## Chapter 13

# Chemistry Ontologies

**Colin Batchelor***

**Royal Society of Chemistry, Thomas Graham House, Cambridge, United Kingdom CB4 0WF**
***E-mail: batchelorc@rsc.org**

I provide an overview of ontologies in chemistry, what they are, how they are used at present, where they might be used in future and where they fall short of what you might hope for. In particular I describe their application in a large drug discovery infrastructure project and how the approach taken there might be applied to providing machine-readable descriptions of chemical experimentation in general.

## Introduction

A growing area of research in recent years has been the application of ontologies to microarray data and integrating such data with other sources of information. Concomitantly there has also been a great deal of interest in the semantic web. As a discipline, chemistry is no exception in producing and disseminating large amounts of heterogeneous data, as exemplified by PubChem, ChemSpider and ChEMBL. A recent innovation has been the use of ontologies in order to classify, disseminate and link this information. An early example of this, linking together genes, proteins, genetic variations, chemical compounds, diseases and drugs is given by Chen et al. (*1*) in the form of the Chem2Bio2RDF project.

In this chapter I will explain what "ontology" means in this context and cover how ontologies are structured, means of representing ontologies, examples of chemical ontologies, including those produced and distributed by the Royal Society of Chemistry as part of its mission to advance the chemical sciences, and applications of ontologies. In terms of applications I will chiefly focus on the data produced by various sources as part of the Open PHACTS project (*2*). This project has been set up as part of the European Union's Innovative Medicines initiative to support drug discovery programs in the public domain

In The Future of the History of Chemical Information; McEwen, Leah R., et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2014.

and in pharmaceutical companies by delivering web interfaces and application programming interfaces (APIs) for providing chemical, pharmacological and biological data about small molecules and proteins. As part of this project at the Royal Society of Chemistry we generate data sets that describe synonyms and identifiers, calculated physicochemical properties for compounds and links between different data sources.

# Background

## What Is an Ontology and How Is It Structured?

First of all, what do I mean by "ontology"? In the Stanford Encyclopedia of Philosophy, Hofweber (*3*) offers the following four possibilities:

(O1) *the study of ontological commitment, i.e. what we or others are committed to,*

(O2) *the study of what there is,*

(O3) *the study of the most general features of what there is, and how the things there are related to each other in the metaphysically most general ways, and*

(O4) *the study of meta-ontology, i.e. saying what task it is that the discipline of ontology should aim to accomplish, if any, how the questions it aims to answer should be understood, and with what methodology they can be answered.*

For my purposes in this chapter I shall modify (O3) and take an ontology to be a machine-readable account of what there is in a given domain and how the things there relate to other things, not necessarily in the most metaphysically general way, but in a way that is consistent with how practicing scientists in a domain understand the relations.

What do I mean here by a machine-readable account? The word "proposition" in the philosophical literature has a number of meanings, but a useful one, and one that is compatible with machine reasoning, is as a bearer of truth-value: a statement that can be evaluated as being true or false. Any proposition must be "about" something, something that makes it true, and we call the things mentioned in a proposition "entities". In this chapter I take a machine-readable account to mean a series of propositions about the entities within a given domain, for example chemistry or stamp-collecting. Why is machine-readability important? Simply because the datasets encountered for cheminformatics applications, particularly virtual screening in the context of drug discovery, often number millions of structures or tens of thousands of scientific articles and this is too many for an unassisted human being to deal with.

One's first thought about machine-readability in cheminformatics might be something expressed in a file format, such as a V2000 mol file, or in a line notation. Line notations express a chemical structure or a reaction involving chemical structures as a linear sequence of characters and five examples of line notations in contemporary use are given in Table I. SMILES notation (*4*) represents molecules in a human-readable way, for example cyclobutane is C1CCC1, indicating that there are four carbon atoms arranged so that the last is bonded to the first, with hydrogen atoms as needed to make up the numbers. The InChI representation (*5*) is InChI=1S/C4H8/c1-2-4-3-1/h1-4H2, which specifies all of the atoms present and their connectivity. This being largely composed of punctuation is ill-suited to indexing in a search engine and so the InChIKey was introduced to fill the gap. To specify parts of molecules, the SMARTS specification has been built upon SMILES, and the SMIRKS notation combines this with atom indexing in the reactants and products in order to provide atom–atom mapping. On their own, however, expressions in line notation are neither true nor false. For this reason, a string written in a line notation does not count as a proposition on its own. However, it might be part of a proposition, for example "the SMILES string for benzene is c1ccccc1".

**Table I. Line notations in cheminformatics**

| *Line notation* | *Example* | *Interpretation* |
|---|---|---|
| SMILES | C1CCC1 | Cyclobutane molecule |
| InChI | InChI=1S/C4H8/c1-2-4-3-1/h1-4H2 | Cyclobutane molecule |
| InChIKey | PMPVIKIVABFJJI-UHFFFAOYSA-N | Cyclobutane molecule |
| SMARTS | [CX1]#[NX2] | Nitrile group |
| SMIRKS | [c:1][C:2](=O)O>>[c:1][C:2]=C(=O)O | Perkin reaction |

The sorts of proposition we find most often in chemistry ontologies include:

