SYNAPSE-KG – Symbolic and Neural Academic & Organizational Platform for Semantic Exploration

Roxana Dogaru[1], Eleonora Laurenza[1] and Emanuel Sallinger[1]

[1]TU Wien, Vienna, Austria

Abstract

We present SYNAPSE-KG, an active, in-use industrial system for academic and organisational analytics. Being the result of a strategic project involving C-level leadership, researchers, engineers and innovation managers, we showcase in detail the industrial setting and the system. At the core of it is a rule-based declarative AI system providing transparency for decision making. The system currently integrates 44 heterogeneous data sources and demonstrates how the interplay between symbolic and subsymbolic methods enables advanced, scalable, and domain-independent analytical capabilities for academic decision-making support.

1. Introduction

In this paper, we report on Synapse-KG, a large-scale neurosymbolic system deployed in one of the most complex and knowledge-intensive industries: Academia. Combining rule-based reasoning and machine learning, the system operates at a strategic institutional level at TU Wien, involving approximately 15 FTEs, including C-level leadership, researchers, engineers, and innovation managers. This distinctive industry faces the challenge of simultaneously pursuing educational, societal, and economic objectives under conditions of increasing uncertainty; contributing to innovation and long-term societal development, while guiding themselves towards optimal functioning [1, 2, 3]. While graph-based approaches to scholarly platforms have been created at many institutions, such as very successfully at EPFL and ETH Zurich, what makes the approach presented here particularly relevant is the paramount importance of rules and rule-based reasoning, especially for complex reasoning patterns.

Although universities are among the primary producers of advanced knowledge and technological innovation, they invest comparatively little in developing formal system-oriented analysis and decision-support infrastructures for their own internal governance and optimization of organizational processes and decision-making capabilities [4, 5]. Academic institutions must also balance financial sustainability and resource efficiency with long-term research goals, societal impact and academic freedom, which can conflict each other. This motivates the need for novel approaches to decision support in academia, capable of leveraging the richness of available data of these *Knowledge Organizations* also for themselves. In this, especially rule-based systems are effective methodologies to operationalize this.

Goal and core challenges. In 2025, TU Wien formalized a strategic vision under the motto "One Mission – One Vision" [6], representing an organizational response to institutional complexity, with the goal of fostering the integration of data, processes, and interoperable digital infrastructures. Such an ambitious vision needs a strong empirical support and the ability to perform advanced multidimensional analysis over the university's activities, structures, and outcomes. Specifically, it calls for decision support systems capable of integrating heterogeneous data and transforming them into actionable insights. This need motivated the development of Synapse-KG, a neurosymbolic system designed to integrate heterogeneous scholarly and organizational data and transform them into actionable institutional intelligence, in particular using rule-based systems at the core.

Solution. To address this need, we built on the extensive expertise in Knowledge Representation and Reasoning (KRR) [7, 8] developed within the RuleML community and at TU Wien, and engineered

roxana.dogaru@tuwien.ac.at (R. Dogaru); eleonora.laurenza@tuwien.ac.at (E. Laurenza); sallinger@dbai.tuwien.ac.at (E. Sallinger)
0000-0002-2786-8163 (E. Laurenza); 0000-0001-7441-129X (E. Sallinger)
© 2026 This work is licensed under a "CC BY 4.0" license.

