UNIVERSITÀ
DEGLI STUDI
DI SALERNO

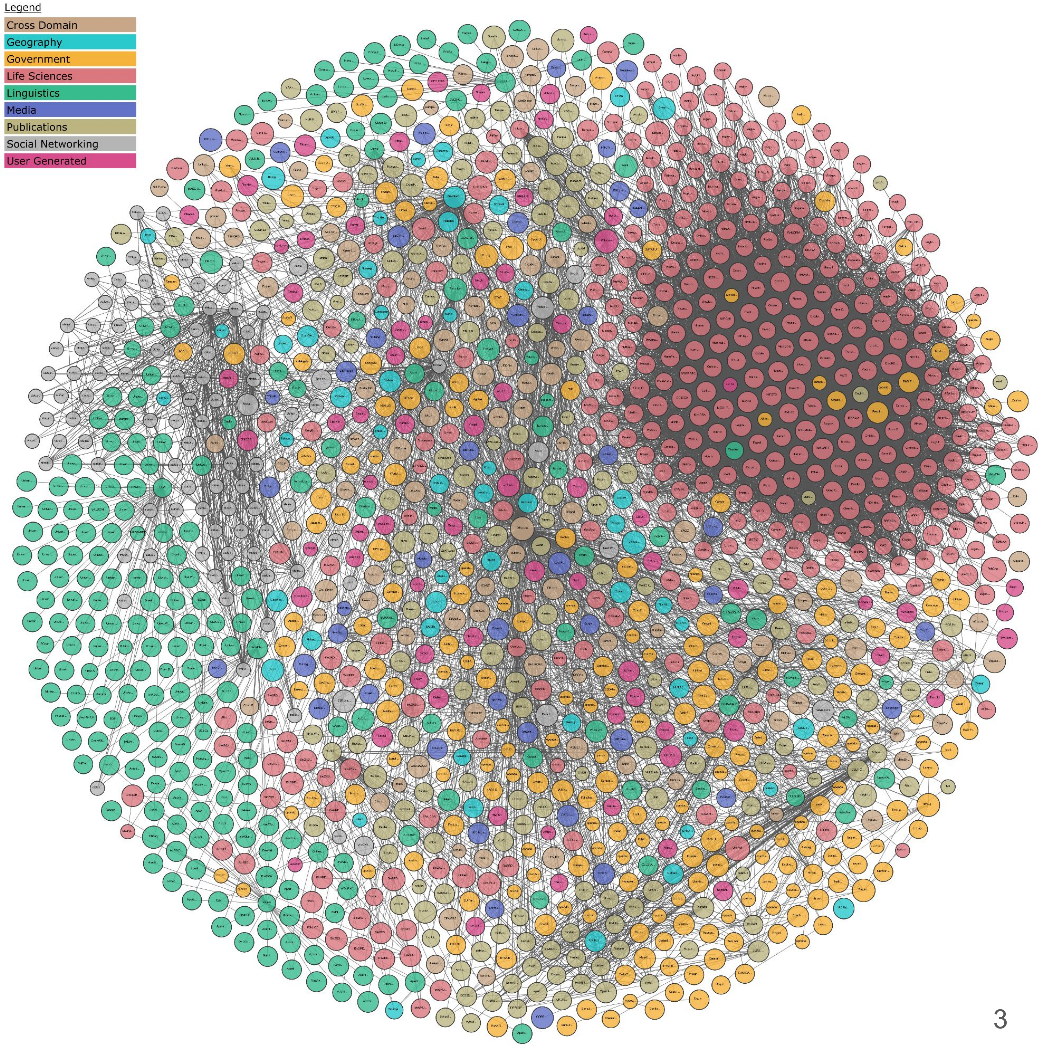
THE WEB CONFERENCE ACM

Navigating the LOD Subclouds: Assessing Linked Open Data Quality by Domain

Gabriele Tuozzo
Dipartimento di Informatica,
Università degli Studi di Salerno
Fisciano, Salerno, ITALY

