

# Selling Decentralized Knowledge Graphs

Luis-Daniel Ibáñez<sup>1</sup> and George Konstantinidis<sup>1</sup>

<sup>1</sup> University of Southampton, Highfield Campus, Southampton, United Kingdom

## Abstract

Knowledge Graphs encode valuable information that has been so far restricted to the companies that developed them, or dependent on public subsidies. Data marketplaces have emerged as platforms where data consumers meet providers to find the data they need and willing to pay for, as an answer to the need of enabling data transactions. In this position paper we examine the specifics of selling Knowledge Graphs with respect to general datasets, and how can we sell decentralized Knowledge Graphs in a decentralized way.

## Keywords

Knowledge Graphs, Decentralization, Data Pricing, Privacy

## 1. Introduction

Knowledge Graphs (KGs) have consolidated as a type of knowledge representation used to represent structured information about entities, their attributes and their relationships [1]. Following Linked Data principles, Knowledge graphs may also be linked to other Knowledge Graphs, enabling the discovery and exploration of further knowledge not necessarily hosted or controlled by a single party. KGs have been used in industry to power a broad range of use cases, from discovery and exploration to causal inference. Crowdsourced textual resources like Wikipedia have been represented as KGs to enable their processing by machine agents and provide advanced query capabilities (YAGO, Dbpedia, Wikidata). Public sector has invested in using KGs to represent catalogs and open government datasets.

Traditionally, usage of Industry-developed KGs has been restricted to the boundaries of the companies that developed them to improve internal services, however, new business cases have emerged that require to sell access to a subset of the KG to 3<sup>rd</sup> parties. On the other hand, KGs constructed by charities or by the public sector depend on subsidies, raising questions about their sustainability.

Enter data marketplaces, platforms to enable data owners to publish descriptions of their datasets and data consumers to discover and integrate published datasets. Alternatively, when there is no need for discovery because partners already know each other, a marketplace provide a trusted space for them to execute a data transaction [2,3]. In this position paper we present the challenges and opportunities of the intersection of KGs and Data marketplaces across two dimensions: (i) How to effectively sell KGs and (ii) how to decentralize the management of KG transactions. We argue that the explicit representation of *semantics* and the potential existence of *dereferenceable links* introduced by the Linked Data Principles provide additional challenges over generic data marketplaces.

## 2. Selling Knowledge Graphs

In this section we present the challenges and problems associated with selling KGs. For each theme we discuss related work on general datasets, identify key differences for KGs, and formulate research questions. When applicable, we also discuss how established techniques for KGs could be adapted to support the sale of a KG.

---

Proceedings of the ESWC 2023 Workshops and Tutorials, May 28–29th, 2023, Hersonissos, Greece

EMAIL: l.d.ibanez@southampton.ac.uk (A. 1); g.konstantinidis@southampton.ac.uk (A. 2)

ORCID: 0000-0001-6993-0001 (A. 1); 0000-0002-3962-9303 (A. 2)



Copyright © 2023 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)

## 2.1. Data Pricing and Valuation

[4] studied the query pricing problem for general databases: assigning a price to a query in such a way that arbitrage, i.e., the possibility of a buyer to buy sub-queries of the target query and aggregate them themselves at a lower price than the original query, is minimized. Their theoretical results are in general negative: to be certain to avoid arbitrage, one needs to overprice many sub-queries which might be more detrimental for business than allowing arbitrage in the first place. The realization that federated learning is the most relevant industrial use case has led to a lot of research attention on how to quantify the contribution of a dataset to the accuracy of a ML model. [5] does a comprehensive survey of data pricing mechanisms. In the Semantic Web community, dealing with non-open data was first discussed as part of the “Linked Closed Data” vision [6], however, research has been dominated by Linked Open Data, probably because is easier to validate and experiment with Open Data. The pricing of a Semantic Web resource re-appears as a “Blue Sky” idea as late as 2018 [7]. A marketplace for the Web of Data is presented in [8], its main assumption is that advertising is useless because the Web of Data is run by automated agents, proposing auction mechanisms as an alternative.

