ICBO: International Conference on Biomedical Ontology
July 28-30, 2011 · Buffalo, NY, USA

What's in an 'is about' Link? Chemical Diagrams and the Information Artifact Ontology

Janna Hastings[1,2], Colin Batchelor[3], Fabian Neuhaus[4,5], Christoph Steinbeck[1]

[1]Chemoinformatics and Metabolism, European Bioinformatics Institute, Cambridge, UK
[2]Swiss Center for Affective Sciences, University of Geneva, Switzerland
[3]Informatics, Royal Society of Chemistry, Cambridge, UK
[4]National Institute of Standards and Technology, Gaithersburg, MD, USA
[5]University of Maryland Baltimore County, MD, USA

Abstract. The Information Artifact Ontology is an ontology in the domain of information entities. Core to the definition of what it is to be an information entity is the claim that an information entity must be 'about' something, which is encoded in an axiom expressing that all information entities are about some entity. This axiom comes into conflict with ontological realism, since many information entities seem to be about non-existing entities, such as hypothetical molecules. We discuss this problem in the context of diagrams of molecules, a kind of information entity pervasively used throughout computational chemistry. We then propose a solution that recognizes that information entities such as diagrams are expressions of diagrammatic languages. In so doing, we not only address the problem of classifying diagrams that seem to be about non-existing entities but also allow a more sophisticated categorisation of information entities.

Introduction

As the importance of ontology in biomedicine grows, the attention of ontologists is being pressed to the tasks of disambiguation of domain terminology and clarification of underlying hierarchies and relationships in an ever-wider network of interrelated domains [2, 10]. Some issues are emerging as similarly problematic in many of these different domains. One such is the clear definition and distinction of foundational types such as *processes* and *dispositions* [1]. Another is the confusion between *information entities*, such as computer simulations, models and diagrams, and the entities that they are models and diagrams *of*. It is to this latter problem that we turn in this paper.

Chemical graphs are the molecular models that are used throughout chemistry to succinctly describe chemical entities and allow for computational manipulations [12, 6]. Chemical graphs are typically depicted graphically as schematic illustrations – chemical diagrams. Chemical graphs and chemical diagrams are examples of information entities in the chemical domain, and their use has become so pervasive that language used by chemists to refer to chemicals regularly interchanges words for information (such as 'graph') with words for actual chemicals [6].

The Information Artifact Ontology (IAO) [8] is an ontology being developed for the domain of information entities of relevance in biomedicine. The fundamental criterion by which information entities are defined and categorised in the IAO is their *aboutness*, that is, the types of entities that they are *about*. A diagram illustrating the chemical structure of caffeine molecules, for example, is about the class of caffeine molecules. While in this case the chemical diagram corresponds to something in reality (caffeine molecules), there are many other useful and scientifically relevant chemical diagrams that are not about something that exists. Thus, these chemical graphs are not information entities as currently defined in IAO. A similar scenario applies to many other models used in biomedicine, for example pathway diagrams and the mathematical models used in quantitative systems biology. Using chemical diagrams as examples, we will argue that information entities in IAO are defined too narrowly. Since information entities may not necessarily be about something, they cannot be categorized merely by what they are about. But, as we will argue, they should rather be categorised by what sort of information entities they are in their own right.

