What's in a Name? Quite a Lot, as it Happens!

by Mark I. Borkum and Jeremy G. Frey

Names are essential to data manipulation and data interpretation. IUPAC standardizes the names that chemists use in their scholarly works, which it publishes as a suite of terminology, nomenclature and ontology, the IUPAC colour books. Currently, machine-accessible representations of these publications are not available on the Web. In this article, we argue the case for Web-based, machine-accessible representations of IUPAC publications.

What's in a name? Names are used to identify whole classes of things, or individual things, either uniquely, or within a given context. Scientific disciplines standardise their terminology (sets of names), nomenclature (rules for the selection of names) and ontology (denotation of names and definitions of associated things) to ensure that their scholarly works have unambiguous interpretations. Names are also an essential component of the architecture of the Web, where they are used to identify Web resources.

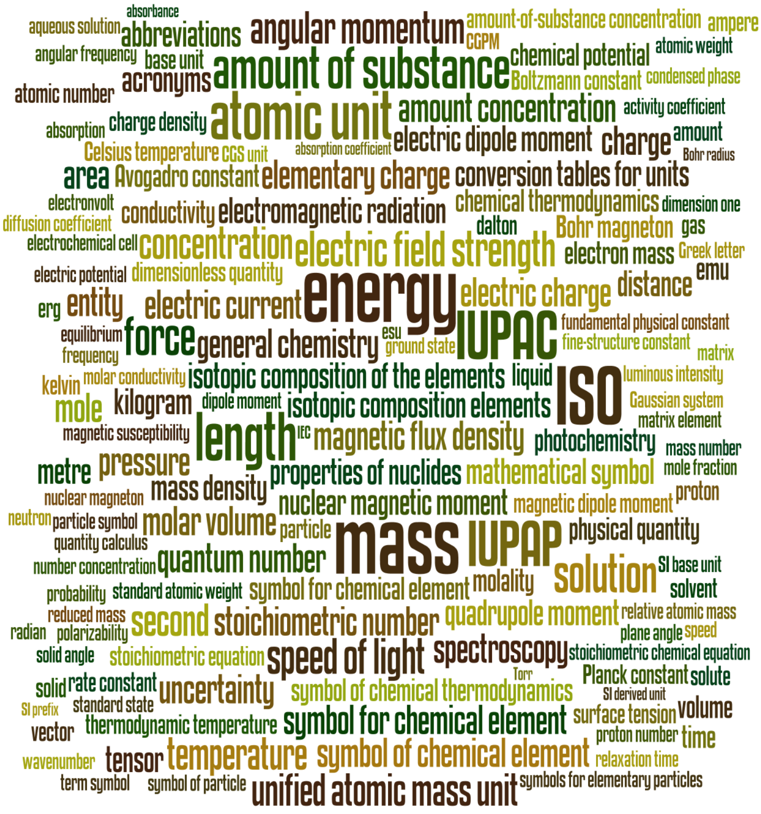
Today, an increasing number of chemists, working around the world, disseminate their scholarly works using the Web. Some, with the assistance of specialist publishers. Unfortunately, instead of being readily available for data integration, much of the world's chemical information is "trapped" inside of vast "data silos", whose contents are accessible to humans, but not machines. The *lingua francas* of the Web (HTML, PDF, e-book, etc.) are rudimentary emulations of paper and ink, designed for data presentation, not data communication. As these data formats do not codify names, it is not possible to delineate and explicate data structure, and hence, the information content of the resultant Web resources is inaccessible to machines.

One of the main roles of IUPAC is to standardise the names that chemists use in their scholarly works. Accordingly, it publishes the IUPAC colour books: a suite of terminology, nomenclature and ontology for chemistry. Currently, only five of these publications are available online [1,2], represented as a mixture of unstructured and semi-structured Web resources that cannot be easily reused by software developers. These projects are *ad hoc* and their outputs mutually incompatible, lacking long-term planning and centralised coordination, as demonstrated by the fact that the only comprehensive list of Web resources is provided by Wikipedia, and not IUPAC's own homepage (which, as a particular example, omits compendia like the Silver book).

But all is not lost! A critical observation is that many, if not all, IUPAC publications are typeset using software-based document preparation systems, meaning that, given some preprocessing, the information content of these publications, such as the subject indices, can, in principle, be made available for data integration.

