Supporting Next-Generation Science with a Semantic Ecosystem

Sabbir M. Rashid[1], John S. Erickson[1], Jamie P. McCusker[1], Henrique Santos[1], Paulo Pinheiro[2], Shruthi Chari[1], Matthew Johnson[1], Jade S. Franklin[1], Kelsey Rook[1] and Deborah L. McGuinness[1]

[1]*Rensselaer Polytechnic Institute, 110 8th St, Troy, NY 12180, USA*
[2]*Instituto Piaget, Av Jorge Peixinho 30, Almada, Portugal 2805-059*

Abstract

The evolving Tetherless World Constellation collection of resources contributes to a complete semantic ecosystem supporting data lifecycle components including dissemination, transformation, management, investigation, and visualization. The tools that form this semantic stack view metadata as first-class citizens. Data descriptions in the form of ontologies and knowledge graphs inform the logical operations inherent to the intelligent use and understanding of the data. This allows the use of data integration, inference, and analysis to explore and discover implicit knowledge already inherent in the data. These tools, ontologies, and methods can foster interest in semantic technologies and make the semantic web more accessible for those unaware of its potential for the creation of end-to-end intelligent applications.

Keywords

Semantic Ecosystem, Data, Ontologies, Knowledge Graphs, Transformation, Knowledge Management

1. Introduction

AI-enabled agents and applications are being utilized and developed more than ever, especially in use cases requiring high precision. In high-precision use cases such as law, finance, and healthcare, there emerge needs for provenance awareness, explainability, and domain knowledge representation. A semantic ecosystem such as ours—the Tetherless World Constellation (TWC) Semantic Ecosystem—allows data to be ingested, represented, and analyzed. Using a semantic ecosystem promotes reusability across projects, enables standardization between tools,

Posters, Demos, and Industry Tracks at ISWC 2024, November 13–15, 2024, Baltimore, USA
rashis3@rpi.edu (S. M. Rashid); erickj4@rpi.edu (J. S. Erickson); mccusj2@rpi.edu (J. P. McCusker); oliveh@rpi.edu (H. Santos); paulo.pinheiro@ipiaget.pt (P. Pinheiro); charis@rpi.edu (S. Chari); johnsm21@rpi.edu (M. Johnson); frankj6@rpi.edu (J. S. Franklin); rookk@rpi.edu (K. Rook); dlm@rpi.edu (D. L. McGuinness)
https://tw.rpi.edu/person/SabbirRashid (S. M. Rashid); https://tw.rpi.edu/person/JohnErickson (J. S. Erickson); https://tw.rpi.edu/person/JamieMcCusker (J. P. McCusker); https://tw.rpi.edu/person/HenriqueSantos (H. Santos); https://tw.rpi.edu/person/PauloPinheiro (P. Pinheiro); https://tw.rpi.edu/person/ShruthiChari (S. Chari); https://tw.rpi.edu/person/MattJohnson (M. Johnson); https://tw.rpi.edu/person/jade-franklin (J. S. Franklin); https://tw.rpi.edu/person/kelsey-rook (K. Rook); https://tw.rpi.edu/person/Deborah_L_McGuinness (D. L. McGuinness)
0000-0002-4162-8334 (S. M. Rashid); 0000-0003-3078-4566 (J. S. Erickson); 0000-0003-1085-6059 (J. P. McCusker); 0000-0002-2110-6416 (H. Santos); 0000-0001-8469-4043 (P. Pinheiro); 0000-0003-2946-7870 (S. Chari); 0000-0001-5212-8100 (M. Johnson); 0000-0002-1986-0989 (J. S. Franklin); 0000-0003-0471-7159 (K. Rook); 0000-0001-7037-4567 (D. L. McGuinness)