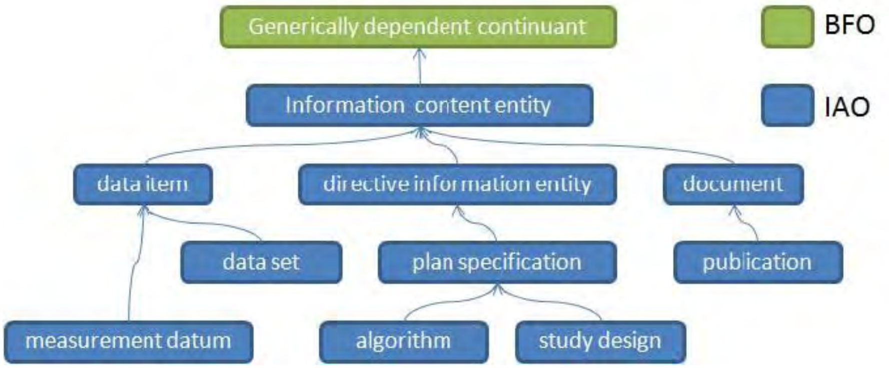


Figure 1. An overview of the Information Artifact Ontology

The remainder of this paper proceeds as follows. In the next section we briefly describe the IAO and the theory of chemical graphs and their related diagrams. Thereafter, we highlight the insufficiency of aboutness in defining types of diagrams. We go on to introduce some semantics for the representation relationship between chemical diagrams and chemical entities; and finally, we propose a modified approach to information ontology that is free of the problems with the current approach.

1 Background

1.1 The Information Artifact Ontology

The Information Artifact Ontology (IAO) [8] is an ontology of information entities being developed in the context of the Open Biological and Biomedical Ontologies (OBO) Foundry [9], beneath the upper level ontology Basic Formal Ontology (BFO) [11, 5]. Within this context, information entities are defined as:

Definition 1. *An information content entity* (ICE) *is an entity that is generically dependent on some artifact and stands in the relation of* aboutness *to some entity.*

The generic dependence on an artifact (i.e., a human creation) in the above definition restricts the scope of the domain to human-created information entities. The 'generic' part of the dependence captures the intuition that information can be copied, that is, reproduced in multiple bearers, in a way that hair colour, for example, cannot. The textual definition also refers to a relation of 'aboutness', which is further supplemented by the axiom:

ICE `subClassOf` **is about** `some` *Entity* (1)

The above is given in the Manchester Web Ontology Language (OWL) syntax, in which the existential quantification (∃) is expressed using the infix some operator. This should not, however, obscure the strong existential dependency claimed, namely: for every *ICE*, there exists some entity to which the *ICE* is related by the **is about** relationship.

A hierarchical overview of the IAO together with some examples of information content entities (*ICE*s) is illustrated in Figure 1.

1.2 Chemical Graphs and Diagrams

The principal object of graph theory is a graph, which consists of a set of objects and the binary relations between them. Graph theory has found many applications in chemistry and is used to represent molecular entities through the molecular graph. These graphs represent the constitution of a molecule in terms of nodes (usually atoms, but in some cases groups of atoms) and edges (chemical bonds) [12].

For the purposes of this paper we define chemical graphs as follows[1] .

Definition 2. A chemical graph, *denoted* CG, *is a tuple* (*V*, *E*) *in which each vertex* $i \in V$ *corresponds to an atom in a molecule; and each undirected edge* $\{i, j\} \in E$ *corresponds to a chemical bond between the atoms i and j.*

These *CG*s are based on the valence bond model of quantum mechanics [7]. For many of the molecules most relevant to the pharmaceutical industry this model reasonably accurately represents (1) by atoms, those

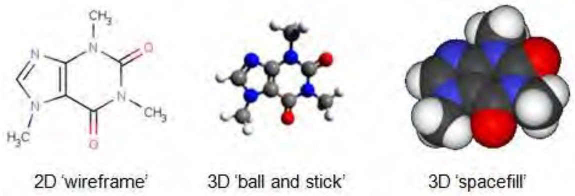
[1] We ignore additional complexity such as the representation of stereochemistry.

portions of the molecules that chemists associate with particular atoms, and (2) by bonds, those portions of the molecules that have high electron probability density. Cheminformatics software uses these to make useful predictions about the chemical properties of a molecule so represented and the physical properties of an ensemble of those molecules. They also enable the schematic representation of molecules in diagrams.

