Modelling the European Space Sector with Knowledge Graphs

Audrey Berquand^{*,a} **and Dominik Dold**^{*,b}

^{a}European Space Agency, Software Technology Section
^{b}European Space Agency, Advanced Concepts Team

**Both authors contributed equally to this work.*

Abstract. Launched in a *New Space* era, the space economy is experiencing rapid growth with an ever-increasing number of new commercial entrants. In this context, we present an approach based on Natural Language Processing (NLP) and Knowledge Graph Embedding (KGE) to model and analyse the socio-economic landscape of the European space sector. In this prototype study, we present the initial results obtained by extracting information from 3 databases containing information on R&D studies led at the European Space Agency (ESA), and on European companies. This information is merged in a semantically compatible way as a Knowledge Graph (KG). Combining NLP and KGE, we predict new links to complete and clean this KG, enabling novel insights into the integrated databases. The presented results demonstrate the potential of our approach for enhancing ecosystem monitoring, mapping existing capabilities, and identifying technology gaps. Although we obtain encouraging results, we also identify several challenges for adapting such an approach in production, to be solved in future studies.

1 Introduction

With lower barriers to entry, an increasing number of novel stakeholders[1] are arising across the world driving a commercialisation-driven *New Space* economy [6]. The role of the European Space Agency (ESA) is to shape and lead the development of Europe's space capability. ESA is an international organisation involving 22 member states and a public funding budget of, e.g., €7.15B in 2022. Monitoring a fast growing and diverse ecosystem is a challenge that is key for the agency to address. In fact, a sound ecosystem overview significantly accelerates and facilitates state-of-the-art (SOTA) reviews, the identification of existing capabilities and technology gaps – which are all essential to roadmap design.

Our study aims at leveraging Natural Language Processing (NLP), Knowledge Graph (KG), and Knowledge Graph Embedding (KGE) methods to generate – from existing databases – an overview of the European space ecosystem. The combination of NLP and KGE allows us to (i) complete the information missing in the used databases, and (ii) predict new links between entities to uncover novel and hidden relationships in the ecosystem. Although encouraging, the results presented here are preliminary and of limited scope. Therefore, they are not to be used to derive any conclusions about countries and stakeholders that appear in this work.

The contributions of this study are summarised as follows:

1. We create a KG structure representing the various aspects of the space ecosystem dynamics, which we populate from several heterogeneous public databases.
2. We demonstrate how KGE can assist in the completion of domain-specific KGs.
3. We demonstrate the potential of utilising this cleaned KG to unveil new insights into domain-specific ecosystems.

2 Related Work

KGE is an approach for statistical relational learning on KGs. It has been used in a variety of industrial applications, e.g., as an engineering design assistant for industrial products [9], for detecting anomalous behaviour in automated systems [7] and for building web-scale recommender systems [17]. Moreover, a specific form of graph embedding algorithm, so-called Graph Neural Network (GNN), have become a prominent algorithm for machine learning on graph-structured data (e.g., [15],[1], and [10]).

While previous studies [5, 2] have applied NLP embedding and KG mapping to space-related content (e.g. utilising large language models to autonomously construct a KG, see [2]), to our current knowledge, KGE methods have not yet been applied to space-related challenges – even though the space sector offers huge amounts of historical data (e.g., scientific projects and mission designs) that could be semantically integrated in a KG and analysed using KGE (or similar methods). Thus, in this study, we assess a combination of KG, NLP and KGE techniques to model the socio-economic landscape of the European space ecosystem – providing a first glimpse at the potential of these technologies for the space sector.

3 Assembling the Knowledge Graph

In the following, we describe the steps required to construct a KG. The two main building blocks for constructing a KG are (i) a manually defined schema describing the structure of the KG (Section 3.1) and (ii) data sources which are used to populate the KG (Section 3.2), i.e., we are filling in information from the data sources according to the schema to obtain the final KG.

3.1 Knowledge Graph Structure

The KG data model or schema defines the allowed entities, attributes, and relations in the graph. The schema shown in Figure 1 is tailored to the target use cases of this study. We defined 7 entities: *study*, *application*, *technologyDomain*, *competenceDomain*, *stakeholder*, *country*,

[1] By stakeholders, we mean all legal entities involved in space activities such as agencies, companies, start-ups and research institutes.

doi: 10.25967/610139

and *product*. The *study* and the *stakeholder* entities both have several attributes. All relations are n-ary, meaning a *stakeholder* can be related to *n product* entities or *n study* entities.

