Exploring Wikidata as a Global-Scale Knowledge Graph for Human Resource Management

Fariz Darari[1], Jaycent G. Ongris[1], Berty C. L. Tobing[2], Douglas R. Faisal[2] and On Lee[2]

[1]Faculty of Computer Science, Universitas Indonesia, Depok 16424, Indonesia
[2]GDP Labs, Jakarta 12950, Indonesia

Abstract

Human resource (HR) management is critical to organizational effectiveness but faces persistent challenges in representing, integrating, and analyzing workforce-related knowledge. This paper explores the potential of Wikidata, a large-scale, multilingual, collaboratively maintained knowledge graph (KG), as a global infrastructure for HR knowledge management. We examine how Wikidata's flexible data model supports key HR dimensions, including employment history, skills, education, projects, achievements, and publications. Its support for external identifiers enables seamless integration with platforms such as GitHub, LinkedIn, and the European Skills, Competences, Qualifications, and Occupations (ESCO) framework. We demonstrate how HR-related insights can be extracted via SPARQL queries and visualized using built-in tools. Furthermore, advanced techniques such as GraphRAG, graph-based exploratory data analysis (EDA), and KG embeddings enable innovative ways of consuming HR knowledge. Our findings highlight Wikidata's value as a foundation for intelligent HR knowledge management, with promising applications in semantic search and organizational analytics.

1. Introduction

Human resource (HR) management plays a vital role in ensuring the operational effectiveness of businesses, government agencies, and other organizations. As outlined in [1], HR refers to the people employed by an organization and is often considered its most valuable asset. The impact and success of an organization largely depend on the skills and competencies of its workforce. The field of HR is inherently multifaceted, spanning legal, social, economic, and technological dimensions, creating significant challenges for capturing and using HR knowledge effectively.

A knowledge graph (KG) captures real-world knowledge through a graph structure in which nodes represent entities of interest and edges represent the relationships between them [2]. Due to their ability to represent knowledge in a structured and interconnected manner, KGs are well suited to support HR functions. Prior work has built HR-related KGs from text to enable applications such as employee and job recommendations as well as job-sector classification with Graph Neural Networks (GNNs) [3]. Related efforts enhance talent acquisition by aggregating sources such as LinkedIn, job boards, and internal HR systems to construct KGs that support comprehensive candidate profiles and the identification of top talent [4].

Wikidata'25: Wikidata workshop at ISWC 2025
fariz@ui.ac.id (F. Darari); jaycent.gunawan@ui.ac.id (J. G. Ongris); berty.c.l.tobing@gdplabs.id (B. C. L. Tobing); douglas.r.faisal@gdplabs.id (D. R. Faisal); onlee@gdplabs.id (O. Lee)
© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

Wikidata is a large-scale, open, multilingual, and general-purpose KG, launched by the Wikimedia Foundation in October 2012 [5]. As of this writing, Wikidata contains more than 118 million entities and over 2.3 billion edits[1]. With its global scope, collaborative nature, and semantic richness, Wikidata offers significant potential as a foundation for HR knowledge management, particularly for companies and organizations worldwide. By leveraging Wikidata's entity relationships and multilingual content, it is possible to support unified skill mapping, organizational benchmarking, and other HR-related processes.

In light of the potential outlined above, this paper investigates key aspects of leveraging Wikidata for HR knowledge management:

- **Representation and integration**: We examine how Wikidata's flexible data model, together with its use of external identifiers, enables effective representation and integration of diverse HR-related information.
- **Consumption through querying, visualization, and graph analytics**: We explore techniques for accessing and visualizing HR knowledge through the Wikidata SPARQL query service, as well as advanced methods such as Graph Retrieval-Augmented Generation (GraphRAG) and KG embeddings.

Our study employs an exploratory research design [6]. This is appropriate given the study's aim to develop an initial understanding of Wikidata's potential for HR-related knowledge services and to surface both opportunities and challenges.

Competency Questions

Following the notion of competency questions (CQs) [7], we outline a set of questions that a Wikidata-based HR knowledge base should be able to answer. These CQs serve as a litmus test for scope and level of detail, and as a basis for determining the scope of the Wikidata schema we consider suitable for HR knowledge management, rather than creating a new schema from scratch.

