ver 2021-09-29

OMDI2021

Provided abstracts

(To provide talk abstracts has been optional.
Hence, they are only available for a few talks)

Tuesday, Oct 5 at 9:05am CEST

Scientific Ontologies in the Digital Age

Peter Murray-Rust,
Yusuf Hamied Department of Chemistry, University of Cambridge

This will be a pragmatic exploration of science-based ontologies in the digital, computational, age and the aspects that helps them flourish. I believe that most scientists appreciate the I see ontologies as formalised systems for describing objects and processes, their constitution, their interrelationships, their constraints and their functionality/computability. At a simple level ontologies can be just lists of formal terms (labels) for things. At a complex level they are computable objects that merge into executable code. The good news is that modern software tools (such as the web, Python and its libraries) now make it much easier for scientists to get involved in using and creating ontologies. We are at the start of the golden age of universal digital ontologies and they are embedded in many of the systems we use.

Many of us are ontologists at heart, and love systematising things. At a lexical level Digital Ontologies help with learning and understanding, and translation between different representations (synonymy, ambiguity, identifer systems, human languages). At a semantic level things in science have identifiable properties (categorical, numeric, or more complex) that can be formaliy described, and computed, by ontologies. Taken even further, structured data + ontologies represent a computable model of a scientific experiment or dataset.

Much of our data is not explicitly structured or semantic. A FORTRAN logfile, or a PDF article contain snippets of data which has lost its labels, its internal structure, and its relation to other data. It is extremely tedious, and error-prone to try to recreate semantic data. It could, in large part, be created as born-semantic but this requires the authors (of code or articles) to recognize the value of doing so, and put in the extra effort. This occasionally happens (Acta Crystallographica requires manuscripts to be semantic CIFs) but normally the community (learned societies and publishers) do not see the value, or even see it as harming their informatics markets (e.g. in chemistry).

The potential value of semantic objects (data, documents) is huge. They can often be integrated seamlessly, used for searching, constraining input, transformation, and potentially becoming intelligent objects that can be questioned or computed. "Find all compounds X1, X2 with symmetry S1, S2, property Y1, Y2 , values V1, V2 measured in the tempertaure range T1, T2, listing the analytical methods and the suppliers".

Progress is slow, and I'll give some reasons, mainly people-problems rather than scientific. It needs (selfless) community collaboration; we see this in crystallography and biosciences but not so much in chemistry and materials. Some organizations actively prevent collaboration as it threatens closed markets and vendor-lockins. Ontologies can be used for social control. Sometimes multiple efforts compete, or are too short terrm - ontologies take many years. Design by itself is not enough; we need tools, examples, training, infrastructure and this rarely maps onto "publication success".

But there are promising signs, especially for generic ontologies. Historically ontologies tended to be huge, in languages developed by computer scientists or theoretical ontologies who did not package them for general use. That's changing. The RDF-triple formalism is now widepsread (e.g. in Wikidata) and has many tools, in particular the SPARQL search language. Young scientists have a greater fluency in computable data and are not afraid of ontologies. We can design a system which would work, if it were built.

Major informatics providers (Google, Microsoft, Yahoo, Yandex) created the public schema.org. NIH/NLM have developed the JATS specification for scientifc publications. And Wikimedia have created the 100-million-entry Wikidata ontology. Many common objects (people, organizations, etc.) are already formalized and supported. In Biosciences there is Ontology Lookup Service (which, for example, contains Units of Measure - we don't have to reinvent that). Usable ontologies can be built from re-usable components.

The challenge for materials science is to find "the will and the means" (Bernal) to create convincing prototypes, support them, educate and change data and document creation.

Tuesday, Oct 5 at 5:20pm CEST

Standard Access to Datasets for Training Interatomic Potentials

Ellad B. Tadmor,
Department of Aerospace Engineering and Mechanics,
University of Minnesota, Minneapolis, MN 55455

Atomic interactions in classical molecular simulation are modeled using a function called an interatomic potential (IP) or force field. Traditionally, IPs have been expressed as functional forms that aim to explicitly represent aspects of the bonding and/or geometry of the system and are fitted to relatively small datasets of key material properties. In recent years interest has grown in data-driven IPs (DDIPs) in which machine learning methods are used to interpolate first principles calculations. Due to the lack of explicit physics, DDIPs must be trained on large datasets, and must be frequently retrained when applications fall outside the original dataset. To facilitate this and allow research groups to easily exchange DDIPs along with their training datasets, it is important to develop a standard for archiving and retrieving datasets. This effort is being pursued as part of the ColabFit project (https://colabfit.org) which aims to facilitate the development, exchange and deployment of DDIPs and their datasets through the OpenKIM framework (https://openkim.org).

Wednesday, Oct 6 at 11:00am CEST

Needs for ontologies for experimental databases – experience of the COD

Saulius Gražulis,
Vilnius University, Lithuania

Experimental databases are valuable resources that are designed to record data about measurements performed during actual research. The Crystallography Open Database (COD, https://www.crystallography.net) records data about experimentally determined crystal structures of small molecules and minerals that were solved using variety of diffraction techniques such as single crystal or powder diffraction, using X-rays, neutron or electron radiation. To record these results faithfully and unambiguously in a machine readable form, COD employs the Crystallographic Interchange Framework (CIF) developed and maintained by the IUCr. This framework enables us to record all relevant data and metadata in a machine readable for subsequent data reuse.

As derived results are being produced from the COD data, however, new computational entities emerge, such as reconstructed molecules, coordination complexes and their ligands, polymers (molecular nets) within crystals, representations of disordered molecule forms and the like. The original CIF can be used to represent some of the aspects of these derived data, but not all aspects of it. Since the CIF core dictionaries were not designed to represent data outside the realm of crystallography and crystallographic data exchange, some notions (such as space group notion) become imprecise since they were determined differently in, say, the file with the reconstructed molecules. It is highly desirable to reuse those parts of the CIF framework that are common to both original and derived data and maintain unchanged semantics, but there is the need to develop new semantic designators in order to describe newly generated data. It seems that ontologies, both the generally developed top level ontologies, as well as locally created ad-hoc ontologies could be helpful to describe the new computational objects that emerge during computations and their relations with experimental observations. In the talk different problems and their possible solutions will be brought to discussion in the workshop.