1. Introduction

- **1,656** resources registered in The Linked Open Data Cloud (LOD Cloud) in the November 24, 2024 snapshot.
- **9** different subclouds:
  - Cross domain
  - Geography
  - Government
  - Life Sciences
  - Linguistics
  - Media
  - Publications
  - Social Networking
  - User Generated



The Linked Open Data Cloud from lod-cloud.net

1. Introduction

The contributions of this work are as follows:

- Examining changes in subcloud quality with respect to the pas to identify **persistent trends**, **highlight improvements** and **pinpoint areas of decline**.
- Providing an overview of the quality variation across different subclouds, with a focus on the **six quality categories** measured by **KGHeartBeat**.
- The analysis seeks to answer the following ***Research Question (RQ)***:

Is quality consistent across all subclouds?

2. Background - The quality framework adopted

This study builds upon the quality framework proposed by **Zaveri et al. [1]** and its adaptation by **Pellegrino et al. [2]**, which defines **6 quality categories**, further divided into quality dimensions:

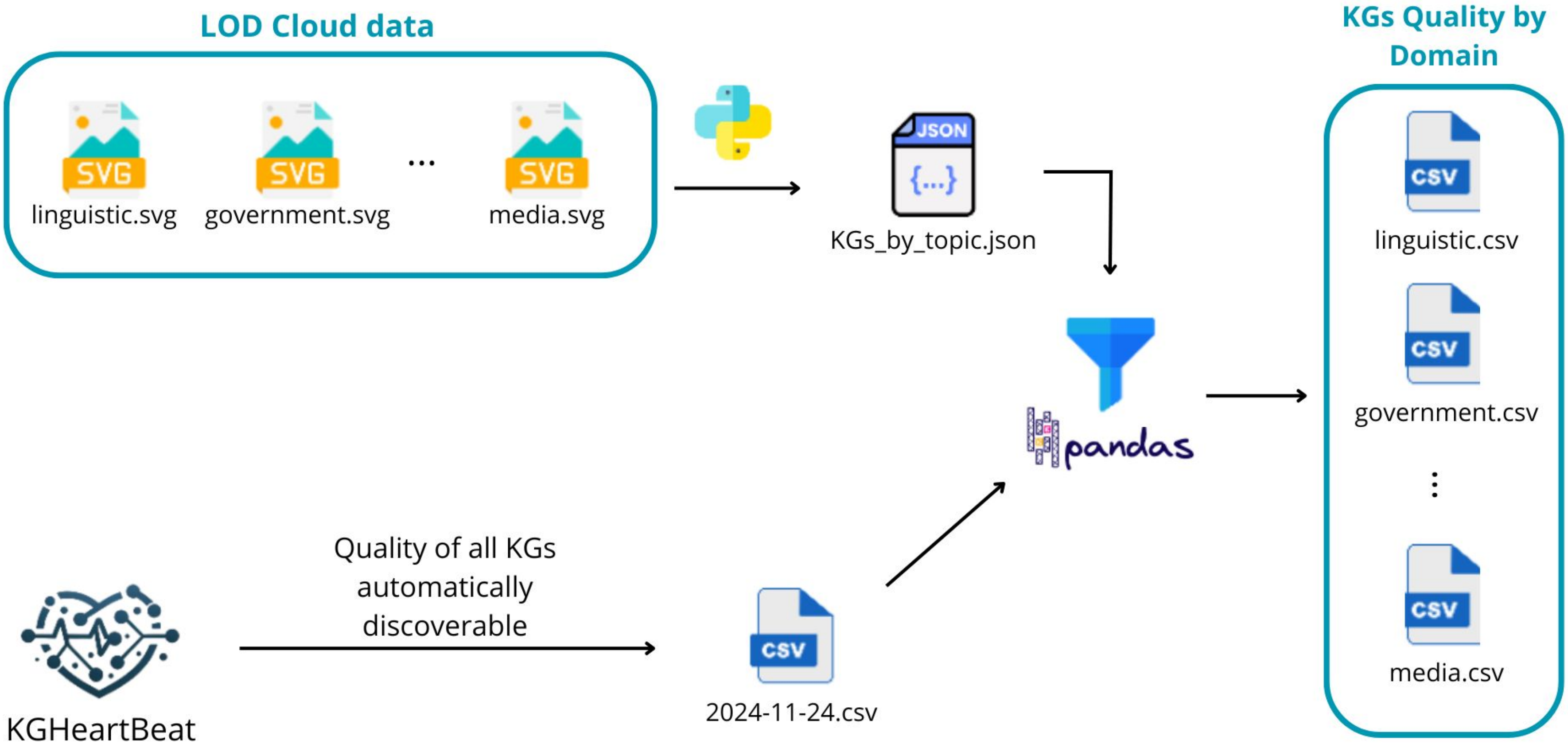
1. ***Accessibility***, covers dimensions related to data access, authenticity, and retrieval.
2. ***Contextual**, focuses on dimensions influenced by task-specific contexts.*
3. ***Dataset Dynamicity**, examines the currency and timeliness of published data.*
4. ***Intrinsic**, includes dimensions independent of user context*
5. ***Representational*** addresses dimensions concerning the design and data presentation.
6. ***Trust** evaluates dimensions related to trustworthiness*