Definition 3. A chemical diagram, *denoted* CD, *is a diagrammatic illustration of the information encoded in a* CG, *which follows an agreed* diagrammatic syntax *for the representation of the graph information.*

Some examples of *CD*s are illustrated in Figure 2. In the 2D wireframe depiction, the diagrammatic syntax used specifies that the *CD* corresponds to the *CG* in that, for each edge $\{i, j\} \in E$ there is a corresponding line, and for each vertex $i \in V$ there is a corresponding *corner* or *line ending* in the *CD*. In the 3D ball and stick diagram, edges are illustrated with lines while vertices are illustrated with coloured, labelled spheres. In the 3D spacefill diagram, vertices are illustrated with large coloured spheres. Both the colours and the radii of the sphere are arbitrary – atoms are much too small to have colours, but the radii are based on experimental averages and are an approximation to the actual molecular structure.

Notice that there is not a one-to-one correspondence between *CD*s and *CG*s, since the same *CG* can be illustrated in many different *CD*s, obeying different syntaxes.



2D 'wireframe' 3D 'ball and stick' 3D 'spacefill'

Figure 2. Some examples of *CD*s for the molecule caffeine

*CD*s, like maps, represent *spatial* information. Let us call spatial representations such as street maps, chemical diagrams, and engineering design models *structural diagrams* and, to a first approximation, assume that they have a direct structural association with a portion of reality, which they are intended to represent.

Definition 4. *A structural diagram* (SD) *is a diagrammatic representation of spatial aspects, such as position, topology and connection, of a structured portion of reality*.

This definition, however, does not suffice, for reasons that will be described in the following section.

2 When 'is about' isn't Enough

The agreed syntax of *CD*s allows their informational content to be reliably understood by all members of the community who use them for exchange of such information.

The agreed syntax also allows for the depiction of molecules, which are

1. Planned, in that the representation is used as a precursor to a synthesis procedure expected to produce a corresponding molecule instance.
2. Hypothesised, in that the representation corresponds to a molecule class for which it is not known whether corresponding instances exist.
3. Chemically infeasible, in that it is known that the representation illustrates a class of molecules for which no instances can exist for a measurable duration of time under normal conditions.
4. Impossible, in that the representation cannot be the structure of any molecule instances, since it violates the rules of molecular compositionality.

In the first two cases the *CD* might or might not be *about* molecules that exist. In the third case chemists expect, and in the fourth case they are certain, that the aboutness criterion of the IAO is violated. Nevertheless, these *CD*s are used by chemists to communicate and exchange information in the same ways as *CD*s that are known to correspond to something in reality. Thus, the way *CD*s are used does not justify treating only a subset of them as information entities. It also indicates that Definition 4 is not along the right lines.

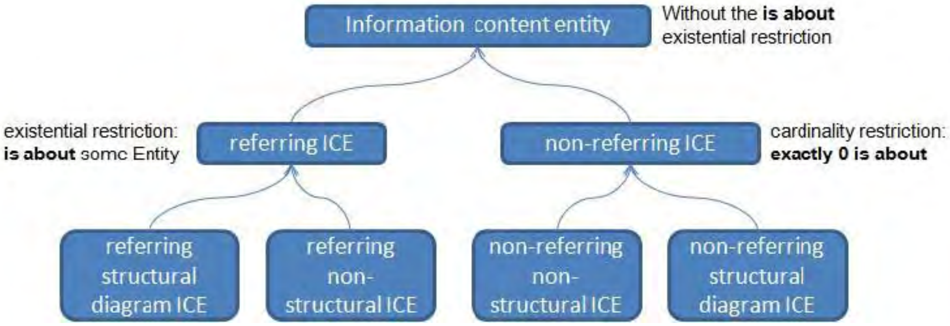

Information content entity
Without the **is about** existential restriction
existential restriction: **is about** somc Entity
referring ICE
non-referring ICE
cardinality restriction: **exactly 0 is about**
referring structural diagram ICE
referring non-structural ICE
non-referring non-structural ICE
non-referring structural diagram ICE

Figure 3. Referring and non-referring information entities in the IAO

A conceptualist resolution to this issue might defend a view of ontology as containing representations of *concepts*, and thereby not be required to differentiate between chemical diagrams for real or impossible molecules, or differentiate at the level of metadata only [4]. However, this seems to overlook the fundamental distinction between these cases, one that chemists recognise. Another strategy for addressing this problem is provided by Ceusters and Smith [3] who distinguish between *referring* and *non-referring* representational units in the context of a mental representation. The application of this distinction to an ontology of SDs beneath IAO is illustrated in Figure 3.