- **CQ1**: Which projects, notable work, and publications is a person associated with, and in what roles (e.g., participant, author, owner)?
- **CQ2**: Which external identifiers can enrich an employee's internal profile?
- **CQ3**: Can we find employees or candidates who satisfy expert-search constraints for a role or project?
- **CQ4**: For workforce planning, can we quantify and compare, by field, the coverage of associated occupations, surface gaps relative to strategic demand (e.g., initiatives or roadmaps), and use the results to prioritize hiring or upskilling?

2. Representation and Integration

To illustrate Wikidata's capacity to represent HR knowledge, we first examine how key concepts such as employment, education, and skills are modeled in its data model. We then show how

[1] https://www.wikidata.org/wiki/Wikidata:Statistics

Wikidata links to external platforms, serving as a hub for integrating HR-related information from diverse sources, and discuss aligning internal HR KGs with the Wikidata schema.

2.1. HR Data Model in Wikidata

Wikidata provides a rich and extensible ontology for modeling a wide range of HR-related concepts, enabling structured representation of individuals' professional profiles, educational backgrounds, competencies, and affiliations. Its property-based model, combined with qualifiers and references, allows for nuanced descriptions of temporal, contextual, and relational aspects of HR data. This subsection supports **CQ1** by showing how roles, education, projects, and publications can be represented for person-centric profiling. The following key entities and properties illustrate how core HR dimensions are captured in Wikidata:

- **Person (or Human)** (Q5) is the central node representing an individual. Attributes include sex or gender (P21), languages spoken (P1412), and citizenship (P27).
- **Company and Organization**. Companies can be modeled as enterprise (Q6881511) or business (Q4830453) nodes, while organizations use organization (Q43229) or its subclasses. Relevant properties include inception (P571), work location (P937), chairperson (P488), organizational divisions (P199), industry or sector (P452), and number of employees (P1128).
- **Employment, Field, and Work Experience**. The node profession (Q28640) captures the general notion of a profession, which can be linked to its sector via field of this occupation (P425). The property occupation (P106) denotes a person's profession, while employer (P108) and position held (P39) capture organizational roles. Start time (P580) and end time (P582) are qualifiers representing temporal boundaries. The property member of (P463) links individuals to professional organizations.
- **Education and Training**. Educated at (P69) represents formal education. Academic degree (P512) and field of work (P101) provide additional context.
- **Skills, Competencies, and Achievements**. The property has certification (P10611) links a person to a certification they have obtained. The node professional certification (Q16023913) captures certificates across fields. Awards received (P166) documents achievements, notable work (P800) links to significant contributions, and significant event (P793) records key life or career events.
- **Projects, Ownership, and Publications**. The node project (Q170584) represents collaborative work toward a specific goal. The property participant in (P1344) links a person to projects, and the property participant (P710) connects a project to its participants. Ownership (e.g., of startups) is represented via owner of (P1830) and owned by (P127). Scholarly article (Q13442814) represents academic publications, with author (P50) linking to contributors.

2.2. Wikidata as a Hub

Beyond internal modeling, Wikidata functions as a central hub by linking to external authoritative platforms through standardized identifiers. This interoperability enables the integration of

HR-related data across systems, enhancing completeness and cross-platform connectivity. This subsection addresses **CQ2** by showing how external identifiers (e.g., GitHub, Google Scholar, LinkedIn) can enrich internal employee profiles, downstream HR workflows, and interlink information across platforms. Examples include:

- **GitHub username** (P2037): Identifies the GitHub account associated with a person or organization.
- **Google Scholar author ID** (P1960): Links a person's Wikidata item to a persistent Google Scholar profile, enabling citation tracking and academic impact analysis.
- **LinkedIn personal profile ID** (P6634): Connects an individual's Wikidata entry to their LinkedIn[2] profile, mapping structured Wikidata data to professional networking records.
- **ESCO skill ID** (P4644): Associates a competency or skill with its ESCO[3] (European Skills, Competences, Qualifications and Occupations) identifier, enabling multilingual and cross-industry skill alignment.
- **ESCO Occupation ID** (P4652): Maps an occupation concept in Wikidata to its ESCO classification, promoting consistency in job taxonomies and semantic interoperability.
- **OpenCorporates ID** (P1320): Provides a unique identifier for companies listed in the OpenCorporates[4] database, allowing Wikidata to reference verified global company profiles for HR analytics.
- **Indeed company ID** (P10285): Connects a company's Wikidata entry to its listing on Indeed[5], linking organizational data to job postings and employer profiles.
- **Crunchbase person ID** (P2087): Links an individual to their Crunchbase[6] profile, including career history, investments, and organization affiliations.