3.2 Data Sources

The data used in this study is extracted and merged from 3 databases: ESA's Nebula library[2], the ESTACA's space database[3], and the Nanosats Database's company table[4]. All database owners were contacted and agreed to sharing their data. Table 1 summarises where the data to populate the graph is sourced from.

The ESA Nebula library is a public database of ESA R&D projects listing 1,512 studies led by the agency from 1989 to today. Each study has a title, a written description, a start and end date, an ESA program it belongs to, an application, and keywords. There are 8 types of application domains defined by the agency: Earth observation, exploration, navigation, space safety, space science, space transportation, telecommunications, and generic technologies. A study is also related to one or several Competence Domains (CDs) and Technology Domains (TDs). To better coordinate its projects, the agency relies on 10 internally defined CDs[5] such as *CD 2: Structures, Mechanisms, Materials, and Thermal* or *CD10: Astrodynamics, Space Debris, and Space Environment*, and 26 TDs [6] such as *Space System Software*, *Propulsion* or *Ground Station System & Networks*. Finally, the database contains information on the companies involved in the study, along with their country of registration. The data is manually submitted by the study lead at the end of the project. It is, however, not mandatory to assign CDs and TDs. Therefore, only 11% of studies include this information.

Members of the French engineering school ESTACA recently released a 'space database' with verified data on 764 space stakeholders, and among them, 306 European entities. Each entity has a creation year, a description of its activities, tags or keywords, and a country of registration. Finally, the Nanosat database links, among other parameters, 272 European stakeholders to their products and services, and to their country of registration.

3.3 Knowledge Graph Population

Data from the 3 databases is extracted, processed, mapped to triples and to TypeQL insert queries. The processing is automated as much as possible, mainly with the use of regular expressions. Eventually, the resulting KG contains 3,167 populated entities (including 1,495 stakeholders), 5,830 links, and 6,068 attributes. The current KG is still preliminary as further cleaning is necessary to remove the remaining duplicate *stakeholder* entities. The overlap of stakeholders is rather weak between the Nebula and the other two databases: 14 entities in common with the Nanosats database and 18 with the ESTACA database. However, the overlap between the ESTACA and the Nanosats databases is higher, with 62 entities in common. Further processing is planned to increase the databases' overlap. Appendix A displays subsets of the KG, visualised with TypeDB Studio.

4 Multi-relational learning

KGs are purely symbolic structures, combining multi-relational information in a human-and computer-readable format. They are usually incredibly sparse, meaning that only a very small subset of true statements is contained in the KG, and the truth about all missing statements (i.e., non-existing links in the KG) is unknown. This is also known as the open-world assumption (OWA). KGE is a prominent approach for working with (noisy and incomplete) KGs, allowing us to infer novel, plausible links as well as detect known, implausible ones. Moreover, KGE enables the usage of machine learning methods such as neural networks or decision trees – which work on data represented by vector spaces – for analysing KGs.

In the following, we first introduce the concept of KGE as well as the link prediction task – one type of inference that can be performed on KGs. Afterwards, as a proof of concept, we demonstrate in preliminary experiments how KGE can be used to predict novel facts about the European space ecosystem – of course limited to the information contained in the 3 databases – from the accumulated KG.

4.1 Knowledge Graph Embedding

The main idea behind KGE is to find, for each element of the KG, a vector-based representation that preserves properties of the original KG's structure. For example, properties of interest are often the existence of links between entities, or the local neighbourhood of entities in a graph. Many methods for KGE exist nowadays. For this initial study, we limit ourselves to the well-known tensor-factorisation model RESCAL [11] that has the benefit of being conceptually simple while providing competitive performance levels. For an overview of current KGE approaches, see, e.g., [14, 8].