[1] Amrapali Zaveri, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann, and Soeren Auer. 2016. Quality assessment for linked data: A survey. Semantic Web 7, 1 (2016), 63–93. https://doi.org/10.3233/SW-150175.

[2] Maria Angela Pellegrino, Anisa Rula, and Gabriele Tuozzo. 2024. KGHeartBeat: An Open Source Tool for Periodically Evaluating the Quality of Knowledge Graphs. In International Semantic Web Conference. Springer, 40–58. https://doi.org/10.1007/978-3-031-77847-6_3

Ref.	Analysis back to…	Focus	Quality Categories					
			A	C	D	I	R	T
State of the LOD cloud [3]	2011	All subclouds (x7)	✓					
Schmachtenberg et al. [4]	2014	All subclouds (x8)	✓					
Debattista et al. [5]	2015	LOD Cloud	✓			✓	✓	✓
Assaf et al. [6]	2016	LOD Cloud	✓	✓	✓	✓	✓	
Debattista et al. [7]	2018	LOD Cloud	✓	✓		✓	✓	
Yamamoto et al. [8]	2018	Life sciences	✓					
Maillot et al. [9]	2020-2021	LOD Cloud	✓	✓			✓	✓
Delgado et al. [10]	2021	Cultural Heritage	✓	✓		✓	✓	✓
Candela et al.[11]	2022	Cultural Heritage	✓	✓	✓	✓	✓	✓
di Buono et al. [12]	2022	Linguistic	✓				✓	
Esposito et al.[13]	2024	Linguistic	✓			✓		
This	2024	All subclouds (x9)	✓	✓	✓	✓	✓	✓