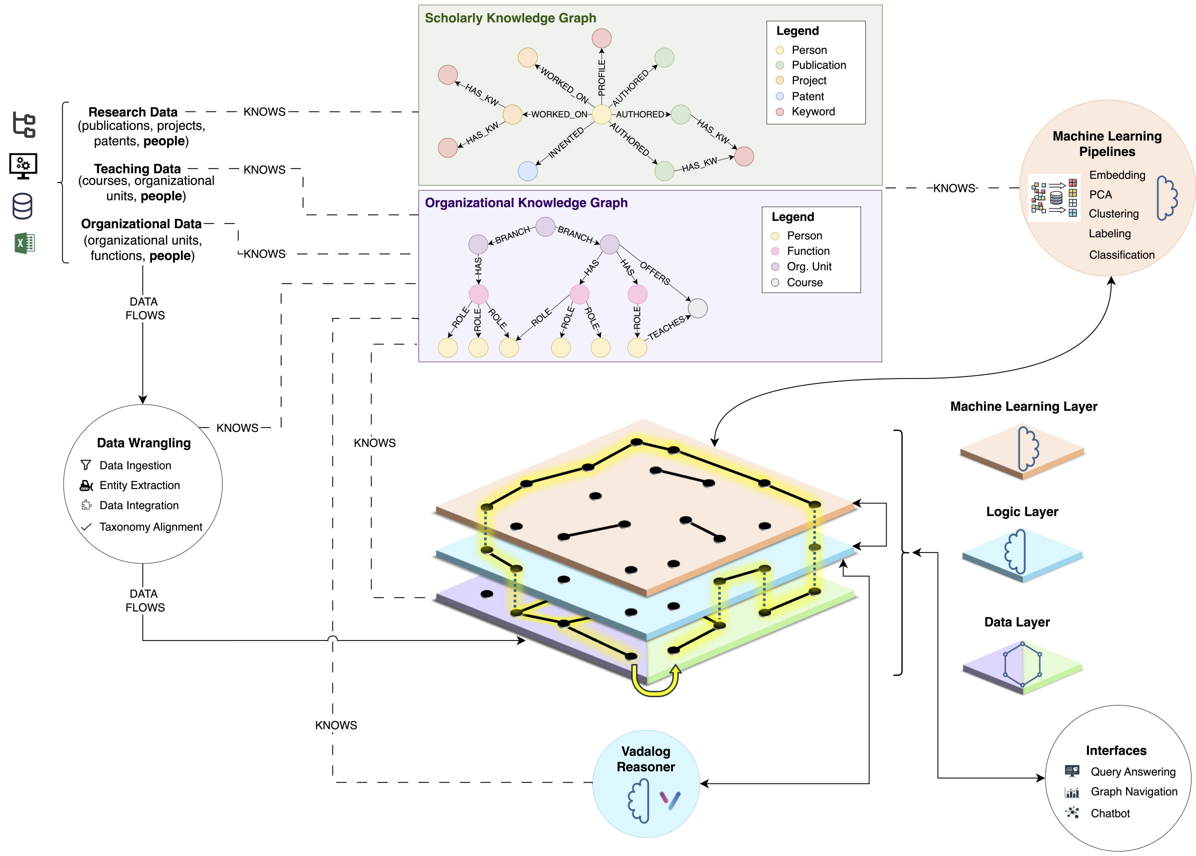

Figure 1: Overview of architecture of the Synapse-KG.

an industrial system that leverages this tradition to support strategic decision-making through three interacting layers within the same rule-based knowledge graph (KG).

We present Synapse-KG (SYmbolic and Neural Academic organizational Platform for Semantic Exploration), a neurosymbolic system designed to support semantic exploration and decision-making in academic institutions. Synapse-KG integrates heterogeneous scholarly and organizational data into a unified KG, combining symbolic reasoning with inductive analysis to capture both explicit and latent knowledge. This enables flexible, multidimensional analysis of complex academic environments, supporting tasks such as researcher profiling, topic analysis, and the discovery of latent collaborations.

We structure the Synapse-KG system in three layers, generalizing the three interacting components of a KG: a data layer, capturing the extensional information that integrates information from multiple data silos; a logical layer, enabling expressive reasoning and inference over integrated data; and a ML layer, supporting the extraction of latent patterns using ML models, as shown in Figure 1.

The **main contributions** of this work can be summarized as follows:

- We present Synapse-KG, an **industrial and large-scale** neurosymbolic system designed and deployed at TU Wien, currently involving approximately 15 FTEs, including research managers, innovation managers, software engineers, and academic researchers, reflecting the interdisciplinary nature of the proposed approach;
- enabling a coherent representation of **people**, their research activities, institutional structures, and their links, achieved by a unified and interconnected scholarly and organizational KG;
- at its core based on **extensive, expressive logical rules** that operate both on the scholarly and organizational dimensions of the graph, supporting advanced reasoning, affiliation analysis, profiling, and collaboration discovery;

- proposing a **neurosymbolic integration** approach that combines symbolic reasoning with inductive models to exploit latent semantic information and domain knowledge at scale, enabling multidimensional, context-aware analysis of heterogeneous academic data.