One obvious problem with this approach is that it leads to a massive level of parallel maintenance, since most types of *ICE* can appear twice in the ontology. A more fundamental objection is that this approach violates the fundamental design principles of BFO: categorization according to *ontological nature*, which does not change. For example, it is impossible for a tree (an independent continuant) to become a temporal region, or for a smile (a dependent continuant) to become a soccer game (an occurrent). However, according to the approach in [3] a *CD* might be a non-referring *ICE* now, but become a referring *ICE* tomorrow, because somewhere in some lab somebody accidentally synthesized the corresponding molecule. Thus, in contrast to the other ontological categories in BFO, it would be possible for non-referring *ICE*s to change their ontological nature. Even worse, the ontological nature of *CD*s would be affected by events that had no causal connection to the *CD* and did not change its structure in any way. Since the ontological nature of an entity is not affected by Cambridge changes, that is to say changes only in its description, we conclude that 'non-referring *ICE*' and 'referring *ICE*' are not true ontological categories.

In summary, we agree with Ceusters and Smith that non-referring *ICE*s are *ICE*s. However, we reject the idea that the distinction between referring and non-referring should be the primary basis for classifying *ICE*s. There are some *ICE*s that are necessarily about something (e.g., photographs). But structural diagrams are information entities in virtue of the fact that they are well-formed *expressions* in a *diagrammatic language*. For each type of SD, there is a vocabulary (the symbols and icons that are used in diagrams of that type), a grammar that regulates how the elements of the vocabulary can be combined, and *compositionality* in the sense that the semantics of a complex expression is determined by the semantics of its components and the way these components are arranged.

The elements of the vocabulary of the diagrammatic language do need to correspond to something existing, otherwise the diagrams will not be scientifically relevant. However, not all combinations of the vocabulary that are permissible by the grammar will correspond to something in reality. It would seem strange indeed, on giving an ontological account of natural language, to divide all sentences into those that are about facts and those that are not. "Submariners love periscopes." is a declarative sentence with a transitive verb regardless of whether it is a fact that submariners love periscopes. The same is true for expressions of diagrammatic languages.

3 The Ontology of Structural Diagrams

Different types of *CD* (such as 2D wireframe, 3D ball and stick) obey different diagrammatic syntaxes. What is essential to distinguish different types of diagrams is thus to provide a definition for these syntaxes.

Definition 5. *A* diagrammatic language $L_D = <V, G>$ *is an ordered pair that consists of the vocabulary V (a set of icons and symbols) and a syntax G of composition rules.*

Definition 6. *An* interpreted diagrammatic language *is a quadruple $IL_D = <V, G, T, \varphi>$ such that $<V, G>$ is a diagrammatic language, T is a set of types that is partitioned set of independent continuants IT and dependent continuants DT, and φ is a function that maps the elements from V onto T .*

Definition 7. *Let IL_D be an intepreted diagrammatic language as above, and let D be a well-formed expression in L_D (i.e., a diagram). D is a* structural diagram *that* **is about** *an entity x iff there is some injective interpretation function ι such that*:

- *for each element of V and each token t of V that is part of D, $\iota(t)$ is an instance of $\varphi(V)$*
- *for two tokens t_1, t_2 that are part of D and $\iota(t_1)$, $\iota(t_2)$ are instances of elements of IT : t_1 is connected to t_2 iff $\iota(t_1)$ is connected to $\iota(t_2)$*
- *for all tokens t, t_1, ... t_n: if $\iota(t)$ is an instance of some element of DT and t_1 ... t_n are all connected to t, then $\iota(t)$ inheres in $\iota(t_1)$ ... $\iota(t_n)$.*
- *there is no part y of x such that y is an instance of some type in T and for all t that are part of D there is no $\iota(t)= y$.*