**Subsumption:** for example, every benzene is an aromatic molecule. (E1)
**Parthood:** for example, every benzene has part some benzene ring. (E2)
**Representation:** for example, this connection table represents such-and-such a molecule. (E3)
**Participation:** for example, every Diels–Alder reaction has participant some diene. (E4)

Within an ontology, the propositions are not, however, represented as sentences as in the above examples. They are represented as a string of textual identifiers for the domain entities and the relations between them. Turtle format (*6*) provides a human-and-machine-readable method for this, listing textual identifiers for the subject (benzene, connection table, Diels–Alder reaction), the predicate (is a, has part, represents, has participant) and the object (aromatic molecule, benzene ring, such-and-such a molecule, diene) separated by spaces and completed with a full stop. Example (E1) rewritten in this format would be:

obo:CHEBI_16716 rdfs:subClass obo:CHEBI_33655 where rdfs:subClass is the "is a" relation and the strings beginning with "obo:" refer to the classes "benzene molecule" and "aromatic molecule". By "class" I mean that the proposition relates to benzene molecules in general, as opposed to a specific benzene molecule under the tip of a scanning–tunneling electron microscope. As you can see, the identifiers are relatively opaque so that they do not have to change if our knowledge about a subject changes or if someone has misspelt something or, for example, a taxonomic species is renamed in the light of new discoveries. This is a feature of biomedical ontologies; the ontologies that computer scientists develop, often to test code that draws inferences based on the propositions in an ontology, to teach people about how reasoning works or to explore the expressiveness of a given ontology language, will have non-opaque IDs because it is important that these propositions should be easily readable by a human being, and revision in the light of new scientific knowledge is less important. One principle of the Semantic Web is that identifiers should have a readily-accessible definition over the web; this implies that URLs should be used.

The prefixes rdfs: and obo: are shorthands for fragments of HTTP URLs. One underlying notion of the Semantic Web is that data should be in some sense self-describing; hence these identifiers should (and in the rdfs: and obo: cases do) resolve over the web to a machine-readable description. Later in this chapter (under Representing Ontologies) I will illustrate some of these machine-readable descriptions and the conventions behind them.

Taken together, these propositions constitute a system that can be checked for internal consistency, for example if the ontology defines somewhere that no protection reaction can also be a deprotection reaction, then deprotections that have been manually misclassified as protections can be identified. This is not a fanciful example; this is something I personally have done by mistake. This can be done programmatically, for example using a reasoner, that is to say a program that draws inferences, or within an ontology editor. A reasoner can also be used to infer things not made explicit in the system. For this sort of reasoning we need quantification, that is to say, what propositions are true of every *x*, some *x* or perhaps no *x*. I will discuss this in greater detail later on.

Even without the propositional structure, the mapping between identifiers and human-readable names, ideally the names that are found in the scientific literature, is in itself a useful artefact that can be used for indexing or more sophisticated forms of text mining, and I will discuss this in more detail in the Applications section.

There are clear similarities between an ontology and a database. One way of thinking about a database is that records contain propositions about entities, the role of the identifiers in an ontology being played by the primary keys in the database. We can think of a query with joins (one that combines data from different tables to extract information that may not have been explicitly put into the database) as being analogous to reasoning over an ontology. To this end, just as there exists SQL, a standard query language for relational databases, so there is SPARQL (*7*), a query language for ontologies and knowledge bases. SPARQL does not allow the underlying knowledgebase to be altered; this has led people, perhaps incautiously, to set up public SPARQL endpoints on the web allowing anyone to query a knowledge base, something which would be extremely hazardous for relational databases as it would enable members of the public to modify and delete information within the database without an audit trail. In fact, "SQL injection", sending an appropriately-formatted string to a website that alters the underlying relational database, is a well-known vulnerability. An analogous vulnerability for a SPARQL endpoint might be a query that returned all of the underlying data. A further important distinction is the contrast between the Closed World Assumption of databases – that anything unknown to the database is false, and the Open World Assumption of ontologies, that anything otherwise unspecified by the ontology we can draw no conclusion from.

In laboratory domains it is often useful to relate the entities to an upper-level ontology, which is a small ontology that typically distinguishes objects (for example molecules) from the processes (for example cyclization) they participate in. This has been used, for example, in the Gene Ontology (*8*) to find errors and inconsistencies. The distinction is less obvious and perhaps less useful when describing software artefacts as computer programs are themselves data. Examples of upper-level ontologies include the Descriptive Ontology for Linguistic and Cognitive Engineering, DOLCE (*9*) and the Basic Formal Ontology, BFO (*10*). In general the former is more popular among ontology researchers and the latter is more popular in the biomedical ontologies community.

## Representing Ontologies

At the moment, ontologies are typically stored in an XML serialization of the Web Ontology Language, OWL (*11*). An example of this is given in Table II. Some of the XML elements are in the owl: namespace, but many others come from the Resource Data Format, RDF (*12*) namespace, as OWL has been built on top of RDF, and as such Table II provides an example of both. The best way of thinking about the two is that OWL is best suited to describing the relations between things in general (types, classes, universals), whereas RDF is better suited to things themselves (tokens, individuals, particulars). The conventional example is that when one talks about Socrates being a man, Socrates is the individual, and man is the type or class.