At the University of Southampton, we are exploring the usage and applications of Semantic Web technologies for chemistry research. In a recent publication [3], we describe the extraction and enrichment of the subject index of the IUPAC Green Book [4]. We note that the subject index is of a particularly high quality. An IUPAC-endorsed, machine-accessible representation would be of considerable interest to software developers. The image [below] is a depiction of the weighted frequency list (or "tag cloud") of the most frequently referenced terms in the subject index of the IUPAC Green Book, rendered using Wordle [5].

"Tag cloud" of the IUPAC Green Book subject index



As a follow-up, and in conjunction with the Royal Society of Chemistry's Chemical Information and Computer Applications Group (RSC CICAG) [6], we organized the one-day meeting, "What's in a name: Terminology and nomenclature, the unsung heroes of open innovation" [7], which was held on 21 October 2014, at Burlington House, London, UK. Presentations covered a wide range of topics of interest to both industry and academia, including: the representation of crystal structures, polymers and chemical reactions; the impact of the Web on the communication of chemical information; and the challenges of managing translational research in an "open" software architecture.

Despite its name, the meeting highlighted the ease with which, from a computer science perspective, many common misunderstandings about names permeate human discussion. For example, it is all too easy to confuse the name of a thing with the thing itself, to ignore the distinction between the processes of identification and resolution, or to forget that the same name can be resolved by more than one identity provider. To paraphrase the Belgian surrealist, René Magritte: "Ceci n'est pas une structure chimique," (ceci est une représentation d'une structure chimique).

In this niche area, chemists risk succumbing to the "curse of knowledge", focusing on the minor details of their own discipline while bypassing the major practicalities of software engineering; an issue that can only be resolved by actively seeking collaboration with computer scientists. There are many fine examples of "chemist-ware" on the Web, but their developers represent an absolutely tiny fraction of the world's chemists, who are presently unable to fully express themselves.

The Web is indispensable to modern chemistry research. It is only a matter of time before the "killer app" for chemistry is successfully developed, "infecting" its end-users with its own potentially problematic interpretation of the discipline. If IUPAC does not take immediate measures [8], leveraging the power of its brand to promote a cohesive vision of chemical terminology, nomenclature and ontology on the Web, then it risks being supplanted as the international authority for chemical sciences.

Semantic Web Technologies

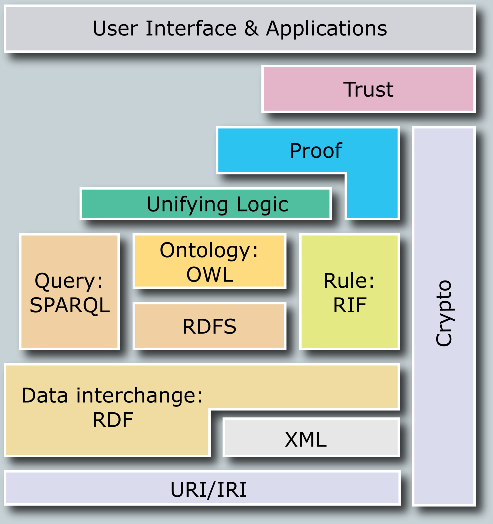
The Semantic Web [a] is a collaborative movement led by the international standards body the World Wide Web Consortium (W3C) [b], whose goal is to transform the human-accessible "Web of documents" into a machine-accessible "Web of data". The Semantic Web is realized as a hierarchy of technologies (see figure), where each successive layer builds upon and extends the capabilities of the preceding layers.

At the base, the fundamental technology of the Semantic Web is the Uniform Resource Identifier (URI) [c], which provides a mechanism to identify the name of a resource. Given identification by one or more URIs, representations of a resource can be exchanged over the network. The most common form of URI on the Web is the URL (the thing that you type into the address bar of your Web browser).

The next layer is the Resource Description Framework (RDF) [d], a family of specifications that collectively define a method for modelling information resources by making assertions about their nature and characteristics. Each assertion takes the form of a "subject-predicate-object" triple, where the subject and predicate are both resources, identified by URIs, and the object is either a resource, identified by a URI, or a literal value, such as a string, number or timestamp.