2.3. Aligning Internal HR KG with Wikidata Schema

Prior work on HR knowledge graphs spans ontology-first modeling and data-driven construction. For instance, Zhang et al. [8] propose a top-down ontology for HR concepts, emphasizing OWL-based structure and conceptual clarity. While such efforts help define scope and relationships, many are light on end-to-end pipelines that connect the ontology to operational data sources and downstream analytics within organizations. In contrast, recent practical approaches (e.g., [3]) leverage large language models (LLMs) to extract entities and relations from HR documents (CVs, job descriptions), and then apply graph methods for tasks such as job matching. Our contribution complements both lines by advocating an internal-first HR KG aligned with Wikidata's schema (classes and properties) for semantic interoperability, while preserving provenance and access control over corporate data.

In our approach, the internal HR KG is aligned with the Wikidata schema. This alignment broadens applicability across HR scenarios beyond any single downstream task: person-centric profiling can unify roles, education, projects, and publications in a queryable form; identity

[2] https://www.linkedin.com/
[3] https://esco.ec.europa.eu/en
[4] https://opencorporates.com/
[5] https://id.indeed.com/
[6] https://www.crunchbase.com/

resolution and cross-platform linking become more systematic; and workforce planning cover occupations across fields. Importantly, Wikidata is not an appropriate repository for private employee data due to privacy, compliance, and governance constraints. Alignment is therefore necessary to obtain semantic interoperability without exposing sensitive information. Accordingly, the internal KG retains authoritative corporate facts, while Wikidata provides the semantic scaffolding and linking layer.

3. Consumption through Querying, Visualization, and Graph Analytics

This section presents practical use cases for accessing HR data in Wikidata through querying, visualization, and graph analytics.

3.1. SPARQL Queries for HR Knowledge Extraction

SPARQL, the query language for RDF, enables targeted retrieval of structured HR data. This subsection addresses **CQ3** by showing how a Wikidata-aligned schema supports internal-first expert search and consolidated profiling via SPARQL. For example, Figure 1 shows the results of a query that extracts professional and educational information about Andrej Karpathy.

propertyLabel	valueLabel
educated at	Stanford University
field of work	machine learning
field of work	computer vision
occupation	computer scientist
occupation	artificial intelligence researcher
employer	Tesla, Inc.
employer	OpenAI
Google Scholar author ID	l8WuQJgAAAAJ
GitHub username	karpathy

Figure 1: Andrej Karpathy's work-related information.

The SPARQL query that generated the results in Figure 1 is shown below:

SPARQL Query

```
SELECT ?propertyLabel ?valueLabel WHERE {
  VALUES ?directProperty {
    wdt:P108  # employer
    wdt:P106  # occupation
    wdt:P69   # educated at
    wdt:P101  # field of work
    wdt:P2037 # GitHub username
    wdt:P1960 # Google Scholar author ID
  }
  wd:Q56037405 ?directProperty ?value . # Andrej Karpathy
  ?property wikibase:directClaim ?directProperty .
  SERVICE wikibase:label { bd:serviceParam wikibase:language "en" } }
```

An expert search seeks a person with specific expertise, going beyond simple keyword matching to leverage semantic relationships. In Wikidata, for example, one can search for a computer scientist who speaks Japanese and has received an award from an AI conference series. The following SPARQL query expresses this need and returns Hiroaki Kitano (Q3915986).

SPARQL Query

```
SELECT ?person ?personLabel WHERE {
  ?person wdt:P106 wd:Q82594 ; # occupation, computer scientist
          wdt:P1412 wd:Q5287 ; # languages spoken, Japanese
          wdt:P166 ?award .    # award received
  ?award wdt:P859 ?conferenceSeries . # sponsor
  ?conferenceSeries wdt:P921 wd:Q11660 . # main subject, artificial intelligence
  SERVICE wikibase:label { bd:serviceParam wikibase:language "en" } }
```

Next, Figure 2 examines how the occupation of AI engineer relates to the concept of artificial general intelligence (AGI) in Wikidata. As shown, the occupation of AI engineer falls within the field of artificial intelligence (AI), which in turn has, among its goals, AGI.