**Table II. A sample of OWL serialized as XML.**

```
<owl:Class rdf:about="http://purl.obolibrary.org/obo/CHEBI_15734">
<rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string">primary
alcohol</rdfs:label>
<rdfs:subClassOf rdf:resource="http://purl.obolibrary.org/obo/CHEBI_24431"/>
<oboInOwl:id  rdf:datatype="http://www.w3.org/2001/
XMLSchema#string">CHEBI:15734</oboInOwl:id>
</owl:Class>
```

A key feature of OWL, which is an evolving standard, is that it is a subset of first-order logic which deliberately hobbles what you can say in order to ensure that any inferential process will actually terminate. This is important for web applications in order to avoid denial-of-service attacks that would involve a website being given a request that was known not to terminate. OWL comes from the Description Logic (DL) tradition. A key difference between conventional first-order logic and DLs is that first-order logic allows definitions that contain variables. Hence an epoxy molecule can be defined as one where an oxygen atom *o* is bonded to exactly two carbon atoms $c_1$ and $c_2$, and those are themselves bonded to *o*. However, DL disallows this. One can only say that there is an oxygen atom that is bonded to exactly two carbon atoms which are themselves bonded to exactly one oxygen atom and one other carbon atom. While there may be workarounds for small systems, in general the problem is intractable and I will come to some potential solutions in the next section.

DL also has some peculiar terminology – this does not limit its power or scope but adds a layer of perhaps unnecessary opacity to papers describing how it works What would normally be called a predicate is called a role (or an object property in OWL), what would normally be a class is called a concept and what would normally be called a proposition is called an axiom.

OWL divides up the universe of discourse as follows: there are classes (in the language of DL, the TBox or terminology box) for example “man”, and there are individuals that instantiate those classes (the ABox or assertion box), for example “Socrates”. We write propositions about those classes and individuals in terms of properties, the division of which is threefold. There are object properties, which relate classes to classes and individuals to individuals (hence “all men are mortal”), data properties, which relate individuals to strings and other instances of data types, such as floating-point numbers, integers and dates (Socrates was born on the 21st of January), and there are annotation properties, which describe the classes and individuals themselves (Socrates is called “Socrate” in French), rather than, for example, their referents.

Particularly popular in the biomedical domain is the Open Biomedical Ontologies (OBO) format (*13*), which has two practical advantages over unadorned OWL. Firstly, the basic format makes detailed provision for synonyms – as such OBO is well-suited to handling terminologies if not disambiguating them. Secondly, the format is readily human-writable using a simple text editor; see Table III for an example. OWL format, on the other hand, is typically best handled using the OWL API (*14*) or a tool such as Protégé.

**Table III. An example class definition in OBO format.**

```
[Term]
id: RXNO:0000036
name: Reformatsky reaction
def: "A carbon-carbon coupling reaction where an aldehyde or amine reacts with a
alpha-halo ester and zinc to form a beta-hydroxy ester." [RSC:cb]
synonym: "Reformatskii reaction" EXACT []
is_a: RXNO:0000002 ! carbon-carbon coupling reaction
relationship: has_part MOP:0000580 ! ketone reduction
relationship: has_part MOP:0001550 ! dehalogenation
```

The bridge between the two has come from both the logical end and the human-interface end. Firstly Golbreich et al. (*15*) have hardened up the previously informal semantics of OBO by providing a mapping to OWL. I will give an example of why this is necessary later on in my discussion of quantification. Secondly, OWL Manchester Syntax has been developed (*16*) to serve two ends: firstly to provide a human-writable way of typing propositions in to an editor, and secondly to provide a more user-friendly way of showing these propositions than the symbolic "German" syntax that had been used previously. Table IV shows an example of this. A particularly exciting use is for providing explanations of inconsistencies (*17*), as in my protection/deprotection example earlier. and a human-readable notation is also particularly useful for explaining the output of a reasoning process. A trivial example might inferring that a cat is a mammal because all cats are felids and all felids are members of Carnivora and all members of Carnivora are mammals.

**Table IV. An example of OWL Manchester syntax taken from (*15*)**

```
Class: VegetarianPizza

EquivalentTo: Pizza and not (hasTopping some FishTopping) and not (hasTopping
some MeatTopping)

DisjointWith: NonVegetarianPizza
```

An important feature of most relations in scientific ontologies, for example parthood and participation, is that they express an ontological dependence, that is to say that it is impossible for something to be a benzene molecule without having as part some benzene ring. However, names and identifiers are not like this. It is neither necessary to the word "benzene" nor to a benzene molecule itself that the other exist. The same is true of words like "wyvern" and "polywater". For many applications we do not want to reason over these inessential properties so OWL provides annotation properties which can be used for indicating synonyms and identifiers, particularly line notation in the case of molecules and reactions.

How would we deal with references to wyverns and polywater in text? We certainly shouldn't define polywater as a kind of polymer or a wyvern as a kind of animal as this would lead to nonsensical entailments, for example any papers that showed that there was no such thing as polywater would contain existential statements along the lines of "there is some *p* such that *p* is polywater and there is no *p*". It is better to define them in terms of the polywater hypothesis, or, in heraldry, the wyvern shape. There are as yet no good automated ways of handling these in text-mining as seldom-referenced dead-end hypotheses such as polywater are not generally amenable to high-throughput analysis.

# Examples

## Example Ontologies in Chemistry: Small Molecules

One might ask, given the power of chemoinformatics methods such as 2D fingerprinting, substructure search and scaffold hopping, why one might need a hierarchical hand-built system for managing knowledge about small molecules.