In RESCAL, each entity i of the graph is represented by a N-dimensional embedding vector $\boldsymbol{e}_i \in \mathbb{R}^N$. In addition, each relation type l (i.e., edge type) is represented by an embedding matrix[7] $\boldsymbol{R}_l \in \mathbb{R}^{N\times N}$. To evaluate the plausibility of a link (i.e., triple) (s, p, o) with entities s and o as well as relation type p, the score $\theta_{s,p,o}$ is calculated from the embeddings

$$\theta_{s,p,o} = \boldsymbol{e}_s^{\mathrm{T}} \boldsymbol{R}_p \boldsymbol{e}_o \,, \tag{1}$$

with $\theta_{s,p,o} \approx 1$ reflecting high and $\theta_{s,p,o} \approx 0$ low plausibility.

Embeddings are obtained using gradient-descent based optimisation by minimising a reconstruction loss $\mathcal{L}$ for the KG

$$\sum_{(s,p,o)\in \mathrm{KG}} \Big[(1-\theta_{s,p,o})^2 + \sum_{(i,j,k)\in \mathcal{C}_{s,p,o}} \theta_{i,j,k}^2 \Big] \,, \tag{2}$$

where $\mathcal{C}_{s,p,o}$ is a set containing $M \in \mathbb{N}$ corrupted triples, i.e., triples where either s or o have been replaced randomly with any other entity[8].

In the KG collected for our study, most attributes such as keywords and tags can be represented as triples as well. However, text features such as study descriptions cannot be represented in this way, and are therefore not accessible for RESCAL during training. In the following, we explain how this downside can be circumvented.

[2] https://nebula.esa.int/

[3] https://estaca-space-systems.notion.site/09044b95aac84076bb11077d124a665f

[4] https://www.nanosats.eu/

[5] https://www.esa.int/Enabling_Support/Space_Engineering_Technology/Shaping_the_Future/ESA_Competence_Domains

[6] https://www.esa.int/Enabling_Support/Space_Engineering_Technology/Technology_Domains

[7] Which is shared by all relations of the same type.

[8] For experiments, we use $N = 64$, $M = 10$, the Adam optimizer with a learning rate of 10^{-3} and regularization strength of 10^{-5} and a batch size of 1000 (triples).

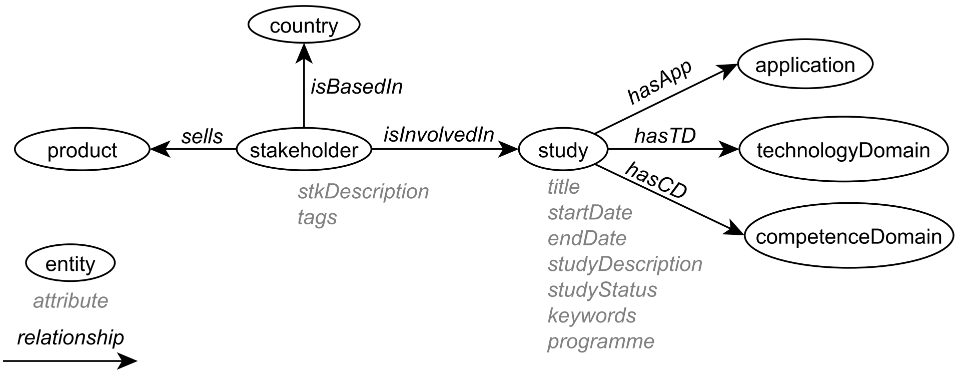

Figure 1. Unformalised KG structure. Information from the databases has to be parsed to be compatible with this structure, e.g., for a stakeholder, studies, products and country of origin are extracted for populating the KG. In addition, entities such as stakeholders or studies contain categorical or textual attributes.

Table 1. Distribution of information over the databases used in this study.

Db	Study	Application	Technology Domain	Competence Domain	Stakeholder	Country	Products
Nebula	✓	✓	✓	✓	✓	✓	-
Space Database	-	-	-	-	✓	✓	✓
Nanosats	-	-	-	-	✓	✓	✓

4.2 *Natural Language Embedding*

Text attributes can be represented as triples as well, with each attribute (i.e., each *studyDescription*) being represented by an individual entity, connected to a relation of type *hasAtt_studyDescription*. The RESCAL vector representations of the text descriptions are given by their sentence embeddings, and they are kept fixed during training such that the remaining KGEs learn to align accordingly with them. For instance, a description text *'This company focuses on EO systems for climate monitoring'* would be represented by a real-valued vector $\mathbf{e}_{text}$ obtained via sentence embedding and used to rank triples such as (*studyA*, *hasAtt_studyDescription*, *'This company focuses on EO systems for climate monitoring'*). To generate the embeddings, the *sentence-transformers* Python framework [13] is used. This framework also provides distillation solutions, notably enabling dimensionality reduction through Principal Component Analysis (PCA)[9]. For the results presented in Sections 4.4 and 4.5, we map the study descriptions into a $N = 64$ dimensional vector space and include them during the training of KGE models.