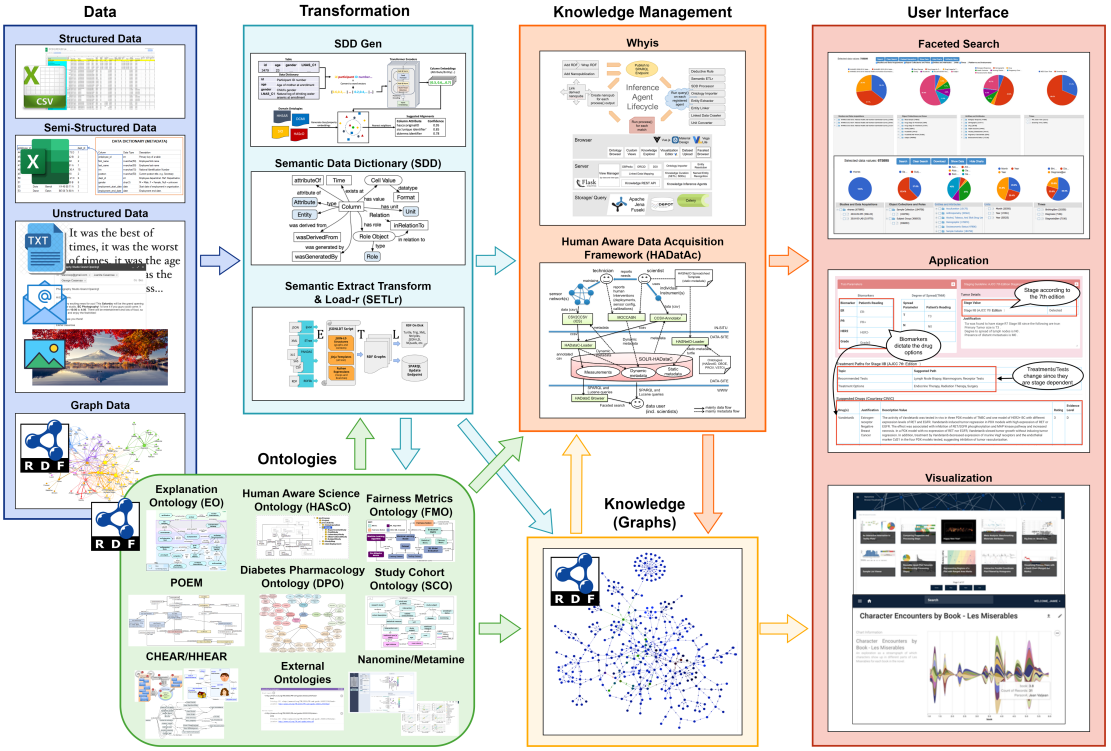

Figure 1: The Tetherless World Constellation Semantic Ecosystem. Using ontologies, data is transformed into knowledge graphs that power knowledge management frameworks. These frameworks are in turn used to create interactive applications.

and allows for the quick development of semantic applications when starting new projects. Furthermore, semantic ecosystems can help address some of the challenges associated with next-generation science.

We present the evolving TWC Semantic Ecosystem, which includes a collection of ontologies, tools, methods, and methodologies that support the data lifecycle. The interplay of the various components involved in this ecosystem is depicted in Fig. 1. The TWC Semantic Ecosystem supports the dissemination, transformation, management, investigation, and visualization of data. Data in many forms from various sources are transformed into a common graphical representation. The integrated data is used to identify implicit knowledge through the use of inference. This knowledge can power frameworks that manage the data to create interactive queries, applications, and visualizations.

This ecosystem supports knowledge representation and reasoning, semantic analysis, FAIRness [1], and explainability. Furthermore, our tools and ontologies have been used to conduct research in various domains including but not limited to materials science, epidemiology, health informatics, nutrition and clinical decision support, and music theory. Specifically, projects that have benefited from our ecosystem include Healthy Birth, Growth, and Development (HBGD) [2], Children's and Human Health Exposure Analysis Resource projects, CHEAR [3] and HHEAR [4], Nanomine [5], MaterialsMine (MM) [6], Health Empowerment by Analytics,

Learning, and Semantics (HEALS) [7, 8], Automated clusteRing Curriculum LearnIng Guided by Human Training (ARCLIGHT) [9], Revised Child Anxiety and Depression Scale (RCADS) [10], and Human Interpretable Attribution of Text using Underlying Structure (HIATUS) [11].

2. Resources

The TWC Semantic Ecosystem provides resources for processing data in various forms, including semantic annotations, creating knowledge graphs, and using the graphical representation for powering applications.

2.1. Ontologies

Ontologies can be viewed as a form of graph data often written using the OWL language containing a collection of concepts and properties, typically used to describe a particular domain [12]. Research conducted at the TWC has resulted in the production of many ontologies — both general-purpose and domain ontologies, such as the Human-Aware Science Ontology (HAScO) [13], the Fairness Metrics Ontology (FMO) [14], the Explanation Ontology (EO) [15, 16], the Psychometric Ontology of Experiences and Measures (POEM) [17], the Diabetes Pharmacology Ontology (DPO) [18], the Study Cohort Ontology (SCO) [19], the Children's Health Exposure Analysis Resource (CHEAR) ontology [20], the HHEAR ontology [4], the NanoMine Ontology [5], and the MaterialsMine (MM) ontology [6], to name a few.

2.2. Knowledge Graphs and Data Transformation

A knowledge graph (KG) is a graphical representation of information that encodes the concepts associated with entities and the relationships between entities. For the TWC Semantic Ecosystem, information from various data sources is integrated into a combined KG. Transformation approaches include the Semantic Data Dictionary (SDD [21], the Semantic Data Dictionary Generator (SDD-Gen) [22], and Semantic Extract, Transform, and Load-er (SETLr) [23].

