Building a Canonical Register of Public Sector Entities: Semantic Linking of Procurement Data at Scale

Roberto Avogadro[1], Ian Makgill[2], Aleena Thomas[1], Ahmet Soylu[3] and Dumitru Roman[1]

[1] SINTEF AS, Oslo, Norway
[2] Spend Network, London, United Kingdom
[3] Kristiania University of Applied Sciences, Oslo, Norway

Abstract

Public procurement generates over $13 trillion annually, yet data about public buyers and suppliers remains fragmented, inconsistent, and difficult to link across jurisdictions. This paper presents a practical industrial solution developed by Spend Network within the European project enRichMyData to semantically enrich and reconcile procurement data at scale. The proposed pipeline combines large language models (LLMs) with knowledge graphs (KGs) to create and maintain a canonical register of public sector entities. It supports multilingual, cross-border integration and is designed to serve both public transparency and commercial applications. The pipeline has been evaluated on a manually curated benchmark of 1,000 procurement-related entities and demonstrates high precision and scalability in real-world settings.

Keywords

Entity Linking, Knowledge Graphs, Large Language Models, Procurement Data, Semantic Technologies

1. Introduction

Government procurement is a key area of public spending and accountability, with over $13 trillion annually spent worldwide. Despite the introduction of standards like the Open Contracting Data Standard (OCDS) [1, 2, 3], data about government buyers and suppliers remains difficult to reconcile due to inconsistencies in naming, multilingual variations, and missing canonical references. This hampers transparency, compliance checks, and cross-border cooperation. Knowledge graphs such as Wikidata [4] provide a foundation for such reference alignment.

In the past, entity matching efforts using standard fuzzy matching techniques (e.g., Levenshtein Distance) often resulted in poor performance, with either high false negatives or false positives, making large-scale reconciliation economically unviable.

To address this challenge, within the enRichMyData project[1], we developed for Spend Network[2] (the largest known collection of OCDS procurement records) a semantic linking pipeline that supports the creation of a canonical, structured, and continuously updated register of public sector entities. This register supports multiple use cases ranging from compliance (e.g., Environmental, Social, and Governance (ESG) reporting or procurement law) and cross-border collaboration to sales intelligence and civil society oversight.

2. Semantic Linking Strategy

The pipeline follows a hybrid architecture combining knowledge graphs with large language models. This approach builds on advances in transformer-based models such as BERT [5]:

ISWC 2025 Companion Volume, November 2–6, 2025, Nara, Japan

roberto.avogadro@sintef.no (R. Avogadro); ian@spendnetwork.com (I. Makgill); aleena.thomas@sintef.no (A. Thomas); ahmet.soylu@kristiania.no (A. Soylu); dumitru.roman@sintef.no (D. Roman)
0009-0005-7404-4123 (R. Avogadro); 0000-0001-6034-4137 (A. Soylu); 0000-0001-6397-3705 (D. Roman)

[1] https://enrichmydata.eu
[2] https://www.spendnetwork.com

1. **Candidate Generation:** Entities are extracted from procurement datasets and matched against canonical references using hybrid search combining vector similarity and approximate string matching.
2. **Ranking and Validation:** LLMs rank and validate candidate entities in context. Entities are approved automatically when confidence is high or reviewed manually in ambiguous cases. This process is informed by prior work on zero-shot and neural-based entity linking [6, 7, 8].
3. **Reconciliation and Enrichment:** Once linked, entities are enriched with structured data from public registers (e.g., URLs, legal identifiers, sectors) and linked to a central reference.
4. **Access and Integration:** The data is made available via APIs and downloadable formats to support dashboards, compliance systems, and bulk integration with private and public tools.

As shown in Figure 1, the full pipeline processes OCDS procurement data through enrichment, linking, and delivery to end-user dashboards. A more detailed view of the semantic linking layer using LLMs and validation workflows is shown in Figure 2.