4.3 *Link prediction*

The main application of KGE in this study is the so-called link prediction task, where the trained embeddings are used to predict new links in the KG. To evaluate how well the model performs in this task, we use the well known mean rank (MR), mean reciprocal rank (MRR) and Hits@k metrics.

Here, we restrict ourselves to predicting either the *competenceDomain* or *application* of studies. This is done in the following way, described for the case of predicting applications:

1. Take a triple (*studyX*, *hasApp*, *Y*) from the test dataset, i.e., a triple we know is true, but has not been used to train the model.
2. Score this triple, as well as all possible alternatives (*studyX*, *hasApp*, *Z*), with *Z* being all other applications as long as the triple does not appear in the training dataset.
3. Create a sorted list, with the highest scored triple being the first on the list.
4. The rank of the test triple is given by its place in this list and its reciprocal rank is the inverse of this value. Hit@k is 1 if the test triple is among the first k entries of the list and 0 otherwise.

MR, MRR and Hits@k are obtained by averaging these quantities over the whole test dataset.

4.4 *Predicting Missing Information*

As stated in Section 4.1, KGs are typically highly sparse, with a lot of information missing. In our case, the study descriptions obtained from Nebula contain only 204 *competenceDomain* descriptions. Moreover, even though every study is assigned an *application*, only 150 studies have one that is different from the *generic_technologies* entity.

In the following, we demonstrate that KGE can be used to clean up the KG by proposing missing competence domains of studies, as well as providing more specific applications to studies by removing the *generic_technologies* entity from the KG and using KGE to predict more appropriate applications. This is possible since KGEs encode information about statistical patterns in the KG, learned from co-occurrences of entities and relations in triples (i.e., links) during training.

[9] https://github.com/UKPLab/sentence-transformers/blob/master/examples/training/distillation/dimensionality_reduction.py

In Table 2 and Table 3, we show the performance of RESCAL on these link prediction tasks. For testing, we put $1/4$ of triples with the *hasApp* and *hasCD* relation, respectively, randomly aside during training. Thus, for training to predict applications, 114 examples are found in the training data and we test on 36 examples. For competence domains, 159 examples are found in the training data and we test on 45 examples. As a baseline, we train both a linear regression model and a random forest[10] to predict applications / competence domains given the Natural Language (NL) embedding of a study's description text. It is important to note that RESCAL is trained on predicting all links in the KG, not just competence domains and applications.

Table 2. Predicting competence domains of studies in ESA's Nebula library.

Model	MR	MRR	Hits@1	Hits@3
RESCAL	1.53	0.86	0.78	0.93
Random Forest	2.40	0.64	0.47	0.80
Lin. Regression	2.58	0.65	0.49	0.71

Table 3. Predicting applications of studies in ESA's Nebula library.

Model	MR	MRR	Hits@1	Hits@3
RESCAL	1.08	0.97	0.94	1.00
Random Forest	1.50	0.86	0.78	0.92
Lin. Regression	1.72	0.83	0.75	0.89

In both cases, RESCAL reaches the best performance, providing reliable predictions for applications and competence domains. Thus, the trained model can be used to find new links, as shown in Table 4, where we assign updated application domains to studies – two of which had the *generic_technologies* and one that had a specific domain assigned before. For cross-reference, we also provide the study titles (not used during training).

Table 4. Applications assigned to studies using KGE. For the first study, the application with the highest score is *Exploration*, which is already contained in the training data. Thus, we show the model prediction with the second highest score as an alternative application label, consistent with the first one. The other two studies were only classified as *generic_technologies* in the raw KG, which was removed from the data before training.

