# Thinking 'Inside the Box' with Description Logics

**by Mike Bergman - Monday, November 10, 2008**

http://www.mkbergman.com/466/thinking-inside-the-box-with-description-logics/

## Linked Data Need Not Rediscover the Past; A Surprise in Every Box

A standard cliché of management consultants is the exhortation to think "outside the box." Of course, what is meant by this is to question assumptions, to think differently, to look at problems from new perspectives.

With our recent release of the (linked open data) 'LOD constellation' of linked data classes based around UMBEL, I have been fielding a lot of inquiries on what the relationship is of UMBEL to DBpedia. (See, for example, this current interview by the Semantic Web Company with me and Sören Auer of the DBpedia project.) This also fits into the ongoing distinction we have made in the UMBEL project between our subject concepts (classes) and named entities (instances).

What has actually most been helping my thinking is to get fully *inside the box* (or, rather, boxes, hehe). Let me explain.

*The problem with urging outside-the-box thinking is that many of us do a less-than-stellar job of thinking inside the box. We often fail to realize the options and opportunities that are blatantly visible inside the box that could dramatically improve our chances of success.*

Naomi Karten [1]

## The Description Logics Underpinnings of the Semantic Web

Description logics are one of the key underpinnings to the semantic Web. They grew out of earlier frame -based logic systems from Marvin Minsky and also semantic networks; the term and discipline was first given definition in the 1980s by Ron Brachman, among many others [2].

Description logics (DL, most often expressed in the plural) are a logic semantics for knowledge representation (KR) systems based on first-order predicate logic (FOL). They are a kind of logical metalanguage that can help describe and determine (with various logic tests) the consistency, decidability and inferencing power of a given KR language. The semantic Web ontology languages, OWL Lite and OWL DL (which stands for description logics), are based on DL and were themselves outgrowths of earlier DL languages.

Description logics and their semantics traditionally split *concepts* and their relationships from the different treatment of *individuals* and their attributes and roles, expressed as fact assertions. The concept split is known as the TBox (for *terminological* knowledge, the basis for *T* in *TBox*) and represents the schema or taxonomy of the domain at hand. The TBox is the structural and intensional component of conceptual relationships.

*Thus, the model is an abstraction of a concrete world where the concepts are interpreted as subsets of the domain as required by the TBox and where the membership of the individuals to concepts and their relationships with one another in terms of roles respect the assertions in the ABox.*

Franz Baader and Werner Nutt [3]

The second split of individuals is known as the ABox (for *assertions*, the basis for *A* in *ABox*) and describes the attributes of individuals, the roles between individuals, and other assertions about individuals regarding their class membership with the TBox concepts. Both the TBox and ABox are consistent with set-theoretic principles.

TBox and ABox logic operations differ and their purposes differ. TBox operations are based more on inferencing and tracing or verifying class memberships in the hierarchy (that is, the structural placement or relation of objects in the structure). ABox operations are more rule-based and govern fact checking, instance checking, consistency checking, and the like [3]. ABox reasoning is generally more complex and at a larger scale than that for the TBox.

Early semantic Web systems tended to be very diligent about maintaining these "box" distinctions of purpose, logic and treatment. One might argue, as I do herein, that the usefulness and basis for these splits has been lost somewhat in our first implementations and publishing of linked data systems.

## ABox and TBox Analogs in the Linked Data Web

Most of the semantic Web work at the beginning of this decade was pretty explicit about references to description logics and related inferencing engines and computational efficiency. Some of the early

commercial semantic Web vendors are still very much focused on this space.

However, with the first release and emphasis on linked data about two years ago, the emphasis seemed to shift to the more pragmatic questions of actually posting and getting data out there. Best practices for cool URIs and publishing and linkage modes assumed prominence. The linking open data (LOD) movement began in earnest and gained mindshare. Of course, many in the DL and OWL development communities continued to discuss logic and inferencing, but now seemingly more as a separate camp to which the linked data tribe paid little heed.