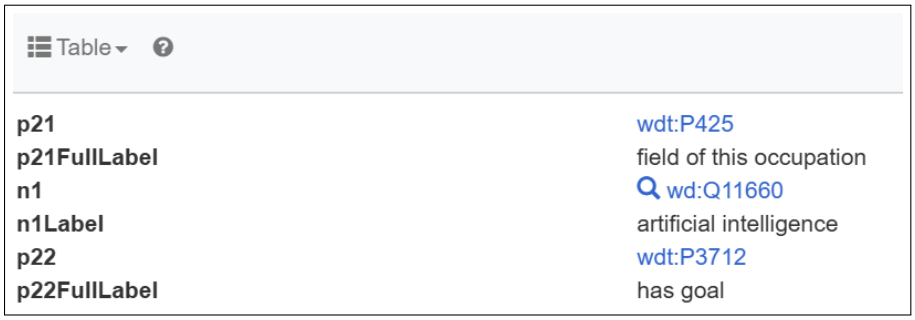

Figure 2: Relation of AI engineer with AGI.

The result shown above comes from the SPARQL query below, which illustrates a multi-hop traversal of HR knowledge. It is designed to retrieve results for 1-hop, 2-hop, or 3-hop connections, depending on which exist in the graph. In this case, only the 2-hop pattern yields results.

SPARQL Query

```
SELECT ?p11 ?p11FullLabel ?p21 ?p21FullLabel ?n1 ?n1Label ?p22 ?p22FullLabel
?p31 ?p31FullLabel ?n2 ?n2Label ?p32 ?p32FullLabel ?n3 ?n3Label
?p33 ?p33FullLabel WHERE {
  # Q126116209 = AI engineer, Q2264109 = AGI
  { wd:Q126116209 ?p11 wd:Q2264109 . # 1-hop query
    ?p11Full wikibase:directClaim ?p11 . }
  UNION
  { wd:Q126116209 ?p21 ?n1 . # 2-hop query
    ?n1 ?p22 wd:Q2264109 .
    ?p21Full wikibase:directClaim ?p21 .
    ?p22Full wikibase:directClaim ?p22 . }
  UNION
  { wd:Q126116209 ?p31 ?n2 . # 3-hop query
    ?n2 ?p32 ?n3 .
    ?n3 ?p33 wd:Q2264109 .
    ?p31Full wikibase:directClaim ?p31 .
    ?p32Full wikibase:directClaim ?p32 .
    ?p33Full wikibase:directClaim ?p33 . }
  SERVICE wikibase:label { bd:serviceParam wikibase:language "en" } }
```

3.2. Visualization Techniques

Beyond textual query results, visual representations provide an intuitive way to explore and communicate structured HR data. Wikidata's SPARQL endpoint supports various built-in visualization modes, such as trees, timelines, and graphs, that can reveal temporal patterns and relational structure.

3.2.1. Field and Occupation Visualization

This subsection addresses **CQ4** by enabling workforce-planning analyses that compare the breadth of occupations across fields. Figure 3 presents a visualization of selected fields and their occupations in Wikidata, showing occupations linked to the fields of esports, marketing, and tennis. The SPARQL query used to produce the tree visualization is shown below.

SPARQL Query

```
#defaultView:Tree
SELECT ?field ?fieldLabel ?occupation ?occupationLabel
(CONCAT("(",STR(?description),")") AS ?occupationDescription) WHERE {
  VALUES ?field { wd:Q300920 wd:Q39809 wd:Q847} # esports, marketing, tennis
  ?occupation wdt:P31 wd:Q28640 . # instance of, profession
  ?occupation wdt:P425 ?field . # field of this occupation
  ?occupation schema:description ?description .
  FILTER(LANG(?description) = "en")
  SERVICE wikibase:label { bd:serviceParam wikibase:language "en" }
} ORDER BY ?fieldLabel
```

Figure 3: Tree visualization of the occupations for esports, marketing, and tennis.