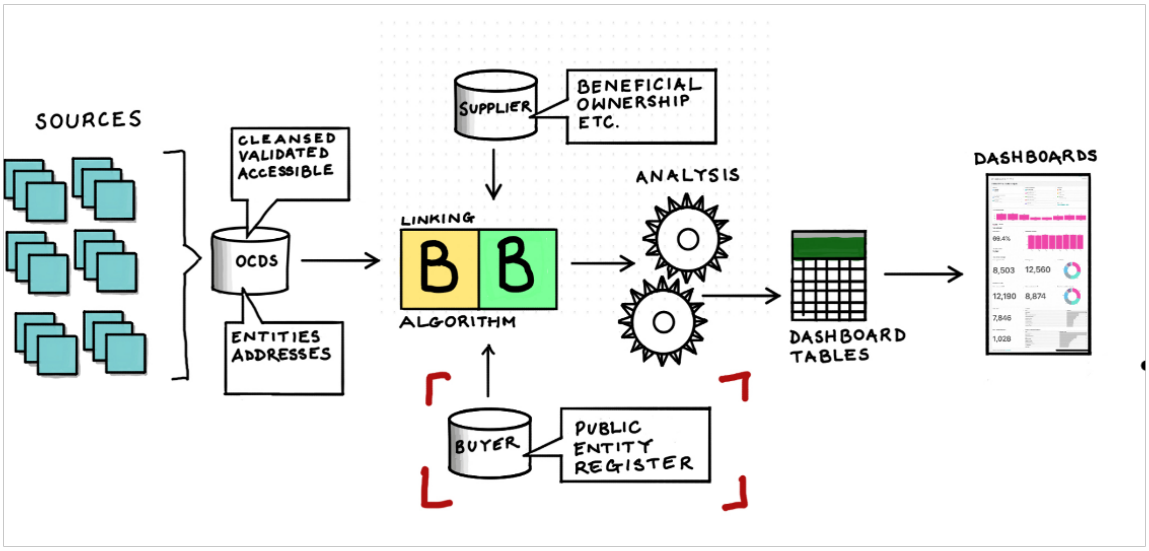

Figure 1: Overview of the procurement data linking pipeline, from OCDS data to dashboards.

3. Deployment Context

Spend Network maintains the largest known collection of OCDS procurement records, with over 180 million entries aggregated from EU and international sources [9]. Within the enRichMyData project, the aim is to build a public register of entities that are legally required to publish procurement records under the EU Procurement Directive. The envisioned service is designed to support both public interest (e.g., transparency, compliance) and commercial use (e.g., data integrations via API or bulk access). It builds on Spend Network's existing infrastructure and leverages the enRichMyData toolbox to support entity discovery, classification, cleansing, and reconciliation.

4. Ground Truth and Evaluation

To evaluate the linking pipeline, we created a ground truth dataset of 1,000 procurement-related entries. Each row is annotated with a Wikidata entity ID or marked as `NIL` where no appropriate match exists. Approximately 22.9% of the dataset contains such NIL cases, reflecting realistic ambiguity and out-of-knowledge scenarios. The experiments were conducted using the Lion Linker[3], an open-source entity linking python library developed within the enRichMyData project.

[3] https://github.com/enRichMyData/lion_linker

We evaluated multiple LLMs across different prompt configurations using precision@1 as our metric. Since each model outputs exactly one prediction per row and we always provide ground truth, this measure effectively captures system accuracy without confounding effects from recall. It aligns with standard evaluation practices in entity linking and retrieval benchmarks [10, 11]. The best-performing model (Gemma3:12b, few-shot prompt) achieved a precision@1 of 77.7%. The ground truth dataset is publicly available on Zenodo.[4]

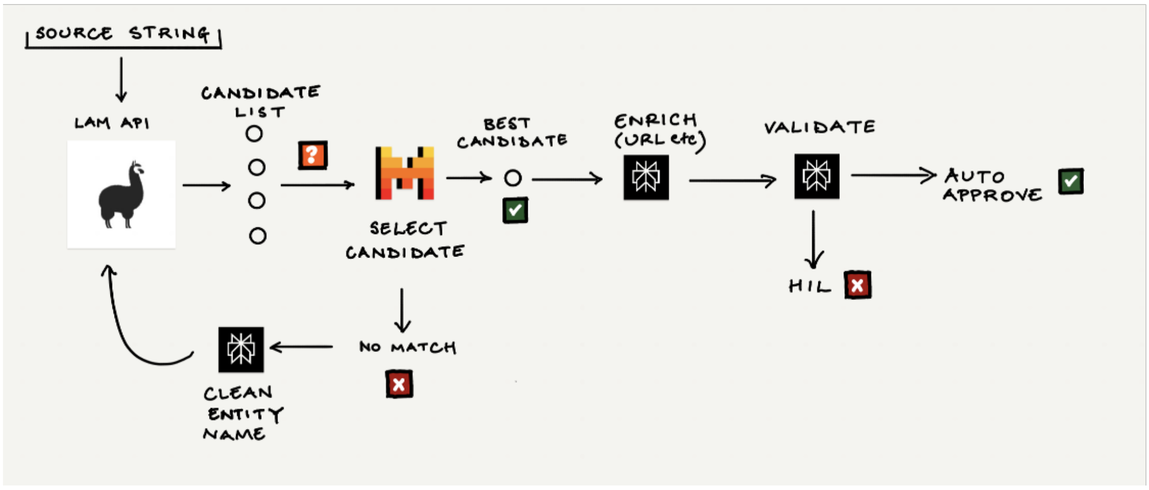

Figure 2: Semantic linking and enrichment process using language models and validation layers.

