Summer 8-6-2024

Knowledge Management of Historical Data: Ontology Development for Chemical Reactions

Quynh D. Tran
Case Western Reserve University, qdt@case.edu

Alexander Harding Bradley
Case Western Reserve University, ach159@case.edu

Balashanmuga Priyan Rajamohan
Case Western Reserve University, bxr261@case.edu

Jonathan E. Gordon
Case Western Reserve University, jeg165@case.edu

Van D. Tran
Case Western Reserve University, vxt101@case.edu

Authors

Quynh D. Tran, Alexander Harding Bradley, Balashanmuga Priyan Rajamohan, Jonathan E. Gordon, Van D. Tran, Kiefer Lin, Erika I. Barcelos, Laura S. Bruckman, and Roger H. French

Materials Data Science for Stockpile Stewardship

COE: US-Department of Energy-NNSA Award

MDS3 CENTER OF EXCELLENCE

Lawrence Livermore National Laboratory · Los Alamos NATIONAL LABORATORY EST.1943 · Sandia National Laboratories · KANSAS CITY NSC National Security Campus · The Nuclear Security Enterprise NNSA

CASE WESTERN RESERVE UNIVERSITY EST. 1826 — think beyond the possible™

Knowledge Management of Historical Data: Ontology Development for Chemical Reactions

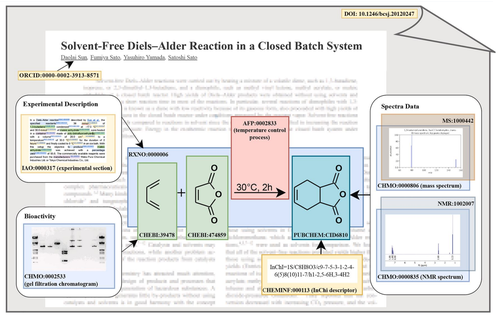
Quynh D. Tran[1,2], Alexander Harding Bradley[1,3], Balashanmuga Priyan Rajamohan[1,3], Jonathan Gordon[4], Van Tran[1,2], Kiefer Lin[1,3], Erika I. Barcelos[1,2], Laura S. Bruckman[1,2], Roger H. French[1,2,3]

[1]Material Data Science for Stockpile Stewardship Center of Excellence, Case Western Reserve University, Cleveland OH 44106 USA
[2]Department of Materials Science and Engineering, Case Western Reserve University, Cleveland OH 44106 USA
[3] Department of Computer Sciences, Case Western Reserve University, Cleveland OH 44106 USA
[4] Department of Physics and Astronomy, Bowdoin College, Brunswick, ME 04011, USA

1. INTRODUCTION

LITERATURE KNOWLEDGE

- **Figures**:
  - Chemical structures and equations
  - Results (NMR, FTIR, UV-Vis, DSC, mechanical, etc.)
- Detailed reaction conditions and setup in **written text**
- Raw data aren't available
- Analyses and data transformations aren't available
- Articles saved in **arbitrary folders** on local laptops



Knowledge is scattered in articles on the web/local computer. How to integrate and manage knowledge efficiently?

EXPERIMENTAL DATA

Samples prepared with user-defined sample IDs → Instrument parameters (recipes) setup & sample analyzed → Metadata saved in notebooks or computer software → Raw data saved as .txt or .xls on local computer/software → Data analyzed containing user-defined variable names → Results published & stacked away

METADATA

- User-defined:
  - Sample & experiment IDs
  - Variable names
  - File names + Folder structures
- Not all metadata needed to reproduce experiments is recorded
- Metadata cannot be found by computers

DATA

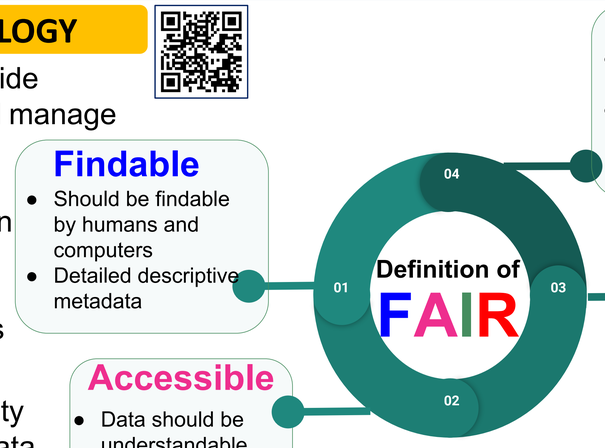
- Data get lost over time
- "Everything" is saved in arbitrary folders on local or instrument computers
- **Integration with other labs/fields is challenging because user-defined variables**
  - Only make sense to the working researcher(s)
- Data and/or required licenses cannot be found by computers

Metadata and data in various formats sources. How to integrate and manage data efficiently?