The central hub of this linked data effort has been DBpedia and its pivotal place within the 'LOD cloud.' What is remarkable about the LOD cloud, however, is that it is almost entirely an ABox representation of the world and its instances. Starting from the core set of individual instances within Wikipedia, this cloud has now grown to many other sources and the central place for finding linked instance data. If one looks carefully at the LOD cloud and its linkages we can see the prevalence of instance-level relationships and attributes.

Linking Open Data's "*ABox*"


Linking Open Data's *"TBox"*

In fact, the LOD cloud diagram to upper right from the Wikipedia article on linked data has become the key visual metaphor for the movement. But, as noted, this view is almost exclusively one at the ABox instance level.

The UMBEL project began at roughly the same time and as a response to the release of DBpedia. My question in looking at the first data linked to DBpedia was, What is this content *about*? Sure, I might be able to find multiple records discussing Abraham Lincoln as a US president regarding attributes like birth date and a list of children, but where could I retrieve records about other presidents or, more broadly, other types of leaders such as prime ministers, kings or dictators?

The intuition was that the linked data and the various FOAF and other distributed instance records it was combining lacked a coherent reference structure of subject topics or concepts with which to describe content. The further intuition was that -- while tagging systems and folksonomies would allow any and all users to describe this content with their own metadata -- a framework for relating these various assignments to one another was still lacking.

In the nearly two years of development leading to the first beta release of UMBEL we have tried many analogies and metaphors to describe the basis of the 20,000 subject concept classes within UMBEL in relation to its role and other linked data initiatives. While many of those metaphors help visualize use and role, the more formal basis offered by description logics actually helps to most precisely cast UMBEL's

role. For example, in today's interview with the Semantic Web Company, I note:

*". . . we have described UMBEL as a roadmap, or middleware, or a backbone, or a concept ontology, or an infocline, or a meta layer for metadata, and others. Today, what I tend to use, particularly in reference to DBpedia, is the TBox-ABox distinction in computer science and description logics. UMBEL is more of a class or structural and concept relationships schema -- a TBox -- while DBpedia is more of an an instance and entity layer with attributes -- an ABox. I think they are pretty complementary. . . "*

The resulting class level structure produced by UMBEL and its mappings to other classes within existing linked data enabled us to create and then publish the 'LOD constellation', a complementary TBox structure to the linked data's existing ABox one. This diagram to the lower right from the Wikipedia article on linked data now shows this complement.

## Completeness and Sufficiency

Description logics have arisen to aid our creating and understanding of knowledge representation systems. From this basis, we can see that the first efforts of the linked data initiative have lacked context, the TBox. At a specific level, the question is not DBpedia *v.* UMBEL or cloud *v.* constellation. Both types of structure are required in order to complete the logical framework. By thinking inside the box -- by paying attention to our logical underpinnings -- we can see that both TBoxes and ABoxes are essential and complementary to creating a useful knowledge representation system.

By more explicitly adopting a description logics framework we can also better address many prior questions of context, coherence and sufficiency. These have been constant themes in my recent writings that I will be revisiting again through the helpful prism of formal description logics.

My interview today with Sören Auer also brought up some important points regarding context. As we have said in other venues, it is important that any TBox be available for context purposes. Whether that should be UMBEL or some other framework depends on the use case. As I noted in the interview, "UMBEL's specific purpose is to provide a coherent framework for serious knowledge engineers looking to federate data." Other uses may warrant other frameworks, and certainly not always UMBEL.

But, in any event, I have two cautions to the linked data community: 1) do not take the suggestion to have a reference framework of concepts as being equivalent to adopting a single ontology for the Web; think of any reference structure as an essential missing TBox, and not some call to adopt "one ontology to rule them all," but 2) in adopting alternative frameworks, take care that whatever is designed or adopted itself be able to meet basic DL logic tests of consistency and coherence.