3. Methodology



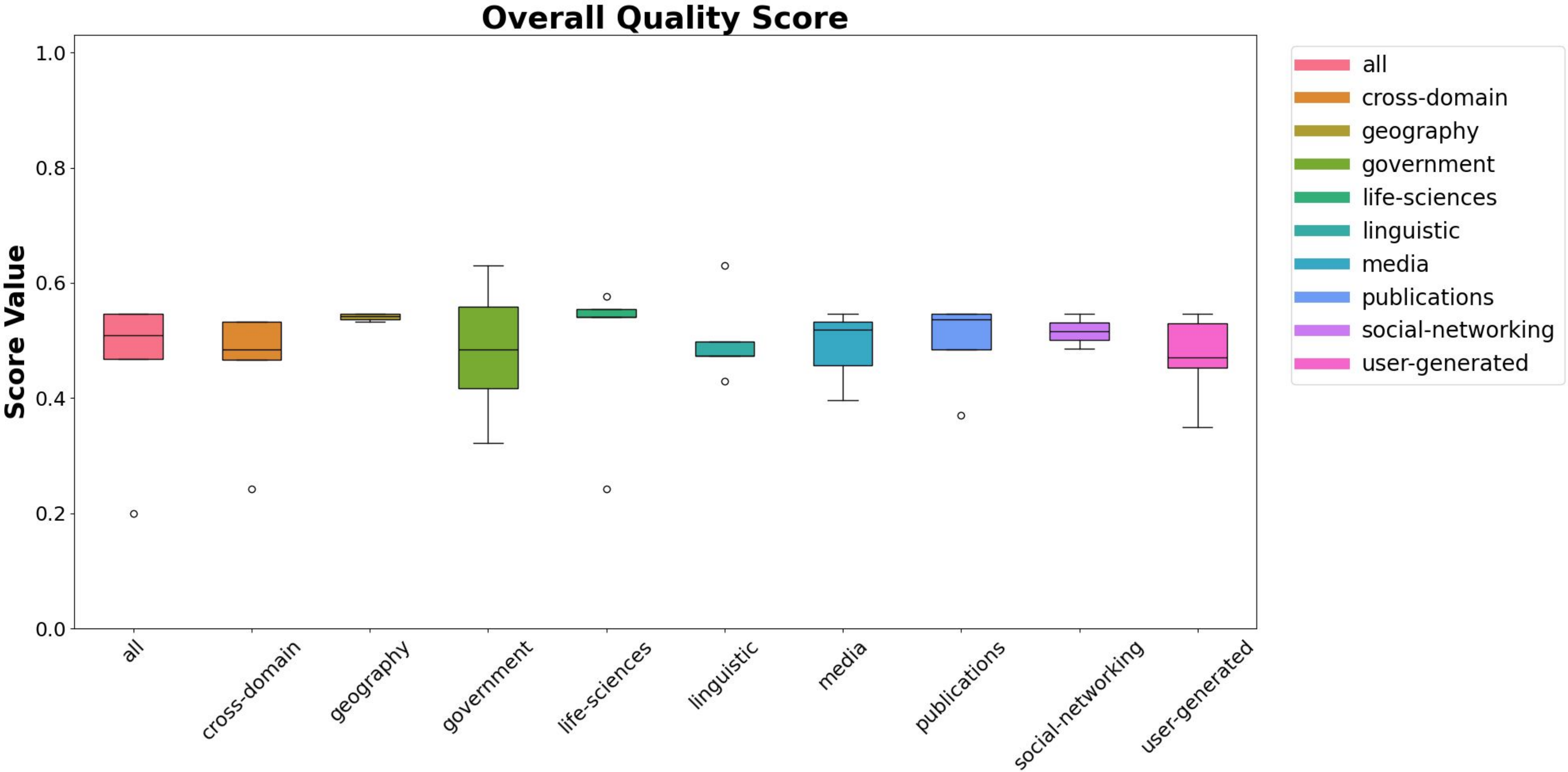
4. How have subclouds evolved over time?

Domain	**Machine-Readable License**			**VoID file availability**		
	[3]	[4]	*This*	[3]	[4]	*This*
Life-sciences	2.44%	3.61%	**24.72%**	7.32%	**36.14%**	1.38%
Media	16.00%	5.41%	***64.86%***	***20.00%***	*0.09%*	*8.10%*
User gen.content	20.00%	10.42%	***76.39%***	***25.00%***	*11.76%*	*1.28%*
Government	14.29%	30.05%	***48.72%***	***42.86%***	*42.08%*	*2.56%*
Cross-domain	19.51%	9.76%	***66.27%***	***21.95%***	*12.20%*	*9.63%*
Geographic	29.03%	0.00%	***68.09%***	***38.71%***	*38.10%*	*8.51%*
Publications	10.34%	4.17%	***48.99%***	***44.83%***	*13.54%*	*4.02%*
Social networking	-	5.38%	***8.25%***	-	*0.96%*	*1.03%*
Linguistic	-	-	*81.53%*	-	-	*7.63%*
Total	14.58%	9.96%	***49.19%***	***32.20%***	*14.69%*	*4.03%*