Hastings et al. (*18*) argue, inter alia, that there is a wealth of information in textual descriptions of classes of molecule, listing examples like "1-alkenoylcyclopropane carboxamides", which better describe the focus of an article or indeed the research agenda of an entire group than any scaffolds one might be able to infer from looking at full structures. Aside from the systematic names there are also natural-product-based names such as "polyketide" or "spongistatin" that reflect the origin of the molecules in question.

As discussed by Richter (*19*), the basics of chemical classification in its modern form date back to around 1840, in terms of parent nuclei, homologous series and functional groups. These have subsequently been refined and enlarged and codified by bodies such as IUPAC as detailed in the Red (*20*) and Blue Books (*21*).

A long-established flagship ontology project in the biomedical domain consists of the Gene Ontology (*8*), and the Gene Ontology annotation (GOA) database (*22*), which together provide a wealth of information about biology from the molecular up to the organismal level. The Gene Ontology provides vocabularies for cellular components, molecular functions and biological processes, while GOA is an abstracting service for the biological literature that annotates gene products, that is to say proteins and messenger RNAs, with their molecular function, where in the cell they do their work, and what broader biological processes these are implicated in.

Chemical Entities of Biological Interest, ChEBI (*23*), started initially as a dedicated chemical classification for those classes in the Gene Ontology that reference small molecules. A detailed description of how the two ontologies interact is given by Hill et al. (*24*). It was subsequently developed as a reference implementation of IUPAC guidelines as specified in the Red and to some extent the Blue Books (*20*, *21*) for chemical nomenclature in the sense that the entries for a given name are to be taken as definitive, although it does not provide a resolution service for unknown names. It has subsequently gained links to patent databases and the natural product literature.

An important development in the transition of ChEBI from being merely a controlled vocabulary to an ontology per se has been its treatment of quantification. An important feature of ChEBI is that its underlying storage medium is a relational database on which humans make queries and the OBO and OWL versions are merely serializations. The interpretation of an entry in the database, for example, that the relation table has a record that connects the oxygen atom with a parthood relation and the water molecule, depends on the chemically-aware reader. Informally one might say that an oxygen atom is part of a water molecule. Initially, based on the Gene Ontology, such relations were expressed in terms of a part_of relation. However, the quantification of the OWL translation of the OBO-style relation "tetracyanonickelate(2-) part_of potassium tetracyanonickelate(2-)" is that "all tetracyanonickelate(2-) ions part_of some potassium tetracyanonickelate(2-) complex", which is patently absurd. Better to say, as ChEBI now does, that "all potassium tetracyanonickelate(2-) complex has_part some tetracyanonickelate(2-) ion" – any complex lacking a tetracyanonickelate(2-) ion cannot be a potassium etracyanonickelate(2-) complex. This is now codified in the OWL files that are available for download from the ChEBI website.

ChEBI is, however curated by hand. As of January 2014 it contains 37271 classes that represent molecules, families of molecules with a detailed structural classification or roles played by those molecules organized into a hierarchy, which is rather too big a number to systematically maintain without automated assistance. There are also many classes within the ontology with only a rudimentary classification which the curators will come to. One example of how its curation might be automated is given by Bobach et al. (*25*) who build a ChEBI-like ontology with a hybrid approach that relies on well-established cheminformatics methods to automatically classify molecular structures. To be precise, they use SMARTS expressions as seen in Table I to specify the connectivity of atoms within a molecular structure. Note that only the *a priori*, structural component can be automated reliably; the process of curating what molecules do inside organisms is a posteriori and is necessarily based on experiment.

A wholly formal-logical approach is demonstrated by Magka (*26*), who expresses chemical structures in terms of propositions in first-order logic, for example, two-place predicates to express bonding between atoms (single($f_{12}$(x), $f_i$(x)) expressing the notion that there is a single bond between atom 12 and atom *i*), and one-place predicates to express the properties of those atoms, for example c($f_i$(x)) indicating that atom *i* is a carbon atom, as shown in Fig. 1, and then shows rules that may be used to determine subclass relations between them. Unlike previous work, for example by Hastings et al. (*27*), the classification code runs in a reasonable amount of time, though it is still slower than implementations based on matching SMARTS expressions.

$$\begin{aligned}\mathsf{ascorbicAcid}(x) \rightarrow\ & \wedge_{i=1}^{13} \mathsf{hasAtom}(x, f_i(x)) \wedge \mathsf{molecule}(x) \wedge_{i=1}^{6} \mathsf{o}(f_i(x)) \wedge_{i=7}^{12} \mathsf{c}(f_i(x)) \wedge \\ & \mathsf{h}(f_{13}(x)) \wedge \mathsf{single}(f_8(x), f_3(x)) \wedge \mathsf{single}(f_9(x), f_4(x)) \\ & \wedge_{i=1,9,11,13} \mathsf{single}(f_{10}(x), f_i(x)) \wedge_{i=5,11} \mathsf{single}(f_{12}(x), f_i(x)) \\ & \wedge_{i=1,8} \mathsf{single}(f_7(x), f_i(x)) \wedge \mathsf{single}(f_{11}(x), f_6(x)) \wedge \\ & \mathsf{double}(f_2(x), f_7(x)) \wedge \mathsf{double}(f_8(x), f_9(x))\end{aligned}$$