3.2.2. Employment Timeline Visualization

Employment timelines display an individual's work history over time. Figure 4 shows the employment trajectory of Andrej Karpathy from 2015 to 2022, as recorded in Wikidata.

Figure 4: Employment timeline of Andrej Karpathy from 2015 to 2022.

The SPARQL query used to generate the employment timeline in Figure 4 is as follows.

SPARQL Query

```
#defaultView:Timeline
SELECT ?employer ?employerLabel ?start ?end WHERE {
  wd:Q56037405 p:P108 ?statement . # Andrej Karpathy, employer
  ?statement ps:P108 ?employer ;
             pq:P580 ?start ;
             pq:P582 ?end .
  SERVICE wikibase:label { bd:serviceParam wikibase:language "en" } }
```

3.2.3. Coworker Graph Visualization

Graph-based visualizations can reveal networks of individuals connected to the same organization. Figure 5 presents a coworker graph of people affiliated with Tesla, Inc., as represented in Wikidata.

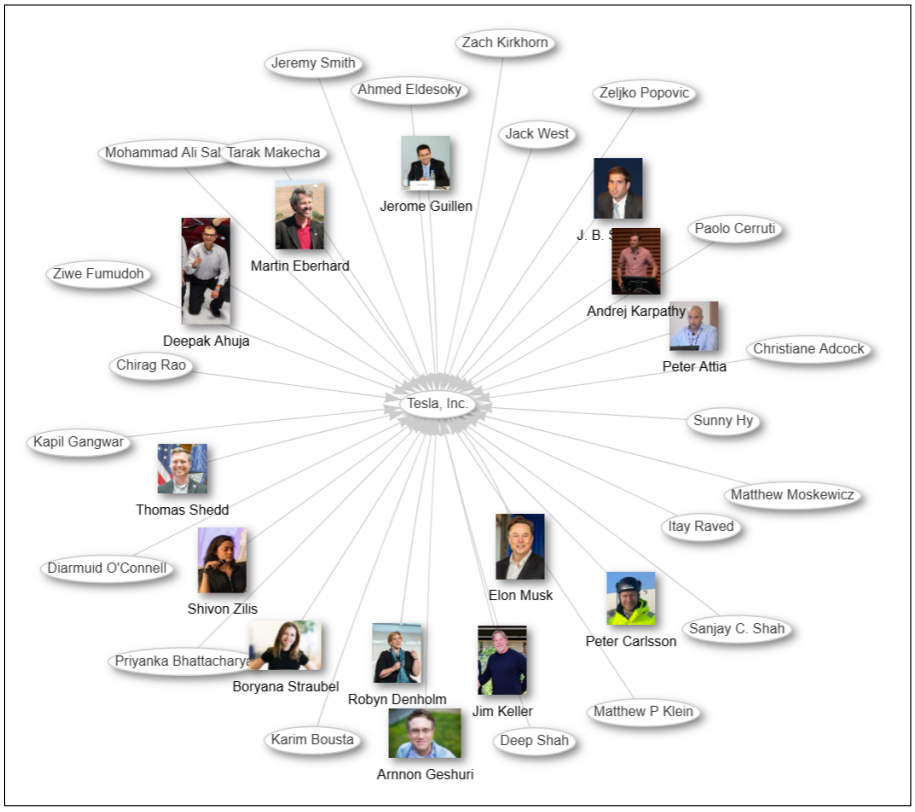

Figure 5: Coworker graph visualization of Tesla, Inc.

The SPARQL query used to create the coworker graph in Figure 5 is shown below.