Chemical diagrams of hypothetical molecules that do not exist are not about anything, but they are still well-formed expressions of an interpreted diagrammatic language. For example, the vocabulary V of the 3D ball and stick language consists of colored spheres and lines. The syntax G describes how these elements can be combined to diagrams. The set IT consists of types of atoms, the set DP consists of the types of chemical bonds that connect atoms within a molecule. The function $\emptyset$ maps the color-coded balls to types of atoms and the links to types of bonds. The second diagram in Figure 2 is a structural diagram of a given instance of a caffeine molecule x, since it is possible to map the spheres of the diagram to the atoms that are part of x and the links of the diagram to the chemical bonds of x such that the connections in the diagrams corresponds to the chemical reality in the molecule. Conversely, if the diagram contains a link that does not correspond to a bond in a given molecule x or if it contains a sphere that is mapped to a type of atoms that do not occur as part of x, then the diagram does not represent x.[2]

To place *SD*s (and therefore *CD*s) as subtypes of IAO's *ICE*, we need to change the fundamental aboutness criterion from Equation (1) to a *value* rather than *existential* restriction:

ICE `subClassOf` **is about** `only` *Entity* (2)

This restriction no longer expresses an existential dependence. Rather, it now has the effect that *if* there is some entity that the *ICE* is about, *then* it must be of the required type to avoid a logical inconsistency. Note that this formula expresses a schema, which will be made more precise for different types of *ICE*. With the inclusion of **conforms to** axioms to relate the *ICE* to the L_D, we are now in a position to provide a better definition for SDs and CDs to replace Definition 4:

SD `subClassOf`*ICE* `and` **is about** `only` *StructuredEntity*
`and` **conforms to** `some` *DiagrammaticLanguage*

CD `subClassOf` *SD* `and` **is about** `only` *MolecularEntity*

We can safely include in the resulting ontology, illustrated in Figure 4, diagrams of planned, hypothetical, infeasible, and impossible molecules.

[2] The second clause of definition 7 is irrelevant in the case of CDs, because in CDs tokens of symbols for independent continuants (the atoms) are always connected by tokens of symbols for dependent continuants (the bonds). However, definition 7 is also intended to be applicable to diagrams where symbols for independent continuants might be connected directly; for example architectural drawings and engineering blueprints.