5. Business Value and Use Cases

The solution addresses concrete needs:

- **Public sector:** Enables compliance checks, inter-agency collaboration, and accurate public registers [9, 12, 13].
- **Private sector:** Supports due diligence, Environmental, Social, and Governance (ESG) reporting, and Customer Relationship Management (CRM) system integration.
- **Civil society:** Empowers journalists, NGOs, and citizens with better transparency tools.

The pipeline powers OpenOpportunities[5] and has been adopted in multiple use cases including a national compliance system in a non-EU European country.

6. Lessons Learned and Future Work

Lessons: Hybrid pipelines improve linking accuracy. Confidence scoring reduces human validation needs. Reconciliation across multilingual and evolving registers remains a challenge.

Next steps: Scale coverage across the EU, increase multilingual robustness, and expand entity types to cover beneficial owners and subnational bodies.

Acknowledgments

This work was supported by the European Union's Horizon 2020 research and innovation programme under grant agreements No. 101070284 (enRichMyData) and No. 101093216 (UPCAST). We thank Spend Network for their contribution.

[4] https://zenodo.org/records/15745734

[5] https://www.openopps.com

Declaration on Generative AI

This paper used ChatGPT (OpenAI) for drafting assistance, grammar checking, and paraphrasing. No AI was used for generating results or conclusions; the authors take full responsibility for the content.

References

[1] A. Soylu, B. Elvesæter, P. Turk, D. Roman, O. Corcho, E. Simperl, G. Konstantinidis, T. C. Lech, Towards an ontology for public procurement based on the open contracting data standard, in: Digital Transformation for a Sustainable Society in the 21st Century: 18th IFIP WG 6.11 Conference on e-Business, e-Services, and e-Society, I3E 2019, Trondheim, Norway, September 18–20, 2019, Proceedings 18, Springer, 2019, pp. 230–237.

[2] M. E. K. Niessen, J. M. Paciello, J. I. P. Fernandez, Anomaly detection in public procurements using the open contracting data standard, in: 2020 Seventh International Conference on eDemocracy & eGovernment (ICEDEG), IEEE, 2020, pp. 127–134.

[3] H. Felizzola, C. Gomez, N. Arrieta, V. Jerez, Y. Erazo, G. Camacho, Enhancing transparency in public procurement: A data-driven analytics approach, Information Systems 125 (2024) 102430.

[4] D. Vrandečić, M. Krötzsch, Wikidata: a free collaborative knowledgebase, Communications of the ACM 57 (2014) 78–85.

[5] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), 2019, pp. 4171–4186.

[6] L. Logeswaran, M.-W. Chang, K. Lee, K. Toutanova, J. Devlin, H. Lee, Zero-shot entity linking by reading entity descriptions, arXiv preprint arXiv:1906.07348 (2019).

[7] O.-E. Ganea, T. Hofmann, Deep joint entity disambiguation with local neural attention, arXiv preprint arXiv:1704.04920 (2017).

[8] I. Jayawardene, R. Avogadro, A. Soylu, D. Roman, Tablinkllm: An llm-based approach for entity linking in tabular data, in: 2024 IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), IEEE, 2024, pp. 206–214.

[9] A. Soylu, O. Corcho, B. Elvesæter, C. Badenes-Olmedo, T. Blount, F. Yedro Martínez, M. Kovacic, M. Posinkovic, I. Makgill, C. Taggart, et al., Theybuyforyou platform and knowledge graph: Expanding horizons in public procurement with open linked data, Semantic Web 13 (2022) 265–291.

[10] R. Usbeck, M. Röder, A.-C. Ngonga Ngomo, C. Baron, A. Both, M. Brümmer, D. Ceccarelli, M. Cornolti, D. Cherix, B. Eickmann, et al., Gerbil: general entity annotator benchmarking framework, in: Proceedings of the 24th international conference on World Wide Web, 2015, pp. 1133–1143.

[11] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, I. Gurevych, Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models, arXiv preprint arXiv:2104.08663 (2021).

[12] A. Soylu, O. Corcho, B. Elvesæter, C. Badenes-Olmedo, F. Y. Martínez, M. Kovacic, M. Posinkovic, I. Makgill, C. Taggart, E. Simperl, et al., Enhancing public procurement in the european union through constructing and exploiting an integrated knowledge graph, in: The Semantic Web–ISWC 2020: 19th International Semantic Web Conference, Athens, Greece, November 2–6, 2020, Proceedings, Part II 19, Springer, 2020, pp. 430–446.

[13] E. Simperl, O. Corcho, M. Grobelnik, D. Roman, A. Soylu, M. J. F. Ruíz, S. Gatti, C. Taggart, U. S. Klima, A. F. Uliana, et al., Towards a knowledge graph based platform for public procurement, in: Research Conference on Metadata and Semantics Research, Springer, 2018, pp. 317–323.