SPARQL Query

```
#defaultView:Graph
SELECT ?item ?itemLabel ?pic ?linkTo ?linkToLabel WHERE {
  BIND(wd:Q478214 AS ?linkTo) # Tesla
  ?item wdt:P108 ?linkTo . # employer
  OPTIONAL { ?item wdt:P18 ?pic }
  SERVICE wikibase:label {bd:serviceParam wikibase:language "en" } }
```

3.3. Graph Analytics: GraphRAG, Graph EDA, and KG Embeddings

Graph Retrieval-Augmented Generation (GraphRAG) combines structured graph data (e.g., Wikidata) with large language models (LLMs) to generate accurate, context-aware text. In the HR domain, GraphRAG enables dynamic summaries of a person's employment history, skill set, certifications, and project involvement. Using LangChain[7], a Python framework for LLM applications, developers can build pipelines that retrieve relevant graph knowledge via SPARQL and feed it to models (e.g., ChatGPT, DeepSeek, Llama, Qwen) for natural language generation. These tools integrate with Wikidata's SPARQL interface (e.g., via `SPARQLWrapper`[8]), enabling responses such as "Jane Doe, a machine learning engineer educated at MIT, worked at Google from 2015 to 2020 on AGI-related initiatives."

We developed a proof-of-concept GraphRAG implementation that integrates several components into a unified workflow. The system performs entity extraction using an LLM, links entities through Wikidata's `wbsearchentities` API[9], stores relevant information in a Chroma-based[10] vector database to support property retrieval, and applies few-shot text-to-SPARQL prompting with an LLM. This text-to-SPARQL interface is particularly useful in organizational settings, where many users are non-experts in KGs or SPARQL, because it lets lay users query HR-relevant information from Wikidata using natural language and benefit from curated KG content without specialized training. Given a natural language HR-related question as input, the system generates a corresponding SPARQL query for execution on Wikidata. For example, the question "What is John von Neumann's field of work?" produces the SPARQL query shown below, where John von Neumann is Q17455 and the field of work property is P101:

SPARQL Query

```
SELECT ?fieldOfWorkLabel WHERE {
  wd:Q17455 wdt:P101 ?fieldOfWork .
  SERVICE wikibase:label { bd:serviceParam wikibase:language "en" }
}
```

When executed, the query returns John von Neumann's fields of work, ranging from mathematics to informatics.

In parallel, exploratory data analysis (EDA) on graph data can be conducted using libraries

[7] https://www.langchain.com/
[8] https://github.com/RDFLib/sparqlwrapper
[9] https://www.wikidata.org/w/api.php
[10] https://www.trychroma.com/

such as `NetworkX`[11]. For example, one analysis sought the Meta (Q380) employee associated with the highest number of distinct fields of work. The result indicates that Tomáš Mikolov (Q24698708) has the most (six fields).

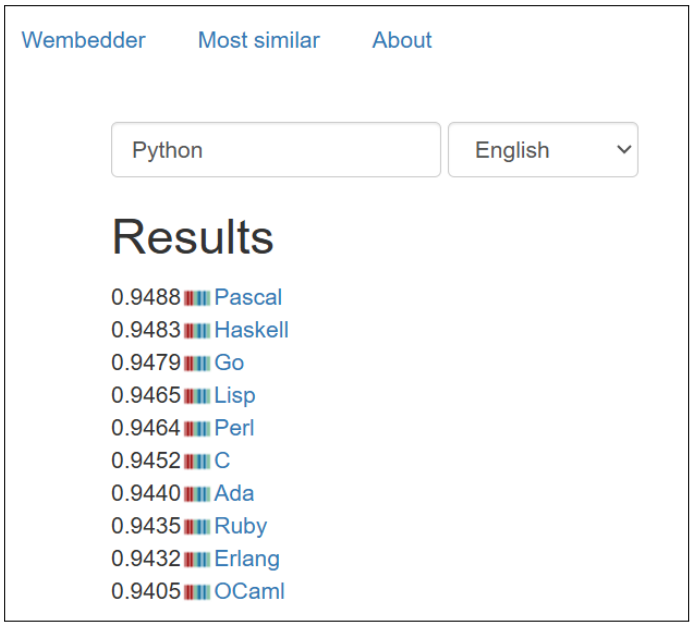

Figure 6: KG embedding-based similarity for the entity Python (programming language).

KG embeddings represent nodes and relations in a continuous vector space, capturing both semantic and structural patterns. Libraries such as `RDF2Vec` [9] generate embeddings by performing random walks over the graph and learning vector representations from the resulting sequences. For instance, Wembedder [10] demonstrates how KG embeddings capture latent semantic relationships: software engineer is close to game programmer; Microsoft is similar to Apple and Amazon; and, as shown in Figure 6, Python clusters with other programming languages such as Pascal, Go, and C.

