use Case, Data Science Journal (2024). doi:10.5334/dsj-2024-052.

[16] N. Brümmer, et al., Better Models, Better Treatment? A Systematic Review of Current Three Dimensional (3D) In Vitro Models for Implant-Associated Infections, Frontiers in Bioengineering and Biotechnology (2025). doi:10.3389/fbioe.2025.1569211.

[17] O. Karras, et al., Comparison of NLP4RE ID Cards, 2025. doi:10.48366/R1584378.