We posit that query pricing in KGs will follow the same general results than those found in databases. A potential research direction is to exploit the semantics of the KG to fine-tune the prices of different subsets to strike an appropriate balance between arbitrage and reasonable prices. Another interesting difference is to assess if the potential price of a de-reference to another KG. In terms of collaborative construction of a KG, we believe Shapley-value techniques as used for ML models will still be relevant, but target metrics will be much more complex than “more accurate”. What are general target metrics for a KG? If they depend on Semantics, what algorithms to derive them?

## 2.2. Metadata and Summary Generation

The first step of a data owner that wants to sell in a data marketplace is to create a description to publish, in a very similar way as when wanting to publish to join the Linked Data Cloud. We expect existing tools for generating metadata [9,10,11] to be applicable. Summary generation techniques both for human and machine consumption will also be of special utility in this context [12,13,14]. We posit the main difference will be the context of advertising for sale: on the metadata front, we could generate metadata to match the requirements of a data buyer, a direction explored for scientific datasets [15]; on the summary front, a summary may reveal too much information, in combination with approaches for valuation, how to determine the most useful summary that reveals the least amount of information? Can we interactively generate increasingly complete paid summaries?

## 2.3. Privacy and usage control

Another important aspect when selling KGs is the definition of privacy and usage policies. This is particularly important for KGs that include personal data. Several works have studied how to define and evaluate usage policies for queries [16,17], and developed vocabularies for expressing and querying privacy [18,19]. We expect the evolution of these vocabularies and evaluation techniques to enable finer granularity (up to the level of a statement) and incorporate elements of explainability: ‘Why I can’t use this dataset?’ and negotiation: What needs to change so I can use it? can I convince the owner to relax the policy in exchange of (money, other data, etc.)?

## 2.4. Service descriptions

Transactions in a data marketplace are not restricted to datasets. The ecosystem is completed with services (aka Data Apps or Data Processing Operations) that can be used to analyze data, leading to the problem of given dataset descriptions, find services that could be used to process them according to certain desired characteristics. The problem relates to that of Web Service Discovery: matching formally described user requests with service functionality satisfying these requests [20,21]. As with datasets, the main blocker is the problem of generating the description as automatically as possible.

### 3. Decentralizing KG transactions

So far, we have assumed that a transaction occurs in a centralized data marketplace controlled by a single operator trusted by the transaction participants. However, from an economics point of view we expect multiple operators develop different marketplaces with different rules and value propositions. Data owners may choose to advertise their datasets in one or many data marketplaces and data consumers may need to use datasets. At a data level, we believe this is an iteration of the Linked Closed Data vision from 2011 [6]. They had the same concerns about usage policies (but using the term *licence*) and foresaw the need for descriptions of cost of premium subsets of the data. More recently, [25] explored the use of Web Monetization on top of decentralized Solid applications. We identify two additional challenges in the modern context that arise when we consider where the dataset will be processed.

#### 3.1. Decentralized execution

The simple assumption is that either the marketplace operator or one of the participants is entrusted to receive all data and algorithms to execute the agreed data processing workflow. Sadly, there are several scenarios where this is not desirable, the most evident being a dataset' usage policies forbidding its transfer outside the owner's infrastructure. The most active related area of research is Federated Learning, where multiple organisations train a Machine Learning Model without leaking data [22]. How to generalize this to any type of data operation or to a sequence of data operation? The problem is reminiscent to that of Workflow Management Systems, that has been revisited to support Big Data processing pipelines [23]. More generally, Trusted Execution Environments [26] have been proposed to deal with this problem, and have been demonstrated in "privacy-preserving" data marketplaces [27].

In a marketplace context, questions about payment arise. What protocols to authorise access and unlock payment based on Service Level Agreement conditions when data and algorithms are decentralized? Are Blockchains like in IoT marketplaces [24] the most efficient solution? Or are there more efficient protocols that only require a minimal amount of trust among them?