2. Approach

The use case we consider is the construction of a complete research profile for a researcher. Creating a complete profile must also account for the broader relational context of organizational and scholarly data, including organizational structures, collaboration and influence networks, topic expertise, and latent research connections. These relationships are not explicitly available in the underlying data sources and must be derived through different kinds of reasoning and analysis.

Complete Research Profile (CRP) of a Researcher. To integrate academic data sources, each of which provides only a partial view and differs in accessibility, structure, quality, and semantics, Synapse-KG uses a large-scale knowledge graph combining scholarly and organizational information, with an extensive data wrangling phase in the data layer. The resulting KG constitutes the system's extensional data layer, where publications, projects, affiliations, organizational entities, and personal profiles are represented as interconnected facts. As illustrated in Figure 1, these facts populate the Scholarly and Organizational KG schemas (green and purple, respectively), providing a unified semantic representation of the institution's research ecosystem. Many of the relationships required to construct a CRP are not explicitly available in the underlying data sources, such as data with multi-level and interleaved hierarchies, collaboration and influence relationships, etc. Determining such information requires exploring large portions of the networks and deriving new relationships from existing ones, raising both reasoning and scalability challenges. To capture these relationships as well as many others, Synapse-KG provides a symbolic, rule-based reasoning layer. This layer derives higher-level relationships, such as indirect affiliations, collaboration structures, and influence networks, making them explicitly available as new facts in the KG. These inferred relations subsequently enrich the data layer and become available for further reasoning and machine learning processes. The symbolic layer of Synapse-KG is implemented using Vadalog [9], a declarative reasoning language based on the Datalog$^\pm$ familyand designed for scalable reasoning over KGs, depicted in the blue circle in Figure 1. Vadalog provides a highly expressive language for navigating and reasoning over KGs, including features of practical utility such as management of temporal [10] and probabilistic rules [11]. In Synapse-KG rules are used both for data wrangling and for analytical reasoning. Let us start our representative running example with simple rules to that effect.

Rules 1-2 (Organization affiliation propagation).

$$has\_indirect\_affiliation(P, O) :- \ hasFunction(P, O). \tag{1}$$

$$has\_indirect\_affiliation(P, O_2) :- \ has\_indirect\_affiliation(P, O_1),\ belongs\_to(O_1, O_2). \tag{2}$$

In Rules (1-2) we reconstruct the direct and indirect affiliation of a person P inside the organization, where O, O_1 and O_2 in the recursive rule denote organizational entities at arbitrary levels of the institutional hierarchy. Similarly, information about the collaborative research relationships of a researcher is typically not explicitly accessible in the original extensional representation, but must be derived through a reasoning process. Relationships such as co-authorship, co-investigation, and latent collaboration emerge from the interactions of multiple heterogeneous entities and therefore need to be materialized as new relations within the KG schema, becoming part of the KG itself and continuously feeding downstream reasoning tasks and machine learning pipelines, in the pink circle in Figure 1. In this way, symbolic reasoning not only consumes the data layer but also actively contributes to its expansion and refinement.

Rule 3-4 (Collaborative relationships).

$$coauth(A, B), coauth(B, A) :- \ auth\_internal(A, P), auth(B, P), A \neq B. \tag{3}$$

$$coinvestig(A, B), coinvestig(B, A) :- \ investig(A, P), investig(B, P), A \neq B. \tag{4}$$

Rules (3-4) capture collaboration-based influence relationships among researchers through shared scholarly outputs. The predicates *auth* and *auth_internal* denote publication authorship relations, with the latter identifying institutionally internal authors. Two authors A and B are in a *coauth* relationship if they co-author the same publication, are distinct, and at least one of them is internal to the institution. Through the reasoning process, the system derives new knowledge that is continuously reinjected into the KG schema, enriching both the data layer and the downstream ML pipelines that in turn, are used for further logical steps as illustrated in Figure 1 where the solid arrows represent operational data flows through the different layers, while the dashed *KNOWS* relationships capture ontological propagation and contextual awareness among the components of the framework.