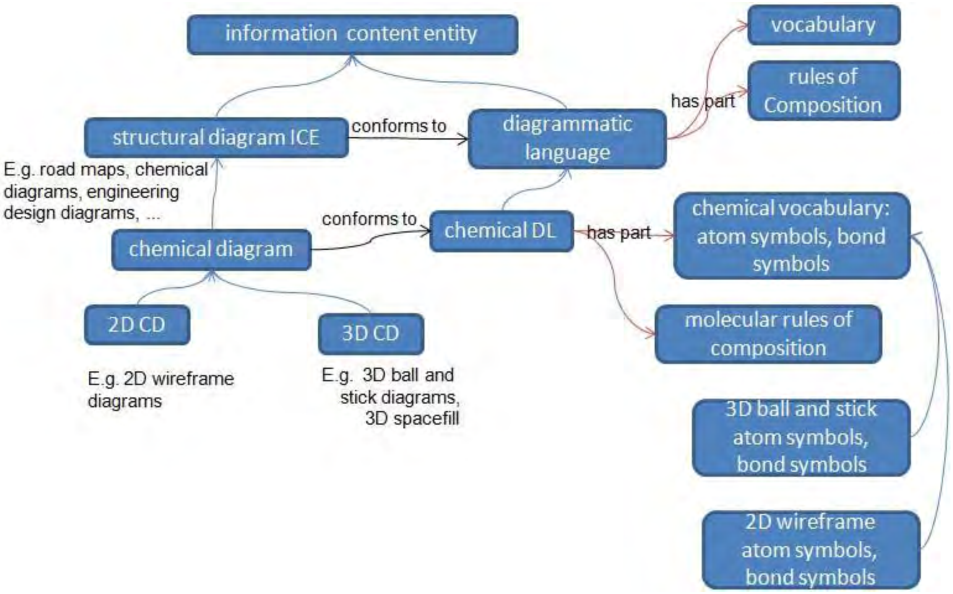

information content entity
structural diagram ICE
conforms to
diagrammatic language
has part
vocabulary
rules of Composition
E.g. road maps, chemical diagrams, engineering design diagrams, ...
chemical diagram
conforms to
chemical DL
has part
chemical vocabulary: atom symbols, bond symbols
molecular rules of composition
2D CD
3D CD
E.g. 2D wireframe diagrams
E.g. 3D ball and stick diagrams, 3D spacefill
3D ball and stick atom symbols, bond symbols
2D wireframe atom symbols, bond symbols

Figure 4. The ontology of chemical diagrams with distinctions for different syntaxes

Now, we can define different types of chemical diagrams regardless of their aboutness, and furthermore express the difference between different *types* of diagrams that are about the same entity (such as 2D and 3D diagrams of caffeine molecules). However, we can go one step further and define a *relationship* between 2D and 3D depictions of the same molecule.

Definition 8. *Let L_1, L_2 be two interpreted diagrammatic languages. Let Θ_1 be a non- empty set of all well-formed expressions of L_1, such that there is at least one diagram D in Θ_1 and one entity x, such that D is about x in L_1. Let Θ_2 be a non-empty set of all well-formed expressions of L_2, such that there is at least one diagram D in Θ_2 and one entity x, such that D is about x in L_2.*

The function m is a coarsening *from Θ_1 (in L_1) to Θ_2 (in L_2) iff*

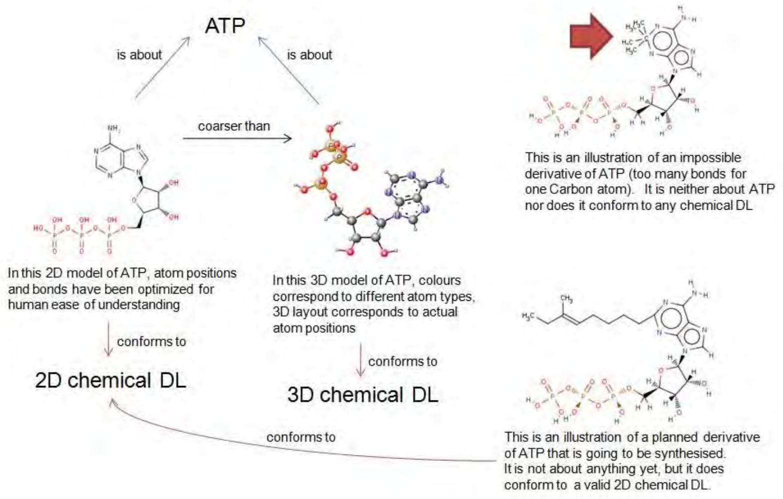
- *m is a function from Θ_1 onto Θ_2; and*
- *for all diagrams D in Θ_1 and all entities x: if D is about x in L_1, then m(D) is about x in L_2; and*
- *for all diagrams D_2 in Θ_2 and all entities x: if D_2 is about x in L_2, then there is a diagram D such that D is about x and m(D) = D_2.*

Coarsening functions map between two different diagrammatic languages, such that if a diagram in one language represents an entity, then it is possible to construct a diagram in the other language that also represents the entity. Typically, coarsening functions are *directed* from a greater to a lesser level of detail; that is, it is possible to map diagrams in a more detailed language to a diagram in a coarser language, but not the reverse. Coarsening functions allow us to define a relationship **coarser than** between *SD*s.

Definition 9. *Let D_1 and D_2 be diagrams conforming to languages L_1 and L_2, respectively. D_2 is* coarser *than D_1 iff*

- *there exists a function m and sets of diagrams Θ_1, Θ_2 of L_1 and L_2, respectively, such that m is a coarsening from Θ_1 (in L_1) to Θ_2 (in L_2) and m(D_1)= D_2; and*
- *there is no function m' such that m' is a coarsening from D_2 (in L_2) to D_1 (in L_1).*

This is illustrated in Figure 5.



