Rensselaer

why not change the world?®

Tetherless World Constellation

SciKG: Tutorial on Building Scientific Knowledge Graphs from Data, Data Dictionaries, and Codebooks

Henrique Santos, Paulo Pinheiro, Jamie P. McCusker, Sabbir M. Rashid, Deborah L. McGuinness

May 28th 2023

The 20th Extended Semantic Web Conference (ESWC-23)

Tutorial Organizers

Henrique Santos, Ph.D.
Director, Semantic Applications Research
Rensselaer Polytechnic Institute

Paulo Pinheiro, Ph.D.
Executive Director
Parcela Semântica Lda

Jamie P. McCusker, Ph.D.
Director, Data Operations
Rensselaer Polytechnic Institute

Sabbier M. Rashid, Ph.D.
Rensselaer Polytechnic Institute

Deborah L. McGuinness, Ph.D.
Tetherless World Senior Constellation Chair
Rensselaer Polytechnic Institute

Acknowledgments

- NIEHS Children’s Health Exposure Analysis Resource (CHEAR), project number 1U2CES026555-01
- NIEHS Human Health Exposure Analysis Resource (HHEAR), project number 5U2CES026555-05
- IBM Health Empowerment by Analytics, Learning, and Semantics (HEALS) through the AI Horizons Network program
- DARPA Machine Common Sense (MCS), grant number N660011924033
- IARPA Human Interpretable Attribution of Text using Underlying Structure (HIATUS)
- NSF Nanomine, award number 1640840
- NSF MaterialsMine, award number 1835648

Publications that documents parts of this tutorial content

- Deagen, M. E., McCusker, J. P., Fateye, T., Stouffer, S., Brinson, L. C., McGuinness, D. L., & Schadler, L. S. (2022). FAIR and Interactive Data Graphics from a Scientific Knowledge Graph. Scientific Data, 9(1), Article 1. https://doi.org/10.1038/s41597-022-01352-z
- Dumontier, M., Baker, C. J., Baran, J., Callahan, A., Chepelev, L., Cruz-Toledo, J., Del Rio, N. R., Duck, G., Furlong, L. I., Keath, N., Klassen, D., McCusker, J. P., Queralt-Rosinach, N., Samwald, M., Villanueva-Rosales, N., Wilkinson, M. D., & Hoehndorf, R. (2014). The Semanticscience Integrated Ontology (SIO) for biomedical research and knowledge discovery. Journal of Biomedical Semantics, 5, 14. https://doi.org/10.1186/2041-1480-5-14
- J, S., P, P., J, M., J, M., S, B., P, K., D, M., & S, T. (2019). The CHEAR Data Repository: Facilitating children's environmental health and exposome research through data harmonization, pooling and accessibility. Environmental Epidemiology, 3, 382. https://doi.org/10.1097/01.EE9.0000610256.39316.c4
- McCusker, J., & McGuinness, D. L. (2023). Whyis 2: An Open Source Framework for Knowledge Graph Development and Research. In C. Pesquita, E. Jimenez-Ruiz, J. McCusker, D. Faria, M. Dragoni, A. Dimou, R. Troncy, & S. Hertling (Eds.), The Semantic Web (pp. 538–554). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-33455-9_32
- McCusker, J., McIntosh, L. D., Shaffer, C., Boisvert, P., Ryan, J., Navale, V., Topaloglu, U., & Richesson, R. L. (n.d.). Guiding principles for technical infrastructure to support computable biomedical knowledge. Learning Health Systems, n/a(n/a), e10352. https://doi.org/10.1002/lrh2.10352

Publications that documents parts of this tutorial content

- McCusker, J. P., Keshan, N., Rashid, S., Deagen, M., Brinson, C., & McGuinness, D. L. (2020). NanoMine: A Knowledge Graph for Nanocomposite Materials Science. In J. Z. Pan, V. Tamma, C. d'Amato, K. Janowicz, B. Fu, A. Polleres, O. Seneviratne, & L. Kagal (Eds.), The Semantic Web – ISWC 2020 (pp. 144–159). Springer International Publishing. https://doi.org/10.1007/978-3-030-62466-8_10
- McCusker, J. P., Rashid, S. M., Liang, Z., Liu, Y., Chastain, K., Pinheiro, P., Stingone, J. A., & McGuinness, D. L. (2017). Broad, Interdisciplinary Science In Tela: An Exposure and Child Health Ontology. Proceedings of the 2017 ACM on Web Science Conference, 349–357. https://doi.org/10.1145/3091478.3091497
- McCusker, J., Rashid, S. M., Agu, N., Bennett, K. P., & McGuinness, D. L. (2018a). Developing Scientific Knowledge Graphs Using Whyis. SemSci@ ISWC, 52–58.
- McCusker, J., Rashid, S. M., Agu, N., Bennett, K. P., & McGuinness, D. L. (2018b). The Whyis Knowledge Graph Framework in Action. International Semantic Web Conference (P&D/Industry/BlueSky).
- McGuinness, D. L., Pinheiro, P., Santos, H., Klawonn, M., & Chastain, K. (2015). Semantic Support for Complex Ecosystem Research Environments. AGU Fall Meeting Abstracts, 33. http://adsabs.harvard.edu/abs/2015AGUFMIN33F..02K
- Pinheiro, P., Bax, M., Santos, H., Rashid, S. M., Liang, Z., Liu, Y., McCusker, J. P., & McGuinness, D. L. (2018). Annotating Diverse Scientific Data with HAScO. Proceedings of the Seminar on Ontology Research in Brazil 2018 (ONTOBRAS 2018). São Paulo, SP, Brazil.
- Pinheiro, P., McGuinness, D. L., & Santos, H. (2015, October). Human-Aware Sensor Network Ontology: Semantic Support for Empirical Data Collection. Proceedings of the 5th Workshop on Linked Science. Bethlehem, PA, USA.