Study title	Application / Score
Local sleep episodes during wakefulness and long term space travel.	Space Science ✓ 0.72
An AI-based system for an active tracking of Earth features from OPS-SAT.	Earth Observation ✓ 0.62
Integration of optical detectionin microfluidic systems for space exploration missions.	Exploration ✓ 0.41

4.5 Inferring Application Domain of Entities

The presented approach can further be used to detect trends in the data that are otherwise hard to detect with conventional methods like visual inspection and KG querying. For instance, in the following the learned KGEs are used to predict the application domain of companies and countries.

For a company, we obtain a measure of its affinity to a certain application domain by averaging over the scores of all its studies, i.e., evaluating triples of the form *(studyX, hasApp, Y)* using the learned embeddings, where *studyX* has the company as a stakeholder. This is demonstrated in Table 5, with ✓ denoting that the prediction is correct, ✗ denoting that its incorrect, and $(\cdot)$ denoting uncertainty in our estimation. As a cross-reference, we also provide brief company descriptions extracted from the official company webpages.

In cases where the KG contains almost no information about studies a company is part of (e.g., no keywords and no description), the quality of the predictions is drastically reduced, as shown for the company *Snecma* in Table 5. This could be mitigated by having additional information from another domain available, e.g., information about products a company sells, as contained in the ESTACA and Nanosat databases. Although we are merging these databases into our KG, they are not yet perfectly aligned, meaning that many entities appear in Nebula and the other databases under different names. However, we are confident that the presented results will greatly improve after solving this alignment problem.

The same approach used for companies can be applied to analyse the application focus of countries as well. Most importantly, since the learned embeddings are used to predict – from the information available in the KG – the most likely application for each study, this has a huge impact on the result compared to simply counting the few application domains available in the KG, as shown in Figure 2. First, as an example that missing information is recovered correctly using KGE, we remove all application domains of studies with stakeholders located in Sweden from the training dataset (2× *Earth Observation*). As shown in Figure 2, our model successfully recovers this information. In cases where several application domains are already present in the KG, due to most studies having no application assigned, we obtain a strong shift in the application landscape. For instance, different from the raw data where Germany seems to mostly focus on *Exploration* activities, the KGE model also assigns it a strong focus on *Space Safety* (e.g., debris removal) and *Space Transportation*. France, which in the data is also strongly biased towards *Exploration*, is assigned more weight for *Space Science* as well as *Telecommunication*.

5 Discussion

We demonstrate that KGs are a promising technology for modelling the European space ecosystem. In particular, KGs allow a meaningful integration of information about the space ecosystem from different domains, e.g., product information and scientific activities of stakeholders – which in turn can be accessed via modern machine learning methods to extract novel insights about the space ecosystem.

Although the shown results are promising, we identified two challenges that are being addressed in ongoing work. First, in its current state, the resulting KG does not perfectly align the information from the 3 databases – mostly due to differences in how stakeholders are named. This prohibits, for example, KGE methods to fully utilise information about company products to predict their competence domain as defined in ESA's R&D ecosystem. Solving this will require more pre-processing of the data (e.g., using NLP) combined with manual validation.

Secondly, the predictions generated by using NLP and KGE methods on the KG are hard to verify, and can currently only be used to get initial insights for further, manual data analysis. This can be mitigated by including Explainable Artificial Intelligence (XAI) approaches that, e.g., in addition to a prediction, also return the sub-KG that has been most important for this output [18], or leverage Language Models to provide human-readable explanations [16]. In addition, a human-in-the-loop solution could be applied as well, where human-verified model outputs are integrated into the original KG, and

[10] With ensemble size 200 and maximum tree depth of 6.