2.1. The Rule+ML Squeezebox

Once influence relationships and direct and indirect affiliations have been identified, the next step in constructing a CRP is the discovery of **latent collaboration opportunities**. By latent collaborations, we refer to potential collaborations that are not directly observable in the available scholarly data. More specifically, these are pairs of researchers who have not yet interacted through co-authored publications or shared research projects, but whose research activities suggest a high likelihood of future collaboration. In a statistical sense, such opportunities correspond to *latent information*: patterns that are not explicitly represented in the data but can be inferred from underlying regularities and hidden semantic structures.

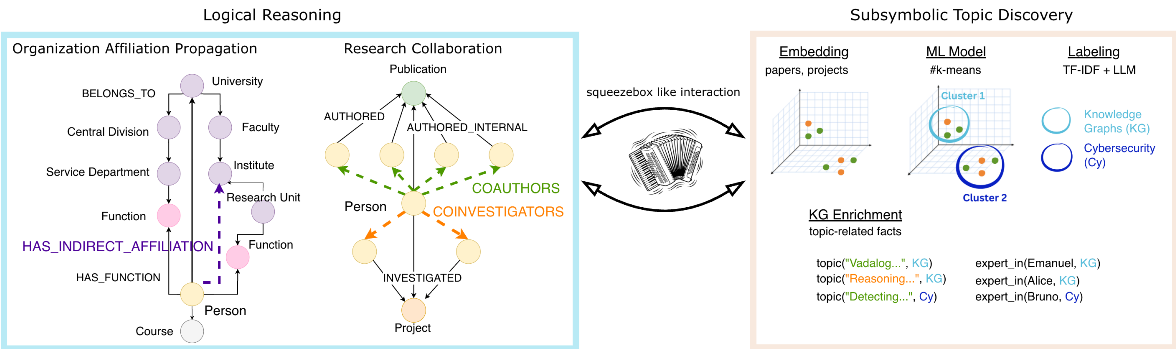

Figure 2: The Rule-based Reasoning and Machine Learning Squeezebox

Researchers working on closely related topics may belong to a sort of latent **research cluster**, not explicitly represented in the KG. Such information is embedded within research outputs (e.g., publications, projects, and patents) and cannot be extracted through symbolic rules alone. Discovering these latent structures requires subsymbolic analysis capable of identifying regularities within large collections of researchers. To this end, the machine learning layer employs clustering, labeling and language model techniques to uncover hidden topic structures, which subsequently become part of the enriched KG representation. ML alone, however, is insufficient to identify meaningful collaboration opportunities, as semantic similarity does not necessarily imply actionable (or institutionally relevant) collaborations. For this reason, we combine inductive models with logical reasoning through a **squeezebox-like** approach that alternates **restrictive** and **expansive** reasoning phases. Restrictive rules narrow the search space by selecting collaboration candidates who align with research topics, expertise, and institutional context. Expansive rules subsequently expand this space, recovering potentially relevant collaboration opportunities that would otherwise be discarded by a purely machine-learning-based analysis. Restrictive and expansive rules may be purely logical or incorporate extensional predicates, whose facts represent the output of corresponding ML models (denoted by predicates prefixed with #). In the case at hand, we will show clustering and textual similarity models, while our system provides a library of other models suitable for multiple applications. The interaction between symbolic and subsymbolic components is illustrated in Rules 5–10.

Rule 5-10 (Squeezebox-like latent collaboration discovery: restrictive and expansive rules).