4. How have subclouds evolved over time?

Domain	**SPARQL endpoint**			**Data Dump**		
	[3]	[4]	*This*	[3]	[4]	*This*
Life-sciences	-	**24.10%**	11.66%	-	**15.66%**	15.00%
Media	-	0.00%	***8.10%***	-	*4.55%*	***29.72%***
User gen.content	-	6.25%	***6.94%***	-	*2.08%*	***19.33%***
Government	-	**31.15%**	*10.25%*	-	***31.15%***	*15.38%*
Cross-domain	-	4.88%	***18.07%***	-	*4.88%*	***28.91%***
Geographic	-	**14.29%**	*8.51%*	-	*19.05%*	***25.53%***
Publications	-	**12.50%**	*8.72%*	-	*4.17%*	***18.79%***
Social networking	-	0.77%	***2.06%***	-	*0.19%*	***5.15%***
Linguistic	-	-	*13.65%*	-	-	*56.22%*
Total	**68.14%**	9.96%	*10.70%*	***39.66%***	*8.19%*	*24.67%*

4. Holistic Quality Assessment of SubClouds



4. Holistic Quality Assessment of SubClouds

- **Accessibility:** Publications is the top performer due to high score in the *Availability* dimension; Government shows low median values and minimal variability.
- **Contextual**: Overall quality is low; Geography and Government perform slightly better, but this is the least maintained category.
- **Dataset Dynamicity**: Government shows slightly better performance than the entire LOD Cloud average. Media and Cross domain perform poorly due to the lack of update frequency metadata.
- **Intrinsic**: Geography, Life Sciences, and Media score above the entire LOD Cloud average. Geography leads in *Accuracy*, while Life Sciences excels in *Conciseness*. Social Networking performs worst.

4. Holistic Quality Assessment of SubClouds

- **Representational**: Linguistics leads in *Versatility* and *Interpretability*. User Generated ranks lowest, mainly due to poor *Versatility*.
- **Trust**: Media, Publications, and Government show the best *Believability* scores. User Generated performs worst, with very low *Verifiability* and *Believability*.

5. Discussion

Shift in Data Access Trends:

- While **SPARQL endpoint availability** remains a **concern** since 2014 [4], data **dump availability** has notably **increased**.
- Contrary to earlier findings, more dataset **now offer data dumps than SPARQL endpoints**, as also confirmed by Debattista et al. [7].

Licensing improvements:

- The **license metric** has shown significant improvement compared to previous assessments

Metadata Effort and Decline:

- The Government and Publications domains initially invested heavily in metadata (VoID files), but struggled to sustain this effort by 2024.

6. Conclusion

- Quality **varies notably** by subcloud (**RQ**), **no subcloud excels across all quality dimensions**.
- Life Sciences, Government, and Geography maintain consistently **good quality across most categories**.
- User Generated, Social Networking, and Cross domain are the **lowest performers**.
- As the data within the dataset **becomes more heterogeneous**, the overall quality **tends to decrease**, while the **domain-specific focus** enables **higher quality** through targeted curation.
- Therefore, **quality improvement efforts must be tailored to each domain**, as domain-specific factors play a crucial role and **uniform strategies are unlikely to be effective**.

6. Limitations and Future Works

Limitations:

- This study focuses on LOD Cloud subclouds, **excluding dataset from other aggregators** (e.g. DataHub, Zenodo, GitHub).
- **Unlabeled dataset** in the LOD Cloud are not considered.

Future works:

- Developing methods to **improve** subcloud quality.
- Proposing **interactive tools** to support diverse communities in curating heterogeneous data.
- Creating **domain-specific best practices** and **tailored manuals** to guide the dataset development and enhance standardization.

