Collaborative Development of a Process Chemistry Domain Ontology, PROCO

Wes Schafer[1], Vincent Antonucci[1], Yongqun Oliver He[2] , Anna Dunn[3], Zach E.X. Dance[3], Jan Nespor[4] and Lama Saeeda[4]

[1] Research and Development Sciences IT, Merck & Co., Inc., Rahway, NJ, USA
[2] University of Michigan Medical School, Ann Arbor, MI, USA.
[3] Analytical Research & Development, Merck & Co., Inc., Rahway, NJ, USA
[4] IT Eng., Dev. & Integration, MSD, Prague

Abstract

Process chemists embracing data mining, machine learning and artificial intelligence rapidly discover that the lack of structured data hampers their efforts. Although raw and processed instrument data have been addressed with public ontologies such as Allotrope and somewhat by commercial enterprise content management solutions, the chemical context of that data has been largely neglected. Recognizing this foundational gap and the impact it would have on developing new and better scientific data capture systems like electronic laboratory notebooks, Merck chemists reached out to peers in other companies and academia to develop PROCO, a domain ontology around process chemistry that studies the development and optimization of the production processes for chemical compounds. The scope was set using specific use cases and is being rounded out modeling public databases such as ORD.

Development was based on "up-scaling" the chemists semantic skill sets and the domain knowledge of the ontologists. Simplified public ontology tools such as WebProtege provided a collaborative online space for ontology developers and subject matter experts with features like commenting, suggesting and approving changes. To increase internal adoption and applicability of PROCO, the ontology was loaded into Merck's master ontology for discovery, pre-clinical and early development space (MDO) via its CENtree ontology management system. Specific applications use MDO as the master source to develop their 'application ontologies'. As an example, ELN application ontology covers experimental metadata and feeds it into the Perkin-Elmer Signals notebook to define values of drop-down lists in the notebook. This enables standardized data capture which consequently makes the data interoperable and reusable for analytics / data science. Using ontologies as a metadata input for data capture enables data to be 'born FAIR' (findable, accessible, interoperable, and reusable) which is significantly more efficient and less expensive than FAIRifying the data at later stages.

This industrial and academic collaboration proved to be an effective means of achieving better structured data with limited enterprise resources. The final PROCO ontology has been submitted to the OBO Foundry to broaden the development pool and usage of the ontology.

Keywords
PROCO, Process Chemistry Ontology, process chemistry, ontology.

ICBO 2022, September 25-28, 2022, Ann Arbor, MI, USA
EMAIL: yongqunh@med.umich.edu (A. 1)
ORCID: 0000-0001-9189-9661 (A. 1)