$$
\begin{aligned}
\mathit{candidates}(P_1, P_2), \mathit{candidates}(P_2, P_1) &:- \mathit{coauth}(P_1, P_2). && (5)\\
\mathit{coauthors}(P_1, P_2) &:- \mathit{coinvestig}(P_1, P_2). && (6)\\
\mathit{collab}(P_1, P_2) &:- \mathit{expert\_in}(P_1, T_1), \mathit{expert\_in}(P_2, T_2), \\
&\quad\ \mathit{research\_output}(P_1, R_1), \mathit{research\_output}(P_2, R_2), \\
&\quad\ \mathit{topic}(R_1, T_1), \mathit{topic}(R_2, T_2), \\
&\quad\ \#k{-}\mathit{means}(R_1, R_2), \mathit{candidates}(P_1, P_2), && (7)\\
\mathit{candidates}(P_1, P_3) &:- \mathit{collab}(P_1, P_2), \\
&\quad\ \mathit{affiliated\_with}(P_1, O),\ \mathit{affiliated\_with}(P_2, O), \\
&\quad\ \mathit{affiliated\_with}(P_3, O). && (8)\\
\mathit{candidates}(P_1, P_3) &:- \mathit{collab}(P_1, P_2), \mathit{professorship}(P_2, S_2), \\
&\quad\ \mathit{professor}(P_3), \mathit{professorship}(P_3, S_3), \\
&\quad\ \#\mathit{LLMSame}(S_2, S_3). && (9)\\
\mathit{latent\_collab}(P_1, P_2) &:- \mathit{collab}(P_1, P_2), \neg\, \mathit{coauth}(P_1, P_2), \\
&\quad\ \neg\, \mathit{coinvestig}(P_1, P_2). && (10)
\end{aligned}
$$

Rules (5) and (6) first construct an explicit collaboration network from scholarly activities. Existing co-authorship and co-investigation relationships are transformed into symmetric *candidates* relations, creating an initial exploration space for latent collaboration discovery.

Rule (7) introduces the **restrictive** reasoning phase that also leverages the inductive layer of Synapse-KG; the predicate $\mathit{collab}(P_1, P_2)$ identifies potential collaboration relationships between researchers associated with semantically related research outputs; the predicates *expert_in*, *research_output*, and *topic* connect researchers to latent research topics, while $\#k{-}\mathit{means}(R_1, R_2)$ captures proximity between research outputs induced through clustering. In particular, rather than exhaustively comparing all pairs of researchers, we leverage logical rules to first identify promising candidates based on shared expertise. Then the predicate $\mathit{candidates}(P_1, P_2)$ further constrains the search space to relevant researchers. Thus, this first phase acts as a first restrictive filtering mechanism.

Rules (8) and (9) implement **expansive** reasoning phases, progressively enlarging the candidate space through additional contextual signals. Rule (8) propagates candidate relationships toward additional researchers P_3 sharing the same organizational unit O. In this way, the framework reintroduces potentially relevant collaboration opportunities that may not emerge solely from semantic similarity in the inductive layer. Rule (9) further expands the exploration space through similarity derived from LLM-based representations of professorship descriptions, captured by the predicate $\#\mathit{LLMSame}(S_2, S_3)$. Here, the focus is not on the specific similarity model adopted. While in this example we leverage an on-premise open-weights LLM (selectable through TU Wien's Aqueduct LLM gateway)), which is also crucial given the usual GDPR requirements, the system is modular and accommodates multiple models.

Finally, Rule (10) $\mathit{latent\_collab}(P_1, P_2)$ derives the facts by filtering out researchers who have already collaborated using a further **restrictive** phase of reasoning. Our Vadalog program terminates by construction, as it explores, in the worst case, all the pairs of candidates. In the average case, there is a trade-off between width of clusters—the smaller the faster—and accuracy—the larger the more accurate. In the limit case, clusters coinciding with the entire search space yield maximum accuracy. Recall depends on the underlying clustering model, which might miss potential candidates.

The resulting latent collaboration network captures unexplored collaboration opportunities from the interaction of symbolic and subsymbolic layers of the system. Thus, not only does Synapse-KG support descriptive profiling but also forward-looking insights on scholarly and organizational data.

2.2. Domain-independent Exploration Strategies and Analysis

The interaction between the introduced layers enables flexible and **domain-independent exploration strategies** for advanced analytical tasks. The goal is to discover relationships that are not explicitly represented in the data but emerge from the interaction of multiple heterogeneous signals. Some of these hidden relationships can be identified through ML models. In our setting, we cluster semantically related research outputs according to a similarity measure and a distance threshold, thereby uncovering latent conceptual structures. However, each ML model captures only a specific notion of similarity, the one encoded in its representation and objective function. Logical reasoning complements this process by introducing additional forms of domain knowledge encoded as rules. Such knowledge involves organizational structures, collaboration patterns, temporal constraints, or other deterministic relationships that are difficult to capture through only inductive models. The results produced by ML are injected into the logical reasoning process, which derives new candidate relationships. These new candidates can be re-evaluated by the ML layer, giving rise to a progressive exploration process.