6. References

[1] Amrapali Zaveri, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann, and Soeren Auer. 2016. Quality assessment for linked data: A survey. Semantic Web 7, 1 (2016), 63–93. https://doi.org/10.3233/SW-150175.

[2] Maria Angela Pellegrino, Anisa Rula, and Gabriele Tuozzo. 2024. KGHeartBeat: An Open Source Tool for Periodically Evaluating the Quality of Knowledge Graphs. In International Semantic Web Conference. https://doi.org/10.1007/978-3-031-77847-6_3

[3] Cyganiak R. Bizer C. Jentzsch, A. 2011. State of the LOD cloud (September 2011). https://web.archive.org/web/20160323120153/lod-cloud.net/state/#structure.

[4] Max Schmachtenberg, Christian Bizer, and Heiko Paulheim. 2014. Adoption of the linked data best practices in different topical domains. (ISWC). https://doi.org/10.1007/978-3-319-11964-9_16

[5] Jeremy Debattista, Sören Auer, and Christoph Lange. 2016. Luzzu—a methodology and framework for linked data quality assessment. Journal of Data and Information Quality (JDIQ) 2016. https://doi.org/10.1145/2992786

[6] Ahmad Assaf, Aline Senart, and Raphaël Troncy. 2016. Towards an objective assessment framework for linked data quality: Enriching dataset profiles with quality indicators. (IJSWIS) 12, 3 (2016), 111–133. https://doi.org/10.4018/978-1-5225-5191-1.ch021

[7] Jeremy Debattista, Judie Attard, Rob Brennan, and Declan O'Sullivan. 2019. Is the LOD cloud at risk of becoming a museum for datasets? Looking ahead towards a fully collaborative and sustainable LOD cloud. In Companion Proceedings of The 2019 World Wide Web Conference. 850–858. https://doi.org/10.1145/3308560.33170

6. References

[8] Yasunori Yamamoto, Atsuko Yamaguchi, and Andrea Splendiani. 2018. YummyData: providing high-quality open life science data. Database 2018 (2018). https://doi.org/10.1093/database/bay022

[9] Pierre Maillot, Olivier Corby, Catherine Faron, Fabien Gandon, and Franck Michel. 2023. IndeGx: A model and a framework for indexing RDF knowledge graphs with SPARQL-based test suits. Journal of Web Semantics 76 (2023). https://doi.org/10.1016/j.websem.2023.100775

[10] Yusniel Hidalgo-Delgado, Yoan A López, Juan Pedro Febles Rodríguez, and Amed Leiva Mederos. 2021. Quality assessment of library linked data: a case study. In Iberoamerican Knowledge Graphs and Semantic Web Conference. https://doi.org/10.1007/978-3-030-91305-2_8

[11] Gustavo Candela, Pilar Escobar, Rafael C Carrasco, and Manuel Marco-Such. 2022. Evaluating the quality of linked open data in digital libraries. Journal of Information Science 48, 1 (2022), 21–43. https://doi.org/10.1177/0165551520930951

[12] Maria Pia di Buono, Hugo Gonçalo Oliveira, Verginica Barbu Mititelu, Blerina Spahiu, and Gennaro Nolano. 2022. Paving the way for enriched metadata of linguistic linked data. Semantic Web 13, 6 (2022), 1133–1157. https://doi.org/10.3233/SW-222994

[13] Pasquale Esposito, Maria Angela Pellegrino, Vittorio Scarano, and Gabriele Tuozzo. 2024. The Linguistic Linked Open Data Cloud: Phenomenal Cosmic Powers... Itty Bitty Quality Space!. In Proceedings of the ISWC 2024 Posters, Demos and Industry Tracks: From Novel Ideas to Industrial Practice co-located with 23nd ISWC, Vol. 3828. CEUR-WS.org. https://ceur-ws.org/Vol-3828/paper24.pdf

Thank you for your attention!

Any questions?

Code on GitHub

Gabriele Tuozzo

PhD student in Computer Science and Information Technology | Knowledge ...

gtuozzo@unisa.it