The SDD allows for the creation of semantic annotations for columns in a data set, categorical or coded cell values, and intrinsic concepts implicit in the data [21]. SDD-Gen is a semantic tabular interpretation algorithm that uses context information from data dictionary descriptions to align tabular concepts to ontology terms [22]. SETLr provides an approach for converting structured and semi-structured data into RDF [23].

2.3. Knowledge Management

Knowledge management refers to an approach for capturing, storing, utilizing, and analyzing information. Knowledge management frameworks provide a governing system for conducting knowledge management. The TWC Semantic Ecosystem has two knowledge management frameworks, Whyis [24] and the Human-Aware Data Acquisition (HADatAc) [25].

Whyis is a nano-scale knowledge graph publishing, management, and analysis framework that supports the open-ended development, management, and curation of knowledge from many different sources [24]. HADatAc is an infrastructure for integrating data and metadata

from multiple scientific studies to promote scalability, provenance-awareness, and freedom from schema restrictions [25].

A user interface (UI) is required for an end-user to interact with the underlying data behind the framework without necessarily conducting direct database operations. A semantic ecosystem supports the incorporation of UI components and in turn allows for the construction of faceted search browsers, applications, and visualizations.

3. Conclusion

All of the resources discussed in this paper have resulted from research at the TWC and have been published openly by adhering to the FAIR guiding principles [1]. Therefore, each component included in the semantic ecosystem described is open-sourced and available for public use. These resources form key contributions to semantic web research as they allow others to follow an end-to-end workflow as outlined in our semantic ecosystem. For more information on these tools, visit https://tw.rpi.edu/tools.

The tools, methods, and ontologies developed at the TWC have been used by various organizations to design end-to-end intelligent applications that leverage semantic web technologies. Using these resources can result in technological advancements and innovations since they enable inference, analysis, and visualization. Data and ontologies combine to form knowledge graphs that allow for logical operations inherent to the intelligent use and understanding of the data. With the aid of knowledge management frameworks, exciting new applications can be implemented that cater to the interests of the end-user. This research supports next-generation science by highlighting how semantic technologies can be used to enhance the data lifecycle.

Acknowledgements

We acknowledge all of the members of the Tetherless World Constellation for their feedback and contributions to this research. This work is partially supported by the following projects:

- NIEHS-funded *Children's Health Exposure Analysis Resource (CHEAR)*, project number **1U2CES026555-01**
- NIEHS-funded *Human Health Exposure Analysis Resource (HHEAR)*, project number **5U2CES026555-05**
- NIMH-funded support for the *RCADS Data Collection Measure*, project number **75N95022C00018-0-9999-1**
- IBM-funded *Health Empowerment by Analytics, Learning, and Semantics (HEALS)* through the AI Horizons Network program
- DARPA-funded *Machine Common Sense (MCS)*, grant number **N660011924033**
- DARPA-funded *Environment-driven Conceptual Learning (ECOLE)*, grant number **HR00112390059**
- IARPA-funded *Human Interpretable Attribution of Text using Underlying Structure (HIATUS)*, grant number **2022-22072200002**
- NSF-funded *Nanomine*, award number **1640840**
- NSF-funded *MaterialsMine*, award number **1835648**

References

[1] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al., The fair guiding principles for scientific data management and stewardship, Scientific data 3 (2016) 1–9.

[2] R. P. Institute, Rensselaer polytechnic institute launches initiative on healthy birth, growth, and development knowledge: Semantic and data analytic support, https://tw.rpi.edu/media/rensselaer-polytechnic-institute-launches-initiative-healthy-birth-growth-and-development, 2015. Accessed: 2024-08-04.

[3] J. Stingone, P. Pinheiro, J. Meola, J. McCusker, S. Bengoa, P. Kovatch, D. McGuinness, S. Teitelbaum, et al., The chear data repository: Facilitating children's environmental health and exposome research through data harmonization, pooling and accessibility, Environmental Epidemiology 3 (2019) 382.

[4] Human health exposure analysis resource, https://bioportal.bioontology.org/ontologies/HHEAR, 2024. Accessed: 2024-04-04.

[5] J. P. McCusker, N. Keshan, S. Rashid, M. Deagen, C. Brinson, D. L. McGuinness, Nanomine: A knowledge graph for nanocomposite materials science, in: International Semantic Web Conference, Springer, 2020, pp. 144–159.

[6] M. E. Deagen, J. P. McCusker, T. Fateye, S. Stouffer, L. C. Brinson, D. L. McGuinness, L. S. Schadler, Fair and interactive data graphics from a scientific knowledge graph, Scientific Data 9 (2022) 239.