*Figure 1. Formal-logical representation of ascorbic acid according to Magka (26). (Reproduced with permission from reference (26). Copyright 2012)*

There has been comparatively little work on larger systems; however the NanoParticle Ontology (*28*) is a promising piece of work for the nanosciences which contains not only a set of classes of nanoparticles, but also their properties (in the familiar sense of the word), for example the surface area, chemical composition, surface charge and zeta potential of a nanoparticle surface. It also provides relations that can be used to specify the composition of a nanoparticle (has_entrapped_component_part, has_encapsulated_component_part and so forth). It is being used by the United States National Cancer Institute's Nanotechnology Working Group in their work on the rational design of nanomaterials and in finding nanomaterial structure–activity relationships. As of December 2013 it has 1904 classes.

## Example Ontologies in Chemistry: Processes

The domain of the CHEMINF ontology (*29*) is chemoinformatics. To this end it represents both chemoinformatics algorithms and the data that they process and output. The algorithms mentioned include those to calculate the polar surface area of a molecule, to calculate partition coefficients and to standardize chemical structures according to some set of rules. The data items include molecular connection tables, molecular formulae and numbers of freely-rotating bonds. Importantly the ontology includes references to software packages, which provides a means to give the provenance of a given calculation. It is being used in the Open PHACTS project (*2*) as I will describe below. As of January 2014, it has 652 classes.

The Chemical Methods Ontology (CHMO) (*30*) was initially based on the IUPAC Orange Book (*31*) and intended to cover the methods described therein for collecting analytical data, such as mass spectrometry and electron microscopy. Subsequently it has been extended to cover the methods to prepare and separate material for further analysis, such as sample ionization, chromatography and electrophoresis, to synthesize materials, such as epitaxy and continuous vapour deposition, the instruments used in these experiments, like mass spectrometers and chromatography columns, and their outputs. It now (December 2013) has 2745 classes. It was initially developed for text mining as part of the RSC's Project Prospect, this text-mining being ongoing, but should be usable for describing all aspects of an experiment. The Golm Metabolome Database, a reference library of GC-MS experiments (*32*) uses CHMO to describe some of the parameters in gas chromatography and mass spectrometry experiments

As for small-molecule reactions, Ingold's nomenclature for reaction mechanisms goes back to before the Second World War. Carey et al. (*33*) offer a categorization of small molecule reactions and classify reactions from the databases of AstraZeneca, GlaxoSmithKline and Pfizer against them. This categorization, excluding "miscellaneous", has 11 categories, which are focussed on the chemical transformations from a synthetic point of view rather than the precise mechanism. These are heteroatom organylation, acylation, carbon–carbon bond forming, aromatic heterocycle formation, deprotection, protection, reduction, oxidation, functional group interconversion, functional group addition and resolution. This has not, however, been formalized into an ontology that can be reasoned over. That falls to RXNO, the name reactions ontology (*34*), which has 511 classes as of January 2014. The top levels of the "intentional" classification in slight contrast to Carey et al.'s classification are cleaving, condensation, functional modification, joining, rearrangement, ring breaking, ring contraction, ring expansion, ring formation and ring rearrangement. The "intentional" classification is based on two principles: firstly comparing the unbroken carbon chains in the reactants and products and secondly considering whether a ring system is created, destroyed or altered. These are all worked out from the perspective of an organic chemist

Here, as in the case of small molecules, we come up against the limits of the DL approach. To take the Diels–Alder reaction, it is necessary but not sufficient for a reaction to be a Diels–Alder reaction if it involves the reaction of a diene with a double-bonded system producing a cyclohexadiene. The ring itself must consist of those atoms that previously constituted the diene part of one reactant and the double-bonded system of the other. We can express this using SMIRKS notation because SMIRKS notation allows us to number atoms and hence provide a mapping from the reactants to the products, but not with the resources available to us within OWL as it is impossible within the definitions allowed in a DL framework to talk about a given atom as we saw in the epoxy example previously. Any approach would have to, like in Magka's work on chemical structures, go outside the DL framework and this is not currently supported by well-established web standards.

## Applications

### Text Mining

The most straightforward application of an ontology, and one that does not require any of the logical apparatus, is in named-entity extraction to provide a controlled vocabulary of terms found in text and identifiers for those. The generic named-entity extraction process works roughly as follows: a document is segmented into sentences, then those sentences are tokenized (split on spaces and relevant punctuation) and those tokens or token sequences assessed for their likelihood of being named entities relevant to the domain. The simplest way of doing this is to compare them to a pre-existing dictionary, for which ontologies are pre-eminently suitable. It is worth mentioning in passing that in chemical documents most of the compounds of interest will be brand new and hence in no dictionary, so in general a name-to-structure approach will be needed.

An example explicitly using chemical ontologies is provided by Batchelor and Corbett (*35*) who describe in detail how named entity recognition based on ontology identification can be applied to annotate a journal article stored as XML by adding more XML elements to it, but more general examples abound outside chemistry, particularly one hand-built example by Shotton *et al.* (*36*). Ontologies do not inherently help with the task of word-sense disambiguation beyond providing different identifiers for the same name, which is useful to distinguish the senses of a word like ‘cell’, which could refer to a biological cell, an electrochemical cell, a solar cell or possibly in an environmental monitoring journal a room that accommodates prisoners, or “plant” in a botanical context as opposed to a manufacturing context. Ontologies also provide textual definitions which could be used in examining the immediate textual context of a name to provide clues as to its referent. In (*37*) Corbett *et al.* use the word ‘pyridine’ to exemplify the more tractable case where chemical names may have more than one reading and how they may be disambiguated.