2. METHOD

FAIR AND ONTOLOGY

- **FAIR principles**[2] provide guidelines to collect and manage metadata.
- **Ontology**
  - a formal representation of knowledge;
  - composes of a set of standardized concepts and definitions;
  - Increase interoperability
  - Organize metadata, data, and knowledge in a meaningful and efficient manner.



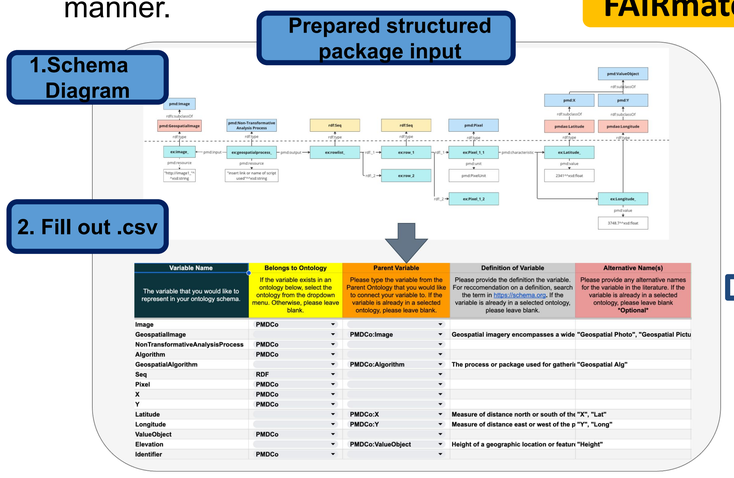
Findable
- Should be findable by humans and computers
- Detailed descriptive metadata

Accessible
- Data should be understandable
- Data should also be in a repository

Interoperable
- Should be integrable with other data
- **Ontologies** tie data to a standardized vocabulary

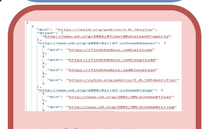
Reusable
- Clear and accessible data usage license
- Data and metadata meet community standards

FAIRmaterials Package[3][4]



An image for users to visualize the ontology schema & ensure its correctness

An OWL File with the ontology schema and populated ontology data



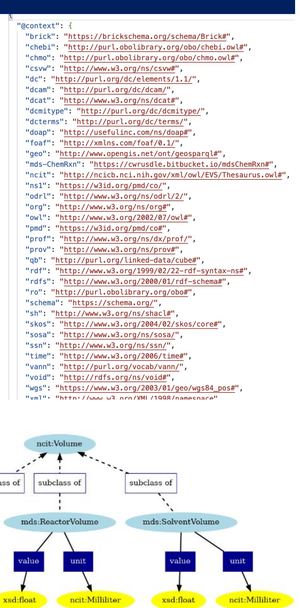
A JSON-LD File with the ontology schema and populated ontology data.

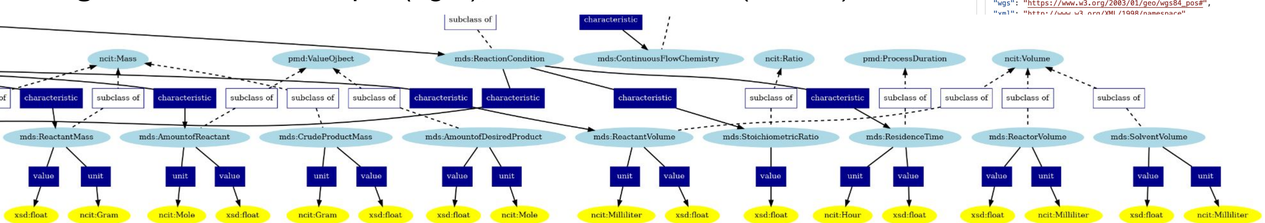
An HTML File with the documentation for the output ontology.