Publications that documents parts of this tutorial content

- Pinheiro, P., Santos, H., Liang, Z., Liu, Y., Rashid, S. M., McGuinness, D. L., & Bax, M. P. (2018). HADatAc: A Framework for Scientific Data Integration using Ontologies. Proceedings of the ISWC Posters & Demonstrations Track.
- Rashid, S. M., McCusker, J. P., Pinheiro, P., Bax, M. P., Santos, H., Stingone, J. A., Das, A. K., & McGuinness, D. L. (2020). The Semantic Data Dictionary – An Approach for Describing and Annotating Data. Data Intelligence, 2(4), 443–486. https://doi.org/10.1162/dint_a_00058
- Santos, H., Dantas, V., Furtado, V., Pinheiro, P., & McGuinness, D. L. (2017). From Data to City Indicators: A Knowledge Graph for Supporting Automatic Generation of Dashboards. The Semantic Web, 94–108. https://doi.org/10.1007/978-3-319-58451-5_7
- Santos, H., Furtado, V., Pinheiro, P., & McGuinness, D. L. (2015, October). Contextual Data Collection for Smart Cities. Proceedings of the Sixth Workshop on Semantics for Smarter Cities. Sixth Workshop on Semantics for Smarter Cities, Bethlehem, PA, USA.
- Santos, H., Pinheiro, P., & McGuinness, D. L. (2022, September). Knowledge Graph Construction from Data, Data Dictionaries, and Codebooks: The National Health and Nutrition Examination Surveys Use Case. 4th U.S. Semantic Technologies Symposium, Michigan State University, East Lansing, MI. https://us2ts.org

Past KG Building tutorials

- Knowledge Graph Construction @ESWC-22
- Tools for Creating and Exploiting Large Knowledge Graphs (KGTK) @ISWC-21
- Knowledge Graph Construction using Declarative Mapping Rules @ISWC-20
- How to build large knowledge graphs efficiently (LKGT) @ISWC-20
- Generating and querying (Virtual) Knowledge Graphs from heterogeneous data sources @ESWC-19
- Building Enterprise-Ready Knowledge Graph Applications in the Cloud (EKG) @ISWC-19

How is SciKG different?

- We will be working with scientific data collected in the context of scientific studies
- Preservation of contextual knowledge that is often lost after data acquisition
- Quality matters
- Standards matter
- Semantic rigour
- Dependable support for data analysis

Requirements

- Not a lot!
- Basic notion of the Semantic Web and RDF
  - Resources
  - Vocabularies
  - Ontologies
- If you want to try on your own:
  - Spreadsheet editor
  - Git
  - Docker
  - And some patience…

What is your motivation?

Agenda

9:00 - 10:30	Part 1: Studies, Data, and Documentation
10:30 - 11:00	Break
11:00 - 12:30	Part 2: Scientific and Biomedical Ontologies
12:30 - 14:00	Lunch
14:00 - 15:30	Part 3: Semantic Data Dictionaries
15:30 - 16:00	Break
16:00 - 18:00	Part 4: Knowledge Graph Frameworks

Part 1: Studies, Data, and Documentation

- Scientific studies and their data acquisition activities
- Scientific data organization
- Scientific data publishing
- Documentation
  - Data dictionaries
  - Codebooks
  - Methods
- National Health and Nutrition Examination Surveys (NHANES)
  - Semantics of NHANES data

Part 2: Scientific and Biomedical Ontologies

- The role of standardized terminology in science
- Semanticscience Integrated Ontology (SIO)
- Human-Aware Science Ontology (HAScO)
- Disease Ontology (DOID)
- Chemical Elements of Biological Interest (ChEBI)

Part 3: Semantic Data Dictionaries

- Introduction to Semantic Data Dictionaries (SDDs)
  - Structure
  - Examples
- Practical section with NHANES datasets

Part 4: Knowledge Graph Frameworks

- The Human-Aware Data Acquisition Framework (HADatAc)
- Whyis
- Using SDDs to bootstrap Knowledge graphs
- Navigating, querying and using the KG

Logistics and other Information

- This is an interactive tutorial
- Please interrupt at any point