The first distinction is between lab-scale and molecular-scale. “Pyridine” may refer to the substance in a bottle or to a given molecule. In (*37*) the authors leave this unresolved as it is in practice a less important distinction than the second, which is between “pyridine” the cyclic molecule with formula $C_5H_5N$ and “a pyridine”, any molecule containing an unfused aromatic $C_5H_5$ ring. They call these the EXACT and CLASS readings respectively. A third, practically-driven sense is that of “pyridine” in “pyridine ring”, which they call the PART reading. This is related to the CLASS sense in that it refers to the aromatic $C_5H_5$ ring *per se* rather than merely a molecule that contains one. This threefold distinction is honoured in ChEBI where “pyridine”, “pyridines” and “pyridine ring” are different classes and have different identifiers.

The dissemination of these annotations is not restricted to the “Rich HTML” view on the RSC Publishing Platform as described in (*38*). As most readers still prefer to read PDFs rather than HTML, Pettifer et al.’s Utopia Documents PDF reader (*39*) uses information from the RSC’s web services to show the annotations found by text mining within the PDF on screen.

## Open PHACTS and Other Datasets of Pharmacological Interest

As part of the Open PHACTS project (*2*), the Royal Society of Chemistry pulls together molecular structures from a variety of databases, chiefly ChEBI, ChEMBL and DrugBank, validates them and produces linksets between them. We use the Vocabulary of Interlinked Datasets (VoID) (*40*) to specify what predicates are used and what sorts of subject and objects are being interlinked. These are particularly valuable because RDF documents can be huge, containing millions of triples, and the VoID provides a concise summary of many triples using each predicate there are. We also use the Open PHACTS dataset specification (*41*) to specify what the justifications are for each connection – for example the structure–structure mapping is based on the InChIKey (see Table I) and a class from the CHEMINF ontology is used to indicate this. The SKOS vocabulary (*42*) is also used to distinguish those links that hold in all cases and those links that only hold under certain circumstances, such as those produced by disregarding stereochemistry or isotopic substitution.

Another dataset we produce is sets of validated and unvalidated synonyms for free-text querying. Given that these synonyms, especially the “unvalidated” ones (synonyms that have come in to ChemSpider from a chemical vendor and will not have been curated by a human being) are inessential properties of the molecule, we take the chemical identifier classes from the CHEMINF ontology and treat them as OWL annotation properties. This enables us to provide more detail in the RDF about what the identifiers are while keeping the RDF relatively simple and easy to query over.

The EBI provides pharmacological data from the ChEMBL database as RDF, as described by Willighagen et al. (*43*). ChEMBL contains a very heterogeneous set of pharmacological data with over 5000 different kinds of activity being reported, so in order to ensure 100% coverage of the data within ChEMBL, the authors took a non-Semantic-Web approach to the RDF, using textual strings in the RDF to specify the activities instead of making the considerable effort of adding nearly 5000 classes to a pre-existing ontology.

We take a different approach to the ChEMBL RDF for the physicochemical properties. Because we have a relatively small (about two dozen) set of physicochemical properties as listed in Table V, we have minted classes in the CHEMINF ontology for each of them. Then we take a Davidsonian event semantics (*44*) approach, similar to that taken by some groups using the Ontology for Biomedical Investigations (OBI) (*45*). By "Davidsonian" I mean that we base everything around an event, call it *e*, which in our case is a particular execution of an algorithm to calculate a physicochemical property that takes input in the form of a connection table and produces output in the form of a calculated property, both of which we label using classes taken from CHEMINF. We relate the inputs and outputs to the event *e* with relations from OBI, specifically has_specified_input and has_specified_output, which capture those inputs and outputs that are necessary and characteristic of the process. A cheminformatics calculation, for example, is likely to need a molecular connection table and to output some calculated value. It may also take in as input the start time, or the username of the account under which it is running, and output debugging information, heat and a peculiar whirring sound from the hard drive, but those latter are not "specified" inputs or outputs in the sense that OBI uses them.

**Table V. Properties calculated by ACD/Labs software for the Open PHACTS project.**

| *Kind* | *Property* |
|---|---|
| Bulk | log *P*, log *D* at pH 5.5, bioconcentration factor, $K_{oc}$ at pH 5.5, molar refractivity, molar volume, surface tension, density at STP, flash point at 1 atm, boiling point at 1 atm, enthalpy of vaporization at STP, vapour pressure at STP |
| Molecular | polar surface area, polarizability, index of refraction, number of hydrogen bond acceptors, number of hydrogen bond donors, number of freely rotating bonds |

## Summary and Outlook

In this chapter I have given a perhaps necessarily personal overview of the current state of play for ontologies in chemistry, what ontologies are in this context, how they are being used and whom by, giving an inside view focusing on the vast datasets produced as part of the Open PHACTS project. As a scientist by training and as someone working for a learned society and often collaborating for these purposes with other scientists at, for example, the European Bioinformatics Institute, my viewpoint will be somewhat different from that of a computer scientist working on an ontology research project.