In this 2D model of ATP, atom positions and bonds have been optimized for human ease of understanding

In this 3D model of ATP, colours correspond to different atom types, 3D layout corresponds to actual atom positions

This is an illustration of an impossible derivative of ATP (too many bonds for one Carbon atom). It is neither about ATP nor does it conform to any chemical DL

This is an illustration of a planned derivative of ATP that is going to be synthesised. It is not about anything yet, but it does conform to a valid 2D chemical DL.

Figure 5. Some examples of chemical diagrams and their relationships

4 Conclusion

We have argued that the **is about** relationship is not enough to define *CD*s, for two reasons. Firstly, given the possibility of having several different *CD*s corresponding to the same molecule, we see that distinguishing between different types of diagrams, which obey different representational syntaxes, is not possible using only distinctions in what the diagram **is about**. Secondly, a challenge is posed in that *CD*s may be used validly to illustrate classes of molecules *for which no instances exist*. The existential dependency expressed in IAO means that the IAO cannot, in its present form, allow for the inclusion of such non-referring information entities.

We evaluated an approach based on parallel maintenance of IAO hierarchies with differing **is about** commitment. While such parallel maintenance may be a scientifically-valid strategy in some scenarios, it is unable to express the fact that the same representational formalism (i.e., diagrammatic syntax) is used across the hierarchies. Of course, the diagrammatic syntax, if it is to be scientifically-valid, must *typically* represent entities which do exist. But the syntax allows for compositionality and it would be absurd to require the existence of instances for all the complex expressions obtained by composing the elements of the representational vocabulary.

We therefore propose the definition of structural diagrams such as chemical diagrams based on their syntaxes. Any diagram expressed in an interpreted diagrammatic syntax is a valid information content entity regardless of the existence of instances that the diagram **is about**; although the existence of such an instance may be an interesting property depending on the application scenario.

References

1. Batchelor, C., Hastings, J., Steinbeck, C.: Ontological dependence, dispositions and institutional reality in chemistry. In: Galton, A., Mizoguchi, R. (eds.) Proceedings of the 6th Formal Ontology in Information Systems conference. Toronto, Canada (2010)
2. Bodenreider, O., Stevens, R.: Bio-ontologies: current trends and future directions. Briefings in Bioinformatics 7(3), 256–274 (2006)
3. Ceusters, W., Smith, B.: Foundations for a realist ontology of mental disease. Journal of Biomedical Semantics 1(1), 10 (2010)
4. Dumontier, M., Hoehndorf, R.: Scientific realism. In: Galton, A., Mizoguchi, R. (eds.) Proceedings of the 6th Formal Ontology in Information Systems conference. Toronto, Canada (2010)
5. Grenon, P., Smith, B., Goldberg, L.: Biodynamic ontology: Applying BFO in the biomedical domain. In: Stud. Health Technol. Inform. pp. 20–38. IOS Press (2004)
6. Hastings, J., Batchelor, C., Steinbeck, C., Schulz, S.: What are chemical structures and their relations? In: Galton, A., Mizoguchi, R. (eds.) Proceedings of the 6th Formal Ontology in Information Systems conference. Toronto, Canada (2010)
7. Pauling, L.: The shared-electron chemical bond. Proc. Natl. Acad. Sci. USA 14, 359–362 (1928)

8. Ruttenburg, A., Courtot, M., The IAO Community: The Information Artifact Ontology (2010), http://code.google.com/p/information-artifact-ontology/

9. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., The OBI Consortium, Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.A., Scheuermann, R.H., Shah, N., Whetzel, P.L., Lewis, S.: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 25(11), 1251–1255 (Nov 2007)

10. Smith, B., Ceusters, W.: Ontological realism as a methodology for coordinated evolution of scientific ontologies. Applied Ontology 5, 139–188 (2010)

11. Smith, B., Grenon, P.: The cornucopia of formal ontological relations. Dialectica 58, 279–296 (2004)

12. Trinajstic, N.: Chemical graph theory. CRC Press, Florida, USA (1992)