Saved by the BEL

ringing in a common language for the life sciences

Established fields of study preserve their history and support advancement by developing and using a common language, without which those fields could not progress at a reasonable pace. The so-called 'language of mathematics', for example, provides a universal way for mathematicians everywhere to understand the discipline as it stands today, and to contribute their own work such that it can be understood by their colleagues. Similarly, life sciences researchers have a common language for describing the entities in their domain (genes, species, diseases, etc) and the relationships that exist between them. This language provides the knowledge framework for understanding what is currently known and for reporting new discoveries. Unfortunately, in the life sciences this language is exclusively a spoken natural language and as such not one that can readily be used for the computation, which has become indispensable to research in the face of the vast amounts of disparate and complex data that need to be used. This means that, while researchers can use a common language for life sciences when they talk amongst themselves, there is a serious 'Tower of Babel' problem when using computers to work with the same information. Compounding the problem is the fact that current methods of storing life sciences data were mostly developed more than 40 years ago; they are devoid of semantics and were never designed for retrieval in a way that enhances understanding. Here we describe OpenBEL, comprising the Biological Expression Language (BEL), which was specifically designed to serve as a common, computable (but still human-readable), semantically-rich language for representing scientific findings, together with associated software tools, and we propose its adoption and discuss the many advantages to be gained by doing so across the entire life sciences ecosystem.

By Ted Slater and
Dr Diane H. Song

In the life sciences, especially in large organisations, we've been talking about data silos, and cursing them, for a long, long time. Data silos are anathema in a domain as challenging as ours, where critical, complex information exists in innumerable places around the world in vast amounts and in a never-ending array of formats. We have acknowledged data silos in bioinformatics for at least two decades, held conferences on data integration to try to teach ourselves how to overcome them, and still today whenever we build and deploy a new database in pursuit of our research goals, we almost always deploy yet another data silo.

What's so bad about data silos? Well, a data silo is an information repository (usually a relational database) that is not interoperable with other, though still related information repositories. For example, an organisation may have a database containing information relevant to diseases and the drugs used to treat them, and it may have another one also around drugs, and including their molecular targets and the relevant biomarkers as well. If these two databases are traditional relational databases, such as you're likely to find in any organisation, then they probably contain redundant information: they certainly both contain information about drugs, and they will no doubt both contain information about genes, for example. Furthermore, they will not be connected in any significant way, such that not only will researchers have to figure out how to query both databases in order to find related information, it is likely (especially in a large organisation) that they will not even know about the existence of one or the other or even both of those databases. Figuring out how to query those databases can be a research project all by itself, because it is likely that the database schemas are very different from each other, and they may be so difficult to query that an IT specialist has to be retained to query it on behalf of researchers. Because of these and other reasons, every data silo in an organisation costs money, decreases information management agility and efficiency, and slows down research.

Why do we keep building data silos when we know *a priori* what they are and we know we do not want any more of them? Didn't Einstein have something to say about doing the same thing over and over again while expecting different results? Importantly, the reason is that we typically use traditional relational databases for information repositories. Relational databases have their advantages, particularly when we have a lot of data that does not change and we always know exactly how we want to query them. But relational databases were introduced more than 40 years ago, at a time when computer storage space, memory and processor capacity were in very limited supply and were thus very expensive. In order to optimise data storage and retrieval under those difficult conditions, we stripped the semantics from the data and made them implicit in bespoke relational schemas. We traded the ability to readily understand our data for reduced storage requirements and increased query performance. In short, we began four decades of focusing on the container for the data, rather than on the data themselves.

You can think of a relational schema as a way to express information about a particular domain – it is a way of 'talking about' a particular subject, such as customer bank accounts or diabetes, for example. For organisations involved in the life sciences, such as pharmaceutical companies, it is very often the case that within these domains various databases are 'talking about' many of the same things; a relatively large proportion of them might contain information about human genes, for example. The problem is that relational schemas are usually bespoke: they are custom-made per database. What that means is that every time you create a relational database in your organisation, you have effectively created a brand new language in which to talk about the subject of your database. It is as if your diseases/drugs database is speaking German, and your drugs/targets/biomarkers database is speaking Lakota. If you are not a native speaker, you can eventually (after a certain amount of 'translation' effort) figure out what is being talked about in those databases. But those databases are not going to be talking to each other, or to any other databases, directly. If you want to query both of them at the same time, perhaps so that you can learn what biomarkers are relevant to what drugs, you either need to work through a translator every time, or you need to invent a third language into which to translate that content and then use that language from then on (making sure to keep it up to date). Taken together, then, all the relational databases in your organisation amount to one huge 'Tower of Babel' problem. That is why you have data silos.

If you want to stop building data silos in your organisation and realise the promise of flexible, interoperable, semantically-rich data, you have to shift the focus away from the containers for the data directly to the data themselves. Not to be overly dramatic, but you have to liberate your data from the tyranny of bespoke, traditional relational schemas.

Data liberation with BEL

There are several ways to do this, and the most promising ones today are standards-based efforts that are graph-based. This typically means that assertions in the system take the form of subject-predicate-object triples, for example:

Mary likes tennis

In this example, 'Mary' is the subject, 'likes' is the predicate and 'tennis' is the object of Mary's liking; taken together, these things make a statement about the relationship between Mary and the sport of tennis. This is a very straightforward representation, and in fact it reads much like a sentence in natural language.

One interesting thing about subject-predicate-object triples is that sometimes the object of one triple can be used as the subject of other triples, and *vice versa*. In this way, triples can be assembled into mathematical graphs. Graphs are simply sets of vertices (or 'nodes'), some of which are linked to each other by edges. **Figure 1** is a drawing from Wikipedia (http://en.wikipedia.org/wiki/Graph_(mathematics)) showing a graph with six labelled nodes, having seven edges linking them.