These are still early days in the field of chemistry ontologies and there is as yet much untapped potential. The approach detailed in the previous section for handling cheminformatics calculations, based partly on OBI, could, taken together with the process ontologies (CHMO and RXNO) described above, be extended to much of the rest of chemical experimentation. Typically an experimental process will take a physical sample as input, process it in some way, and then make a measurement. In the Davidsonian approach the single event *e* is replaced with a chain of events $e_i$, the specified inputs of each event being an output or outputs of a previous event. Frey *et al.* at the University of Southampton have for a long time espoused a vision whereby a similar sort of machine-readable account of experiments is generated automatically by electronic lab notebooks (*46*) which are integrated with the experimental apparatus. A simple way in which this approach could benefit from ontologies is if different research groups shared the same identifiers for the different stages in their experiments. One clear opportunity is in computational chemistry, where the "experiments" are necessarily born digital. However, ontologies have historically had most traction in fields such as biocuration where the practitioners are skilled enough to take advantage of them but not having so much skill, for example at programming or data processing as to have a large collection of home-grown Perl or Python scripts to satisfy their data needs beforehand. Perhaps computational chemists are closer to the latter category than to the first.

As far as the IUPAC colour books are concerned, the Red Book (inorganic chemistry) and Blue Book (organic chemistry) have been at least partly codified by the ChEBI team, much as the Gold Book (chemistry in general) (*47*) and the Orange Book (analytical chemistry) have been codified in the Chemical Methods Ontology. The Green Book (*48*), however, is pristine and untouched. It is a varied book, acting partly as an *aide memoire* for practicing chemists, partly to instruct new chemists in the correct use of notation and the Greek alphabet, and partly to define terms used in particular fields. In the last role it could well provide a useful addendum to the Chemical Methods Ontology. Aside from careful hand curation of ontologies on inspection of the literature, which has been the route hitherto, it could well be possible to leverage the vast amount of experimental information about machines and operating conditions that is stored in CIFs in a semi-systematic way.

As far as chemists in general are concerned, the impact of ontologies has been limited. The pervasiveness of ontologies in the biomedical realm is largely due to the existence of large databases, and as the chemical sciences move more towards large databases, repositories and data-sharing mandates, and as the boundaries between journal articles, supplementary data and raw data become more fuzzy, it could well be that ontologies take a more central role in organizing chemical data than they have hitherto.

## Acknowledgments

This chapter is based on a talk given at the University of Exeter in April 2013, for which invitation the author thanks Zena Wood. Thanks also to Leah McEwen for the invitation to write this chapter and to Leah and an anonymous reviewer for extensive feedback on how the first draft could be improved.

# References

1. Chen, B.; Dong, Z.; Jiao, D.; Wang, H.; Zhu, Q.; Ding, Y.; Wild, D. *BMC Bioinf.* **2010**, *11*, 255.
2. Williams, A. J.; Harland, L.; Groth, P.; Pettifer, S.; Chichester, C.; Willighagen, E. L.; Evelo, C. T.; Blomberg, N.; Ecker, G.; Goble, C.; Mons, B. *Drug Discovery Today* **2012**, *17*, 1188–1198.
3. The Stanford Encyclopedia of Philosophy. http://plato.stanford.edu/archives/spr2013/entries/logic-ontology/, accessed on 2014-05-09.
4. SMILES – A Simplified Chemical Language. http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html, accessed on 2014-05-09.
5. About the InChI Standard. http://www.inchi-trust.org/about-the-inchi-standard/, accessed on 2014-05-09.
6. RDF 1.1 Turtle. http://www.w3.org/TR/2014/REC-turtle-20140225/, accessed on 2014-05-09.
7. SPARQL 1.1 Query Language. http://www.w3.org/TR/2013/REC-sparql11-query-20130321/, accessed on 2014-05-09.
8. Ashburner, M.; et al. *Nucleic Acids Res.* **2000**, *25*, 25.
9. WonderWeb Deliverable D17; http://www.loa.istc.cnr.it/old/Papers/DOLCE2.1-FOL.pdf, accessed on 2014-05-09.
10. Grenon, P.; Smith, B.; Goldberg, L. In *Ontologies in Medicine*; Pisanelli, D. M., Ed.; IOS Press: Amsterdam, 2004; pp 20–38.
11. OWL 2 Web Ontology Language Document Overview (Second Edition). http://www.w3.org/TR/2012/REC-owl2-overview-20121211/, accessed on 2014-05-09.
12. RDF 1.1 Concepts and Abstract Syntax. http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/, accessed on 2014-05-09.
13. The OBO Flat File Format Specification, version 1.2. http://www.geneontology.org/GO.format.obo-1_2.shtml, accessed on 2014-05-09.
14. OWL API Documentation. https://github.com/owlcs/owlapi/wiki/Documentation, accessed on 2014-05-09.
15. Golbreich, C.; Horrocks, I. In *Proceedings of the OWLED 2007 Workshop on OWL: Experiences and Directions*; CEUR Workshop Proceedings: 2007.
16. OWL 2 Web Ontology Language Manchester Syntax. http://www.w3.org/TR/2012/NOTE-owl2-manchester-syntax-20121211/, accessed on 2014-05-09.
17. Horridge, M.; Parsia, B.; Sattler, U. In *Proceedings of the 16th Automated Reasoning Workshop (ARW 2009)*; University of Liverpool: 2009.
18. Hastings, J.; Magka, D.; Batchelor, C.; Duan, L.; Stevens, R.; Ennis, M.; Steinbeck, C. *J. Cheminf.* **2012**, *4*, 8.
19. Richter, F. *J. Chem. Educ.* **1938**, *15*, 310.
20. *Nomenclature of Inorganic Chemistry, IUPAC Recommendations 2005*; Royal Society of Chemistry: Cambridge, 2005.
21. *Nomenclature of Organic Chemistry, IUPAC Recommendations and Preferred names 2013*; Royal Society of Chemistry: Cambridge, 2013.
22. Camon, E.; et al. *Nucleic Acids Res.* **2004**, *32*, D262–D266.