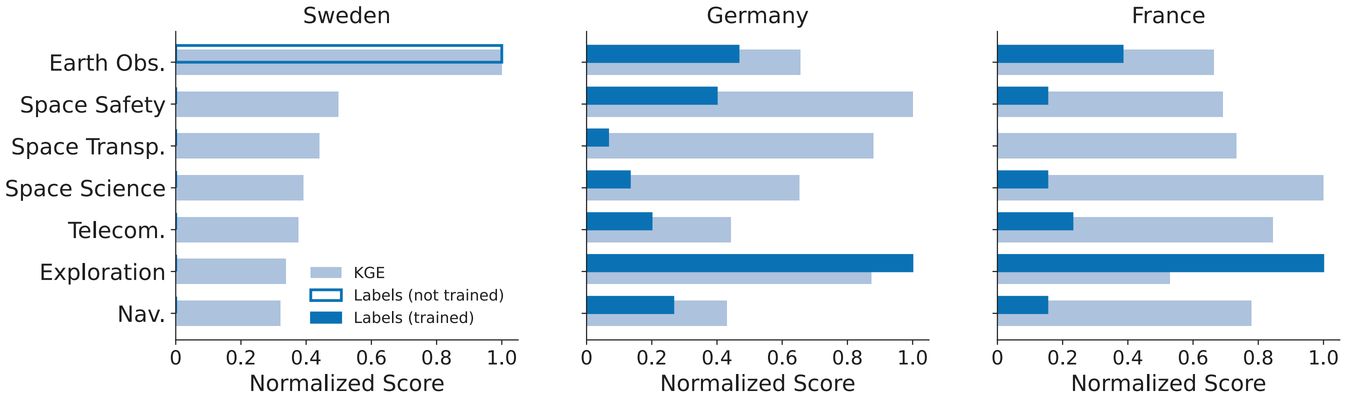

Figure 2. Predicting application domains of countries. Labels are obtained by counting the occurrence of the application domain in the KG. Both label counts as well as the scores obtained from RESCAL are normalised on the respective maximum value.

Table 5. Predicting application domain of companies.

Company	Noveltis	QinetiQ	Snecma
Highest	Earth Obs. ✓	Space Science ✓	Earth Obs. ✗
Lowest	Telecom (✓)	Earth Obs. (✓)	Exploration (✗)
Descriptions (company webpages)	Noveltis significantly contributes to the development of new Earth Observation missions, whether related to the atmosphere, oceans or land surfaces.[11] We (QinetiQ) are a world-centre of excellence in research and development, and act as catalyst for fast-track innovation, offering outstanding experimentation facilities, technical, engineering and scientific expertise.[12] Safran Aircraft Engines *(prev. Snecma)* — A world-class aircraft engines manufacturer.[13]		

in-production evaluation is only performed by querying the KG. Ultimately, to identify the most suitable solution, a survey with domain experts will be performed.

Finally, in future work, the presented KG will be enhanced by adding further data sources to improve our underlying model of the European space ecosystem. Likewise, we will also investigate the suitability of a variety of NLP and KGE methods for the aforementioned challenges in detail.

6 Conclusion

Space, as famously quoted, is not only the final frontier, but an essential and strongly growing part of the European economic system; being both a key enabler for innovative, nature-preserving technologies as well as opening new opportunities for research and companies alike. In this study, we present initial results on modelling the European space ecosystem using KGs and machine learning techniques such as NLP and KGE. We are confident that the proposed framework will yield a fundamental contribution to the field by assisting experts in the space research, technology and economy sector to navigate the European space ecosystem, e.g., by allowing them to connect market needs and stakeholders, but also identify under-represented competences early.

Ethics Statement

The study has been performed following the Ethics Guidelines for Trustworthy Artificial Intelligence of the European Commission. To construct the used KG, only publicly available data has been used with permission of the database owners. The companies and countries specifically used in this study for model evaluation have been selected arbitrarily. No conclusion about the competence and domain expertise of said countries and companies is to be derived from this study, as it represents only preliminary results of limited scope. Finally, the tools developed in this study have the aim of bolstering the economic development of the European space sector, ultimately supporting highly relevant applications such as climate change monitoring, as well as enabling European leadership and autonomy.

Software

For sentence embeddings, we use the *sentence-transformers* Python library [13], *scikit-learn* [12] for the PCA decomposition and cosine similarity, and *nltk* [3] for various processing tasks. For KGE, we use custom implementations for, e.g., data loaders and link prediction evaluation, building upon the source code of the *TorchKGE* [4] Python library. For baseline models, we use *scikit-learn*. For the KG, we use the open-source Vaticle TypeDB database[14]. The visualisation of the KG is done through their TypeDB Studio interface. We use their Python client to read from and write to a KG. TypeQL is the query language of TypeDB.