## A Serendipitous Surprise

No one has yet elaborated the significant advantages from design, performance, architectural and flexibility perspectives from a distinct and explicit separation of TBox from ABox -- but they're there!

The many advantages from separate TBox and ABox frameworks are one serendipitous surprise coming from the early development of linked data. To my knowledge, no one has yet elaborated the significant advantages from design, performance, architectural and flexibility perspectives from a distinct and explicit separation of TBox from ABox. We believe these advantages to be substantial.

Realize, as distributed, UMBEL already has both TBox and ABox components. The TBox component is the lightweight UMBEL ontology, with its 20,000 subject concept classes and their hierarchical and other relationships. This component has a vocabulary (or terminology) for aiding the linking to external ontologies. The vocabulary is quite suitable for extension into new domains as well.

The ABox component is the named entities part of instances drawn from Wikipedia and the BBC's John Peel sessions. Besides being of common, broad interest, these 1.5 million instances (per the current version) are included in the distribution to instantiate the ontology for demonstration and sandbox purposes.

So, UMBEL's world is quite simple: subject concepts (SCs) and named entities (NEs). Subject concepts are the TBox and classes that define the structure and concept relationships. Named entities are the individual "things" in the world (some lower case such as animals or foods) and are the ABox of instances that populate this structure.

In our early efforts, we concentrated on the SC portion of UMBEL. Most recently, we have been concentrating on the NE component and its NE dictionaries. It was these investigations that drew us into an ABox perspective when looking at design options. The logic and rationale had been sitting there for some years, but it took cracking open the older textbooks to become reacquainted with it.

Once we again began looking inside the box, we began to see and enumerate some significant advantages to an explicit TBox-ABox design, as well as advantages for keeping these components distinct:

- Easier understood ontologies with a very limited number of predicates
- Lightweight schema design that is easy to extend
- Ability to "triangulate" between separate SC (concept) and NE (instance) disambiguation approaches to improve overall precision and recall
- Attribute information is kept separate from structural and conceptual relationships
- Easy to swap in varied, multiple and private or public named entity dictionaries
- Relatively easy extension of the schema ontology into specific domains
- A design suitable to computation efficiency (rules for ABox; inference and standard reasoning for TBox), and
- Assignment of NEs to distinct and disjoint "super types" [4] that can bring significant tableaux benefits to ABox reasoning.

We are still learning about these advantages and will document them further in pending work on coherence and named entity dictionary (NED) creation.

## Thinking Inside the TBox and ABox

The two main points of this article have been to: 1) recognize the important intellectual legacy of

description logics and how they can inform the linked data enterprise moving forward; and 2) be explicit about the functional and architectural splits of the TBox from the ABox. Making this split brings many advantages.

There will continue to be many design challenges as linked data proliferates and actually begins to play its role of aiding meaningful knowledge work. The grounding in description logics and the use of DL for testing alternative designs and approaches is a powerful addition to our toolkit.

Sometimes there are indeed many benefits to thinking *inside the box*.

---

[1] See http://www.stickyminds.com/sitewide.asp?Function=edetail&ObjectType=COL&ObjectId=8279.

[2] F. Baader, D. Calvanese, D. McGuinness, D. Nardi, and P. F. Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation and Applications.* Cambridge University Press, 2003. See Chapter 1. Sample chapters may be viewed from Enrico Franconi's Description Logics course notes and tutorial at http://www.inf.unibz.it/~franconi/dl/course/, which is an excellent starting reference point on the subject.

[3] *Ibid*.; see Chapter 2.

[4] These are akin to the lexicographer supersenses that have been applied in WordNet for nouns and verbs (though only nouns are used here). See Massimiliano Ciaramita and Mark Johnson, 2003. Supersense Tagging of Unknown Nouns in WordNet, in *Proceedings of the Conf. on Empirical Methods in Natural Language Processing*, pp. 168173, 2003. See http://www.aclweb.org/anthology-new/W/W03/W03-1022.pdf.

---