#### 3.2. Governance

In an ideal (Linked Data) world and in many IoT visions transactions automated agents consume and follow the published rules and disputes are avoided. When humans and money are involved, dishonest actors may try to breach usage conditions, or simply not deliver the promise of its specification. In a centralized scenario, the data marketplace operator may have the mechanisms to identify and punish dishonest members. But how to achieve the same in a decentralized scenario? If we assume it is not possible to agree on a centralised arbiter that monitors the behaviour of participants, what protocols the federated marketplace operators can implement? What is the minimum amount of information that needs to be exchanged or maintained?

### 4. Conclusion

In this paper, we examined the particularities of selling Knowledge Graphs on the light of the emergence of data marketplaces. We reviewed related work across two dimensions: (i) selling Knowledge Graphs (ii) decentralized sale of Knowledge Graphs. Our intent is this description will inspire researchers to coordinate efforts in this direction

### 5. Acknowledgements

Authors were supported by the European Union's Horizon Europe research and innovation actions under Grant Agreement N° 101093216. More information available at <https://upcastproject.eu/>

## 6. References

- [1] Fensel, D. *et al.* (2020). Introduction: What Is a Knowledge Graph?. In: Knowledge Graphs. Springer, Cham. [https://doi.org/10.1007/978-3-030-37439-6\\_1](https://doi.org/10.1007/978-3-030-37439-6_1)
- [2] R. C. Fernandez, P. Subramaniam, and M. J. Franklin, 'Data market platforms: trading data assets to solve data problems', *Proc. VLDB Endow.*, vol. 13, no. 12, pp. 1933–1947, Aug. 2020, doi: [10.14778/3407790.3407800](https://doi.org/10.14778/3407790.3407800).
- [3] J. Kennedy, P. Subramaniam, S. Galhotra, and R. Castro-Fernandez. Revisiting Online Data Markets in 2022: A Seller and Buyer Perspective. *SIGMOD Rec.* 51, 3, 2022 <https://doi.org/10.1145/3572751.3572757>
- [4] P. Koutris, P. Upadhyaya, M. Balazinska, B. Howe, and D. Suciu, 'Query-Based Data Pricing', *J. ACM*, vol. 62, no. 5, p. 43:1-43:44, Nov. 2015, doi: [10.1145/2770870](https://doi.org/10.1145/2770870).
- [5] J. Pei, "A Survey on Data Pricing: From Economics to Data Science," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 10, pp. 4586-4608, 2022.
- [6] M. Cobden, J. Black, N. Gibbins, L. Carr, N. Shadbolt, 'A Research Agenda for Linked Closed Data'. *Second International Workshop on Consuming Linked Data*, 2011
- [7] H. Paulheim, 'How much is a Triple? Estimating the Cost of Knowledge Graph Creation'. Proceedings of the ISWC 2018 Posters & Demonstrations, Industry and Blue Sky Ideas Tracks co-located with 17th International Semantic Web Conference. 2018
- [8] T. Grubenmann, A. Bernstein, D. Moor, and S. Seuken, 'Financing the Web of Data with Delayed-Answer Auctions', in *Proceedings of the 2018 World Wide Web Conference*, Republic and Canton of Geneva, Switzerland, 2018, pp. 1033–1042. doi: [10.1145/3178876.3186002](https://doi.org/10.1145/3178876.3186002).
- [9] P. Heyvaert, A. Dimou, B. De Meester, T. Seymoens, A-L. Herregodts, R. Verborgh, D. Schuurman, E. Mannens, 'Specification and implementation of mapping rule visualization and editing: MapVOWL and the RMLEditor', *Journal of Web Semantics*, Volume 49, 2018.
- [10] A. Assaf, A. Senart, and R. Troncy. 'Roomba: Automatic Validation, Correction and Generation of Dataset Metadata'. In Proceedings of the 24th International Conference on World Wide Web (WWW '15 Companion), 2015.
- [11] Kacprzak, E. *et al.* Making Sense of Numerical Data - Semantic Labelling of Web Tables. In: Faron Zucker, C., Ghidini, C., Napoli, A., Toussaint, Y. (eds) Knowledge Engineering and Knowledge Management. EKAW 2018
- [12] Goasdoué, F., Guzewicz, P. & Manolescu, I. RDF graph summarization for first-sight structure discovery. *The VLDB Journal* 29, 1191–1218 (2020)
- [13] Troullinou, G., Kondylakis, H., Daskalaki, E., Plexousakis, D. RDF Digest: Efficient Summarization of RDF/S KBs. In: Gandon, F., Sabou, M., Sack, H., d'Amato, C., Cudré-Mauroux, P., Zimmermann, A. (eds) The Semantic Web. Latest Advances and New Domains. ESWC 2015. 2015
- [14] L. Koesten, E. Simperl, T. Blount, E. Kacprzak, J. Tennison, 'Everything you always wanted to know about a dataset: Studies in data summarisation', *International Journal of Human-Computer Studies*, Volume 135, 2020.
- [15] M. Färber and A-K. Leisinger. 'DataHunter: A System for Finding Datasets Based on Scientific Problem Descriptions'. In Proceedings of the 15th ACM Conference on Recommender Systems (RecSys '21). 2021.
- [16] P. Upadhyaya, M. Balazinska, D. Suciu. 2015. Automatic Enforcement of Data Use Policies with DataLawyer. In Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data (SIGMOD '15), 2015
- [17] G. Konstantinidis, J. Holt, and A. Chapman. Enabling personal consent in databases. *Proc. VLDB Endow.* 15, 2, 2021.
- [18] Bonatti, P. A., Bos, B., Decker, S., Fernandez Garcia, J. D., Kirrane, S., Peristeras, V., Polleres, A., & Wenning, R. (2018). *Data Privacy Vocabularies and Controls: Semantic Web for Transparency and Privacy*. 1-1.
- [19] Oltramari, Alessandro et al. 'PrivOnto: A Semantic Framework for the Analysis of Privacy Policies'. *Semantic Web*, vol. 9, no. 2, 2018.