Acknowledgements

The authors thank ESA for its support through the Research Fellowship program. A special thanks goes to Ana Victória Ladeira (ESA),

[14] https://vaticle.com/

Charles Poncet (ESA), and Francisco Faes (Immedia) for their support with accessing the nebula database. We also thank the moderators of the ESTACA database: Thomas Garnier, Théo Hannoun, and Nicolas Thierry and of the Nanosats database: Erik Kulu. Finally, we warmly thank Luis Mansilla (ESA) for his precious advice.

References

[1] Victor Bapst, Thomas Keck, Agnieszka Grabska-Barwińska, Craig Donner, Ekin Dogus Cubuk, Samuel S Schoenholz, Annette Obika, Alexander WR Nelson, Trevor Back, Demis Hassabis, et al., 'Unveiling the predictive power of static structure in glassy systems', *Nature Physics*, **16**(4), 448–454, (2020).

[2] Audrey Berquand and Ana Victoria Ladeira, 'From mission description to knowledge graph: Applying transformer-based models to map knowledge from publicly available satellite datasets.', in *Proceedings of the 10th International Systems & Concurrent Engineering for Space Applications Conference (SECESA)*, (2022).

[3] Steven Bird, Ewan Klein, and Edward Loper, *Natural Language Processing with Python*, O'Reilly Media, Inc., 1st edn., 2009.

[4] Armand Boschin, 'Torchkge: Knowledge graph embedding in python and pytorch', in *International Workshop on Knowledge Graph: Mining Knowledge Graph for Deep Insights*, (Aug 2020).

[5] Paul Darm, Audrey Berquand, Luis Mansilla, and Annalisa Riccardi, 'A system engineering recommendation system based on language similarity analysis: an application to space systems conceptual design', in *Proceedings of the 10th International Systems & Concurrent Engineering for Space Applications Conference (SECESA)*, (2022).

[6] Alessandro de Concini and Jaroslav Toth, *The future of the European space sector. How to leverage Europe's technological leadership and boost investments for space ventures*, European Investment Bank, 2019.

[7] Josep Soler Garrido, Dominik Dold, and Johannes Frank, 'Machine learning on knowledge graphs for context-aware security monitoring', in *2021 IEEE International Conference on Cyber Security and Resilience (CSR)*, pp. 55–60. IEEE, (2021).

[8] Will Hamilton, Zhitao Ying, and Jure Leskovec, 'Inductive representation learning on large graphs', *Advances in neural information processing systems*, **30**, (2017).

[9] Marcel Hildebrandt, Swathi Shyam Sunder, Serghei Mogoreanu, Mitchell Joblin, Akhil Mehta, Ingo Thon, and Volker Tresp, 'A recommender system for complex real-world applications with nonlinear dependencies and knowledge graph context', in *The Semantic Web: 16th International Conference*, pp. 179–193. Springer, (2019).

[10] Borja Ibarz, Vitaly Kurin, George Papamakarios, Kyriacos Nikiforou, Mehdi Bennani, Róbert Csordás, Andrew Joseph Dudzik, Matko Bošnjak, Alex Vitvitskyi, Yulia Rubanova, et al., 'A generalist neural algorithmic learner', in *Learning on Graphs Conference*. PMLR, (2022).

[11] Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel, 'Factorizing yago: scalable machine learning for linked data', in *Proceedings of the 21st international conference on World Wide Web*, pp. 271–280, (2012).

[12] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay, 'Scikit-learn: Machine learning in Python', *Journal of Machine Learning Research*, **12**, 2825–2830, (2011).

[13] Nils Reimers and Iryna Gurevych, 'Sentence-bert: Sentence embeddings using siamese bert-networks', in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, (11 2019).

[14] Daniel Ruffinelli, Samuel Broscheit, and Rainer Gemulla, 'You can teach an old dog new tricks! on training knowledge graph embeddings', in *International Conference on Learning Representations (ICLR)*, (2020).

[15] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling, 'Modeling relational data with graph convolutional networks', in *The Semantic Web: 15th International Conference*, pp. 593–607. Springer, (2018).

[16] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou, 'Chain of thought prompting elicits reasoning in large language models', (January 2022).

[17] Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L Hamilton, and Jure Leskovec, 'Graph convolutional neural networks for web-scale recommender systems', in *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 974–983, (2018).