This **squeezebox-like** interaction alternates between restrictive and expansive phases: the former where restrictive rules focus the search on coherent candidate spaces, the latter enlarges them by introducing additional plausible candidates. The process allows the framework to explore by construction a broader, richer set of hypotheses than would not be possible with any individual part in isolation. Furthermore in this way, the interaction between symbolic reasoning and inductive analysis continuously enriches the data layer, enabling more refined and context-aware institutional insights.

2.3. Validation

The development of Synapse-KG followed an iterative evaluation process with continuous stakeholder engagement. During the project, more than 30 stakeholder meetings were conducted with representatives from C-level leadership, innovation management, research strategy, IT services, and researchers in multiple faculties. This iterative validation process informed the evolution of the framework, helping align the system with institutional needs, operational constraints, and strategic decision-making objectives. The current deployment of Synapse-KG covers the period 2018–2025 and integrates information from 44 heterogeneous internal and external data sources, including 32 organizational systems and 12 scholarly repositories. The resulting KG captures the research and organizational university ecosystem within its 8 faculties, comprising approximately 30k publications, 10k research projects, and 300 patents. Following data wrangling and integration, the KG contains more than 187k nodes and 1.1 million relationships, representing researchers, organizational entities, scholarly outputs, and their interconnections. The system adopted continuous integration and continuous delivery in a cloud-based environment provided by TU Wien and has been evaluated according to a comprehensive, quality-oriented framework that emphasized continuous assessment of design choices, security, and maintainability.

Declaration on Generative AI. During the preparation of this work, the authors used ChatGPT and Grammarly in order to: Grammar and spelling check, Citation management, Paraphrase and reword. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

Acknowledgements. The authors' work was funded by the Vienna Science and Technology Fund (WWTF) [Grant ID: 10.47379/VRG18013, 10.47379/ICT25032, 10.47379/NXT22018, 10.47379/ICT2201, 10.47379/DCDH001], and by the Austrian Science Fund (FWF) 10.55776/COE12.

References

[1] J. E. S. Carmo, D. P. Lacerda, C. O. Klingenberg, F. A. S. Piran, Digital transformation in the management of higher education institutions, Sustainable Futures 9 (2025) 100692.

[2] C. D. Jaldi, E. Ilkou, N. Schroeder, C. Shimizu, Education in the era of neurosymbolic ai, Journal of Web Semantics 85 (2025) 100857.

[3] M. K. Z. Elkaimbillah, Z. Mcharfi, B. E. Asri, Nesy-guidance: A neuro-symbolic kg for academic recommendations combining rule-based reasoning and neural inference, Int. Journal of Modern Education and Computer Science 17 (2025) 48–64.

[4] C. L. Borgman, A. Brand, Data blind: Universities lag in capturing and exploiting data, Science 378 (2022) 1278–1281.
[5] S. Gaftandzhieva, S. Hussain, S. Hilcenko, R. Doneva, K. Boykova, Data-driven decision making in higher education institutions: State-of-play, Int. Journal of Advanced Computer Science and Applications 14 (2023).
[6] TU Wien, Executive summary: The strategy of TU Wien, 2025.
[7] S. Kirrane, M. Simkus, A. Soylu, D. Roman (Eds.), Rules and Reasoning - 8th International Joint Conference, RuleML+RR 2024, 2024.
[8] A. Hogan, K. Satoh, H. Dag, A. Turhan, D. Roman, A. Soylu (Eds.), Rules and Reasoning - 9th International Joint Conference, RuleML+RR 2025, 2026.
[9] L. Bellomarini, E. Sallinger, G. Gottlob, The vadalog system: Datalog-based reasoning for knowledge graphs, Proc. VLDB Endow. 11 (2018) 975–987.
[10] L. Bellomarini, L. Blasi, M. Nissl, E. Sallinger, The temporal vadalog system: Temporal datalog-based reasoning, Theory Pract. Log. Program. 25 (2025) 168–196.
[11] L. Bellomarini, E. Laurenza, E. Sallinger, E. Sherkhonov, Swift markov logic for probabilistic reasoning on knowledge graphs, Theory Pract. Log. Program. 23 (2023) 507–534.