- [20] M. Kifer, R. Lara, A. Polleres, C. Zhao, U. Keller, H. Lausen and D. Fensel, 'A Logical Framework for Web Service Discovery'. Proceedings of the ISWC 2004 Workshop on Semantic Web Services: Preparing to Meet the World of Business Applications, 2004
- [21] Cabral, L., Domingue, J., Motta, E., Payne, T., Hakimpour, F. (2004). Approaches to Semantic Web Services: an Overview and Comparisons. In: Bussler, C.J., Davies, J., Fensel, D., Studer, R. (eds) The Semantic Web: Research and Applications. ESWS 2004.
- [22] D. Roman *et al.*, "Big Data Pipelines on the Computing Continuum: Tapping the Dark Data," in *Computer*, vol. 55, no. 11, pp. 74-84, 2022.
- [23] C. Zhang, Y. Xie, H. Bai, B. Yu, W. Li, Y. Gao, A survey on federated learning, *Knowledge-Based Systems*, Volume 216, 2021.
- [24] Christidis J, Karkazis PA, Papadopoulos P, Leligou HC. Decentralized Blockchain-Based IoT Data Marketplaces. *Journal of Sensor and Actuator Networks*. 2022
- [25] Sebrechts, M. *et al.* (2022). Solid Web Monetization. In: Di Noia, T., Ko, IY., Schedl, M., Ardito, C. (eds) Web Engineering. ICWE 2022. Lecture Notes in Computer Science, vol 13362. Springer, Cham. [https://doi.org/10.1007/978-3-031-09917-5\\_40](https://doi.org/10.1007/978-3-031-09917-5_40)
- [26] M. Sabt, M. Achemlal and A. Bouabdallah, "Trusted Execution Environment: What It is, and What It is Not," *2015 IEEE Trustcom/BigDataSE/ISPA*, Helsinki, Finland, 2015, pp. 57-64, doi: 10.1109/Trustcom.2015.357
- [27] Nick Hynes, David Dao, David Yan, Raymond Cheng, and Dawn Song. 2018. A demonstration of sterling: a privacy-preserving data marketplace. *Proc. VLDB Endow.* 11, 12 (August 2018), 2086–2089. <https://doi.org/10.14778/3229863.3236266>