23. Hastings, J.; de Matos, P.; Dekker, A.; Ennis, M.; Harsha, B.; Hale, N.; Muthukrishnan, V; Owen, G.; Turner, S.; Williams, M.; Steinbeck, C. *Nucleic Acids Res.* **2013**, *41* (D1), D456–D463.
24. Hill, D. P.; Adams, N.; Bada, M.; et al. *BMC Genomics* **2013**, *14*, 513.
25. Bobach, C.; Boehme, T.; Laube, U.; Pueschel, A.; Weber, L. *J. Cheminf.* **2012**, *4*, 40.
26. Magka, D. In *Proceedings of the 5th International Workshop on Semantic Web Applications and Tools for Life Sciences (SWAT4LS 2012)*, Paris, 2012; CEUR Workshop Proceedings: 2012.
27. Hastings, J.; Dumontier, M.; Hull, D.; Horridge, M.; Steinbeck, C.; Sattler, U.; Stevens, R.; Hörne, T.; Britz, K. In *Proceedings of the 7th International Workshop on OWL: Experiences and Directions (OWLED 2010)*, San Francisco, CA, 2010; CEUR Workshop Proceedings: 2010.
28. Thomas, D. G.; Papu, R. V.; Baker, N. A. *J. Biomed. Inf.* **2011**, *44*, 59–74.
29. Hastings, J.; Chepelev, L.; Willighagen, E.; Adams, N.; Steinbeck, C.; Dumontier, M. *PLOS One* **2011**, doi: 10.1371/journal.pone.0025513.
30. Chemical Methods Ontology project home. https://code.google.com/p/rsc-cmo/, accessed on 2014-05-09.
31. *IUPAC Compendium of Analytical Nomenclature*, 2nd ed.; Blackwell Scientific Publications: Oxford, 1987.
32. Kopka, J.; Schauer, N.; Krueger, S.; Birkemeyer, C.; Usadel, B.; Bergmüller, E.; Dörmann, P.; Weckwerth, W.; Gibon, Y.; Stitt, M.; Willmitzer, L.; Fernie, A. R.; Steinhauser, D. *Bioinformatics* **2005**, *21*, 1635–1638.
33. Carey, J. S.; Laffan, D.; Thomson, C.; Williams, M. T. *Org. Biomol. Chem.* **2006**, *4*, 2337–2347.
34. The RSC Name Reaction Ontology project home. https://code.google.com/p/rxno/, accessed on 2014-05-09.
35. Batchelor, C. R.; Corbett P. T. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume: Proceedings of the Demo and Poster Sessions*; Association for Computational Linguistics: 2007; pp 45–48.
36. Shotton, D.; Portwin, K.; Klyne, G.; Miles, A. *PLoS Comput. Biol.* **2009**, e1000361.
37. Corbett, P.; Batchelor, C.; Copestake A. In *Proceedings of Building and Evaluating Resources for Biomedical Text Mining at LREC 2008, Marrakech, Morocco*; 2008.
38. Kidd, R. *Integr. Biol.* **2009**, *1*, 293.
39. Pettifer, S.; McDermott, P.; Marsh, J.; Thorne, A.; Villeger, A.; Atwood, T. K. *Learned Publ.* **2011**, *24*, 207–220.
40. Describing Linked Datasets with the VoID Vocabulary. http://www.w3.org/TR/2011/NOTE-void-20110303/, accessed on 2014-05-09.
41. Dataset Descriptions for the Open Pharmacological Space. http://www.openphacts.org/specs/2012/WD-datadesc-20121019/, accessed on 2014-05-09.
42. SKOS Simple Knowledge Organization System Reference. http://www.w3.org/TR/2009/REC-skos-reference-20090818/, accessed on 2014-05-09.

43. Willighagen, E. L.; Waagmeester, A.; Spjuth, O.; Ansell, P.; Williams, A. J. *J. Cheminf.* **2013**, *5*, 23.

44. Donaldson, D. In *Essays on Actions and Events*, 2nd ed.; Oxford University Press: Oxford, 2001.

45. Brinkman, R. R.; et al. *J. Biomed. Semant.* **2010**, *1* (Suppl. 1), S7.

46. Bird, C. L; Frey, J. G. *Chem. Soc. Rev.* **2013**, *42*, 6754–6776.

47. IUPAC Gold Book. http://goldbook.iupac.org/, accessed on 2014-05-09.

48. *Quantities, Units and Symbols in Physical Chemistry*, 3rd ed.; Royal Society of Chemistry: Cambridge, 2007.