In RDF, every set of assertions induces a labelled, directed graph, where vertices and edges correspond to resources and assertions respectively. A core capability of RDF is that any two graphs can be added together, to yield a third graph of equal or

Depiction of the stack of technologies underlying the architecture of the Semantic Web (the "layer cake"). Source: www.w3.org



Mark I. Borkum <m.i.borkum@soton.ac.uk> is a Postdoctoral Researcher at the University of Southampton. While his academic background is Computer Science, he is a member of the Chemistry department. His research interests include how the machine-readable representation of chemical, laboratory and health and safety information can support Chemistry-themed use cases. Twitter: @markborkum
LinkedIn: https://uk.linkedin.com/pub/mark-borkum/1b/196/89b/

Jeremy G. Frey <j.g.frey@soton.ac.uk > is a Professor of Physical Chemistry at the University of Southampton. His interests include how e-Science infrastructure can support scientific research, with an emphasis on the way appropriate use of laboratory infrastructure can support the intelligent access to scientific data. Twitter: @profechem
LinkedIn: www.linkedin.com/in/jeremygfrey

References

1. Wikipedia, the free encyclopedia. IUPAC book; avail able at http://en.wikipedia.org/wiki/IUPAC_book
2. IUPAC. Nomenclature and Terminology (including IUPAC color books); available at www.iupac.org/home/publications/e-resources/nomenclature-and-terminology.html
3. Borkum, M. I. and J. G. Frey. "Usage and applications of Semantic Web techniques and technologies to support chemistry research". *J. Chem. Inf.*, **6**(1):18, 2014; http://dx.doi.org/10.1186/1758-2946-6-18
4. Cohen, E. R., T. Cvitas, J. G. Frey, B. Holmström, K. Kuchitsu, R. Marquardt, I. Mills, F. Pavese, M. Quack, J. Stohner, H. L. Strauss, M. Takami, and A. J. Thor. *Quantities, Unit and Symbols in Physical Chemistry*, 3rd edition, 2nd printing (IUPAC Green Book). Cambridge: IUPAC and RSC Publishing, 2008.
5. Feinberg, J.. Wordle - Beautiful Word Clouds; available at www.wordle.net
6. Royal Society of Chemistry. Chemical Information and Computer Applications Group; available at www.rsc.org/Membership/Networking/InterestGroups/CICAG
7. Royal Society of Chemistry. *What's in a name?* available at www.rsc.org/events/detail/11834
8. Frey, J.G.. "Digital IUPAC: A vision and a necessity for the 21st century". *Chem Int.* **36**(1):14-15, 2014; http://dx.doi.org/10.1515/ci.2014.36.1.14

greater extent, i.e., the addition operation for graphs is monotonic [e]. Consequentially, when using RDF, data integration is always possible.

The interpretation of labels in RDF graphs is formalised by two related technologies, RDF Schema (RDFS) [f] and Web Ontology Language (OWL) [g]. The former is a vocabulary for RDF, which facilitates the description of rudimentary entity-relationship models. The latter is an extension of the former, founded upon description logic, which enables the description of arbitrarily complex data models.

Another core technology is the Simple Knowledge Organisation System (SKOS) [h], a standard built upon RDF and RDFS for the representation of controlled vocabularies, including, but not limited to, thesauruses, classification schemes, subject-heading systems, and taxonomies. SKOS employs a concept-centric model of vocabularies, where the abstract notions of the vocabulary are represented by instances of the SKOS "\concept" class. SKOS concepts are annotated with RDF properties, including: index terms (labels), synonyms and alternative spellings, common misspellings, definitions, notes and notations.

References

a. Berners-Lee, T., J. Hendler, and O. Lassila. The Semantic Web. *Scientific American*, **284**(5):28-37 (2001).
b. World Wide Web Consortium (W3C); available at www.w3.org.
c. Berners-Lee, T., R. Fielding, and L. Masinter. RFC 3986, Uniform Resource Identifier (URI): Generic Syntax (2005).
d. Cyganiak, R., D. Wood, and M. Lanthaler, editors. *RDF 1.1 Concepts and Abstract Syntax*. W3C Recommendation. W3C, February 2014
e. Hayes, P. J. and P. F. Patel-Schneider, editors. RDF 1.1 Semantics. W3C Recommendation. W3C, February 2014
f. Brickley, D. and R. V. Guha, editors. *RDF Schema 1.1.* W3C Recommendation. W3C, February 2014
g. Hitzler, P., M. Krötzsch, B. Parsia, P. F. Patel-Schneider, and S. Rudolph, editors. *OWL 2 Web Ontology Language Primer* (2nd edition). W3C Recommendation. W3C, December 2012
h. Miles, A. and S. Bechhofer, editors. *SKOS Simple Knowledge Organization System Reference*. W3C Recommendation. W3C, August 2009