- FAIRmaterials package provides ***an easy to use csv template*** for domain experts to construct their own ontologies mapped from any existing ontologies
- Must connect with a top-level ontology to be interoperable
- Some chemical mid-level ontologies:
  - ***ChEBI, NCIt, CHMO, RXNO***
- The output files (PNG, OWL, JSON-LD, and HTML) allow for better visualization, term definitions and relationships, data population, and documentation, respectively.

3. RESULTS

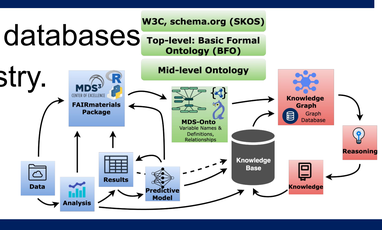
- Investigated mid-level chemical ontologies
  - Must map to Basic Formal Ontology (BFO) as the ISO standard[5]
- Developed an ontology for chemical reaction in nitration of aromatic compounds using continuous flow chemistry
  - A small, concentrated subdomain to ensure the ontology is less overwhelming
  - Specific variable names such as "reactorTemperature" and "reactorFlowRate" instead of just "Temperature" or "flowRate".
- **Figures:** JSON-LD output (right); PNG visualization (bottom).





4. FUTURE WORK

- Build and populate a knowledge base with chemical open-source databases on nitration of aromatic compounds using continuous flow chemistry.
- Equip the knowledge base with a reasoning engine for semantic reasoning.
- Use FAIR and ontology as a backbone of data governance.



5. REFERENCES

1. Strömert, Philip, Hunold, Johannes, Castro, André, Neumann, Steffen and Koepler, Oliver. "Ontologies4Chem: the landscape of ontologies in chemistry" *Pure and Applied Chemistry*, vol. 94, no. 6, 2022, pp. 605-622. https://doi.org/10.1515/pac-2021-2007
2. Jacobsen, A., de Miranda Azevedo, R., Juty, N., Batista, D., Coles, S., Cornet, R., Courtot, M., Crosas, M., Dumontier, M., Evelo, C. T., Goble, C., Guizzardi, G., Hansen, K. K., Hasnain, A., Hettne, K., Heringa, J., Hooft, R W. W., Imming, M., Jeffery, K. G., … Schultes, E. (2020). FAIR Principles: Interpretations and Implementation Considerations. *Data Intelligence*, *2*(1–2), 10–29. https://doi.org/10.1162/dint_r_00024
3. Alexander Harding Bradley, Jonathan E. Gordon, Priyan Rajamohan, Nathaniel Hahn, Kiefer Lin, Arafath Nihar, Hayden Cadwell, Jiana Kambo, Jayvic Jimenez, Kristen J. Hernandez, Hein Htet Aung, Brian Giera, Weiqi Yu, Mohommad Redad Mehdi, Finley Holt, Quynh Tran, Gabriel Ponon, Dan Savage, Don Brown, … Roger H. French. (2024). *FAIRmaterials: Python* (0.4.2.6) [Python; Cross-platform]. SDLE Res. Cntr., Case Western Reserve University. https://pypi.org/project/fairmaterials/
4. Jonathan E. Gordon, Alexander Harding Bradley, Priyan Rajamohan, Nathaniel Hahn, Kiefer Lin, Arafath Nihar, Hayden Cadwell, Jiana Kambo, Jayvic Jimenez, Kristen J. Hernandez, Hein Htet Aung, Brian Giera, Weiqi Yu, Mohommad Redad Mehdi, Finley Holt, Quynh Tran, Gabriel Ponon, Dan Savage, Don Brown, … Will Oltjen. (2024). *FAIRmaterials: R* (0.4.2.1) [R]. SDLE Res. Cntr., Case Western Reserve University. https://cran.case.edu/web/packages/FAIRmaterials/index.html
5. ISO/IEC, "ISO/IEC 21838-2 Ed. 1: 2021 | Information technology - Top-level ontologies (TLO) - Part 2: Basic Formal Ontology (BFO)," ISO 21838-2 Ed. 1: 2021, Nov. 2021. Available:. [Accessed: Nov. 12, 2022]

6. ACKNOWLEDGEMENT

This material is based upon research in the Materials Data Science for Stockpile Stewardship Center of Excellence (MDS3-COE), and supported by the U.S. Department of Energy's National Nuclear Security Administration under Award Number(s) DE-NA0004104. Work supported by the Laboratory Directed Research and Development program at Sandia National Laboratories, a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This work made use of the High Performance Computing Resource in the Core Facility for Advanced Research Computing at Case Western Reserve University.