[7] R. P. Institute, Health empowerment by analytics, learning, and semantics, https://idea.rpi.edu/research/projects/heals, 2022. Accessed: 2024-04-04.

[8] O. Seneviratne, A. K. Das, S. Chari, N. N. Agu, S. M. Rashid, J. McCusker, J. S. Franklin, M. Qi, K. P. Bennett, C.-H. Chen, et al., Semantically enabling clinical decision support recommendations, Journal of Biomedical Semantics 14 (2023) 8.

[9] B. Fusion, Boston fusion awarded darpa ecole contract, https://bostonfusion.com/boston-fusion-darpa-ecole/, 2023. Accessed: 2024-07-01.

[10] H. Santos, K. Rook, P. Pinheiro, D. M. Gruen, B. F. Chorpita, D. L. McGuinness, Facilitating reuse of mental health questionnaires via knowledge graphs, The Healthcare and Life Sciences Symposium, 2023.

[11] IARPA, Human interpretable attribution of text using underlying structure, https://www.iarpa.gov/research-programs/hiatus, 2022. Accessed: 2024-04-05.

[12] E. F. Kendall, D. L. McGuinness, Ontology engineering, Morgan & Claypool Publishers, 2019.

[13] P. Pinheiro, M. Bax, H. Santos, S. M. Rashid, Z. Liang, Y. Liu, J. P. McCusker, D. L. McGuinness, Annotating Diverse Scientific Data with HAScO, in: Proceedings of the Seminar on Ontology Research in Brazil 2018 (ONTOBRAS 2018). São Paulo, SP, Brazil, 2018, pp. 80–91.

[14] J. S. Franklin, K. Bhanot, M. Ghalwash, K. P. Bennett, J. McCusker, D. L. McGuinness, An ontology for fairness metrics, in: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, 2022, pp. 265–275.

[15] S. Chari, O. Seneviratne, D. M. Gruen, M. A. Foreman, A. K. Das, D. L. McGuinness, Explanation ontology: a model of explanations for user-centered ai, in: International

Semantic Web Conference, Springer, 2020, pp. 228–243.
[16] S. Chari, O. Seneviratne, M. Ghalwash, S. Shirai, D. M. Gruen, P. Meyer, P. Chakraborty, D. L. McGuinness, Explanation ontology: A general-purpose, semantic representation for supporting user-centered explanations, Semantic Web (2023) 1–31.
[17] K. Rook, H. Santos, B. F. Chorpita, M. S. Sprung, P. Pinheiro, D. L. McGuinness, Towards an Ontology of Psychometric Measures, 2024.
[18] S. M. Rashid, J. McCusker, D. Gruen, O. Seneviratne, D. L. McGuinness, A concise ontology to support research on complex, multimodal clinical reasoning, in: European Semantic Web Conference, Springer, 2023, pp. 390–407.
[19] S. Chari, M. Qi, N. N. Agu, O. Seneviratne, J. P. McCusker, K. P. Bennett, A. K. Das, D. L. McGuinness, Making study populations visible through knowledge graphs, in: International Semantic Web Conference, Springer, 2019, pp. 53–68.
[20] J. P. McCusker, S. M. Rashid, Z. Liang, Y. Liu, K. Chastain, P. Pinheiro, J. A. Stingone, D. L. McGuinness, Broad, interdisciplinary science in tela: An exposure and child health ontology, in: Proceedings of the 2017 ACM on Web Science Conference, 2017, pp. 349–357.
[21] S. M. Rashid, J. P. McCusker, P. Pinheiro, M. P. Bax, H. O. Santos, J. A. Stingone, A. K. Das, D. L. McGuinness, The semantic data dictionary–an approach for describing and annotating data, Data intelligence 2 (2020) 443–486.
[22] M. Johnson, J. A. Stingone, S. Bengoa, J. Masters, D. L. McGuinness, Complex semantic tabular interpretation using sdd-gen, in: 2024 IEEE 18th International Conference on Semantic Computing (ICSC), IEEE, 2024, pp. 317–322.
[23] J. P. McCusker, K. Chastain, S. Rashid, S. Norris, D. L. McGuinness, Setlr: the semantic extract, transform, and load-r, PeerJ Preprints 6 (2018) e26476v1.
[24] J. P. McCusker, S. M. Rashid, N. Agu, K. P. Bennett, D. L. McGuinness, The whyis knowledge graph framework in action., in: ISWC (P&D/Industry/BlueSky), 2018.
[25] P. Pinheiro, H. Santos, Z. Liang, Y. Liu, S. M. Rashid, D. L. McGuinness, M. P. Bax, HADatAc: A framework for scientific data integration using ontologies, in: Proceedings of the ISWC, 2018, p. 49.