[18] Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec, 'Gnnexplainer: Generating explanations for graph neural networks', *Advances in neural information processing systems*, **32**, (2019).

A Technical Appendix: Knowledge Graph Visualisation

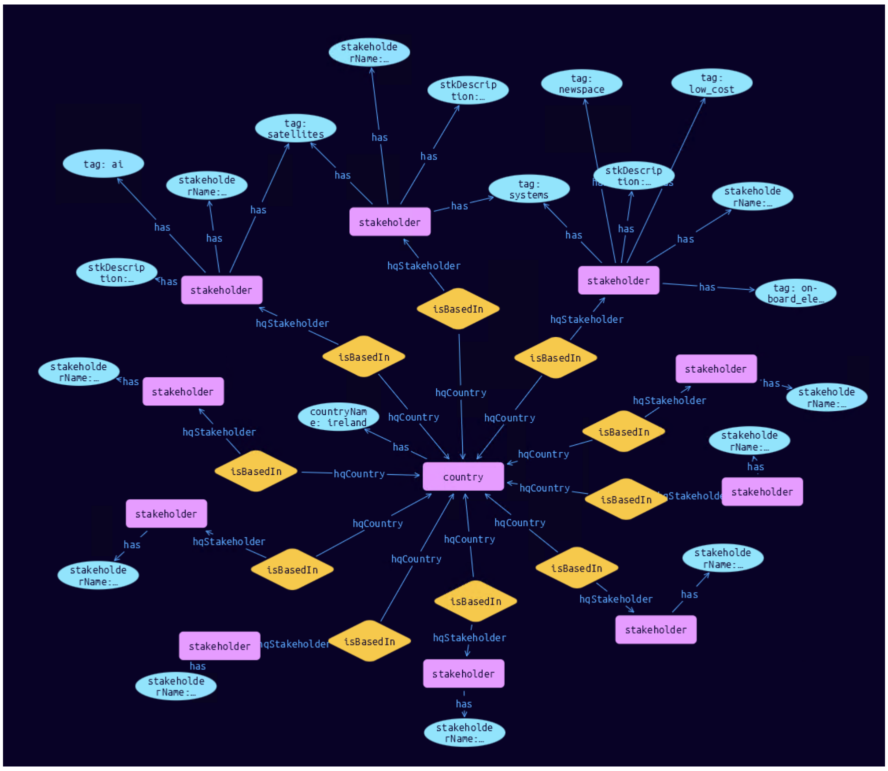

Figure 3. Overview of the stakeholder entities linked to Ireland. Visualised with TypeDB Studio.

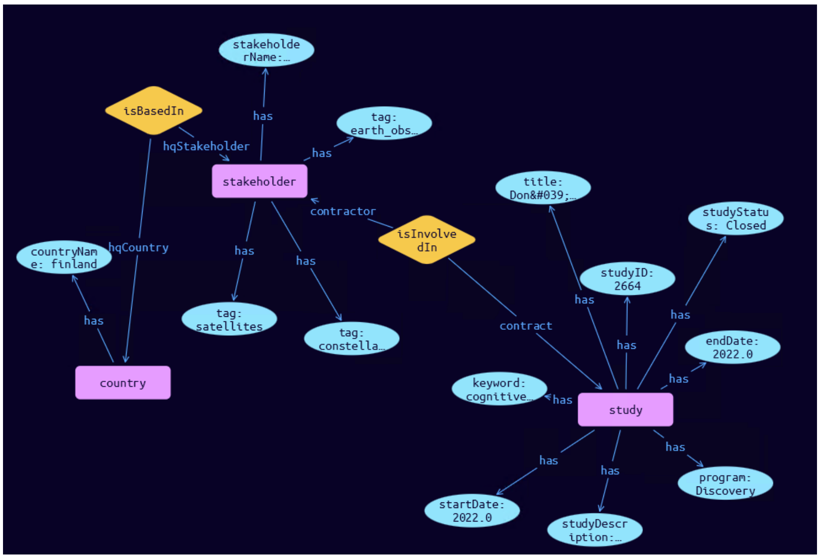

Figure 4. Overview of the entities linked to *Iceye*, a Finnish company. Visualised with TypeDB Studio.