4. Conclusions

This paper has demonstrated that Wikidata provides a valuable foundation for modeling and managing human resource (HR) information at global scale. Its flexible data model can represent core HR dimensions, including employment history, education, skills, certifications, and affiliations, while rich use of external identifiers enables integration with authoritative sources such as GitHub, Google Scholar, LinkedIn, and ESCO. We have shown that SPARQL queries can effectively extract, analyze, and visualize HR data (e.g., timelines, tree views, coworker graphs), and that GraphRAG, exploratory graph analysis, and KG embeddings open additional ways of consuming HR knowledge. A key practical point is that private employee data should not

[11] https://networkx.org/

reside in Wikidata. Instead, aligning an internal HR KG to the Wikidata schema provides semantic interoperability and cross-platform linking while keeping sensitive data within corporate systems. Taken together, these results suggest that Wikidata can serve as a shared semantic scaffold for scalable, transparent, and intelligent HR knowledge management, provided that alignment patterns, provenance, and access control are respected.

Future work will focus on extending this approach to real-world HR systems by adapting HR knowledge from Wikidata and combining it with private corporate data stored in platforms such as cloud storage, CRM, and ERP systems. Additional metadata and alignment techniques can be explored to better integrate heterogeneous data sources, particularly when company data exists only in unstructured forms (e.g., text or documents). Because HR interacts closely with finance, legal, and IT, integrating knowledge across departments toward a unified enterprise KG presents both a challenging and rewarding opportunity. We also plan to investigate additional use cases, including global workforce planning and data-driven organizational decision-making.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT and Gemini for grammar & spelling checking and paraphrasing. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] Whatcom Community College, What is Human Resources?, Online, n.d. URL: https://textbooks.whatcom.edu/bus230/chapter/1-1-what-is-human-resources/, accessed: 2025-08-06.

[2] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. de Melo, C. Gutiérrez, S. Kirrane, J. E. Labra Gayo, R. Navigli, S. Neumaier, A.-C. Ngonga Ngomo, A. Polleres, S. M. Rashid, A. Rula, L. Schmelzeisen, J. F. Sequeda, S. Staab, A. Zimmermann, Knowledge Graphs, Synthesis Lectures on Data, Semantics, and Knowledge, Springer, 2021. doi:10.2200/S01125ED1V01Y202109DSK022.

[3] A. T. Wasi, HRGraph: Leveraging LLMs for HR data knowledge graphs with information propagation-based job recommendation, in: R. Biswas, L.-A. Kaffee, O. Agarwal, P. Minervini, S. Singh, G. de Melo (Eds.), Proceedings of the 1st Workshop on Knowledge Graphs and Large Language Models (KaLLM 2024), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 56–62. URL: https://aclanthology.org/2024.kallm-1.6/. doi:10.18653/v1/2024.kallm-1.6.

[4] Zenia Graph, HR Accelerator – Talent Acquisition, Web page, n.d. URL: https://zeniagraph.ai/products/talent-acquisition/, accessed: 2025-08-06.

[5] D. Vrandečić, M. Krötzsch, Wikidata: a free collaborative knowledgebase, Commun. ACM 57 (2014) 78–85. doi:10.1145/2629489.

[6] J. W. Tukey, We need both exploratory and confirmatory, The American Statistician 34 (1980) 23–25. URL: https://www.jstor.org/stable/2682991.

[7] N. Noy, D. McGuinness, et al., Ontology development 101: A guide to creating your first ontology, 2001. URL: http://www.ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness-abstract.html.

[8] S. Zhang, X. Wang, W. Lu, Y. Lu, B. Deng, Construction of human resource ontology model for knowledge graph, in: 2021 IEEE 4th International Conference on Big Data and Artificial Intelligence (BDAI), 2021, pp. 150–153. doi:10.1109/BDAI52447.2021.9515238.

[9] P. Ristoski, Exploiting semantic web knowledge graphs in data mining, in: Studies on the Semantic Web, 2018. URL: https://api.semanticscholar.org/CorpusID:28273154.

[10] F. Årup Nielsen, Wembedder: Wikidata entity embedding web service, 2017. URL: https://arxiv.org/abs/1710.04099. arXiv:1710.04099.