When we talk about graph-based knowledge bases, we mean a knowledge base that comprises a graph of what we know about the entities in some domain of discourse, and the relationships that we know exist between those entities. Often we need to represent directional ('one-way') relationships between entities in graphs – after all, just because Mary likes tennis does not mean that tennis likes her! Such relationships are illustrated by simply using an arrow to show the directionality of the relationship, and the presence of such relationships make the graph a directed graph.

Representing knowledge as graphs has many advantages, particularly in the life sciences. Many people think about the world in terms of things and the relationships those things can (or cannot) have with each other, which is just how graph knowledge bases represent the world. We have been studying graph data structures for almost 300 years, so there is a lot of previous work upon which to build powerful software tools. Graph data structures can easily model hierarchical relationships between concepts. Graphs lend themselves naturally to network visualisation methods, and these visualisations are usually easily interpretable by humans. Furthermore, as we shall see, the topology of graphs provides an excellent substrate for specialised applications, such as traversal-based if-then reasoning over causal relationships and other artificial intelligence applications.

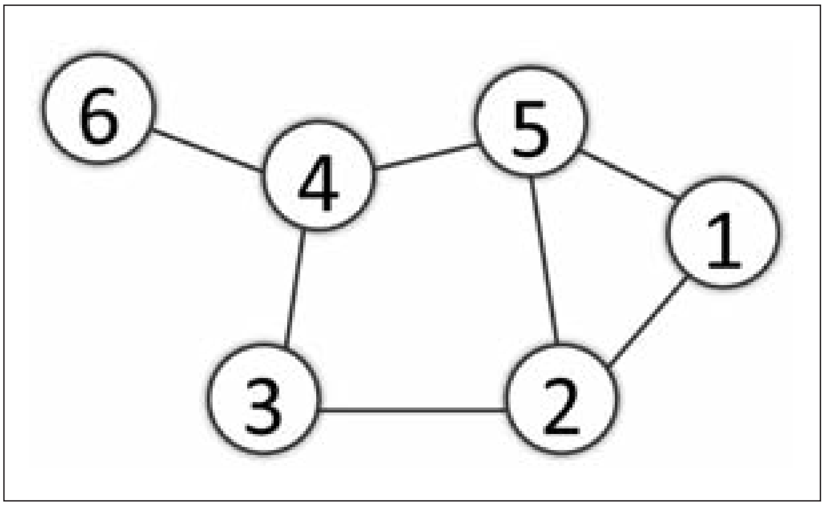

Figure 1
An illustration of a simple graph

There are several sets of standards for graph-based knowledge representation. Prominent among these are the ones put forth by the W3C as part of their Semantic Web initiative: RDF, OWL, SPARQL, and others (http://www.w3.org/2001/sw/). Until recently, however, none of these standards has been particularly usable by researchers in the life sciences, nor have they been sufficiently flexible to represent biology at the multiplicity of scales needed to represent everything from atoms to organisms and beyond. This situation has been remedied by Selventa's release of OpenBEL as an open language standard (BEL) accompanied by a set of open source software tools.

BEL (the Biological Expression Language) is a triple-based language for representing the findings reported in scientific literature. Invented by Dexter Pratt at Selventa (then operating as Genstruct), it has been in continuous development and use for the past 10 years on more than 80 commercial life science projects. It was designed to be human-readable, while being structured enough to be computable and expressive and flexible enough to represent biology in all its wide variety. BEL was specifically designed to capture causal relationships between biological entities as they have been reported in the scientific literature, with the goal of acquiring knowledge with sufficient accuracy and semantics to build knowledge bases suitable not only for querying but also for the powerful visualisation and inference capabilities necessary to support today's life sciences research.

A closer look at BEL

Let's get a better understanding of what BEL is and how it works by looking closely at a BEL statement. A BEL statement is a representation of a particular finding from life sciences literature: a statement of fact that relates two concepts. More often than not, the relationship is a causal one indicating that the subject entity (often a class of molecules) causes some measurable change in the object entity (often a biological process). A typical example of such a causal relationship between entities is, ‘Corticosteroid decreases tissue damage’. **Figure 2** shows what that statement looks like in BEL:

Let’s take this statement apart to understand it better, because there is a bit more going on here than just the original relationship. First, notice that the BEL statement:

```
a(CHEBI:corticosteroid) -| bp(NCI:
        Tissue Damage)
```

is represented in three parts: a term, a relationship and then another term. As explained in the diagram, the first term refers to the abundance of corticosteroid. Abundance is a very important concept in BEL, because it lets you refer to an unspecified amount of a class of item. You can think of it as ‘some amount’, so in this example we are talking about ‘some amount of corticosteroid’.

Abundance here is abbreviated as simply ‘a’ and it is an example of another important aspect of BEL’s functional notation. BEL’s functional notation allows it to represent sophisticated concepts as single terms, and to build these terms from simpler terms when necessary. It allows BEL to be relatively terse, and therefore easy to read. The second term also uses functional notation, but there the function is ‘bp’ which represents ‘biological process’.

Next, notice how the corticosteroid molecule is named in the first term: the name ‘corticosteroid’ is present, but it is prefixed with a namespace (‘CHEBI’) and a colon (‘:’). BEL doesn’t have an ontology of its own, but rather allows you to use whatever external ontologies you would like to use. In this case, the author of the BEL statement has chosen to represent the corticosteroid molecule as it is found in the ChEBI (Chemical Entities of Biological Interest database), and that source is denoted by the namespace value. ‘Tissue damage’ is nominated in the second term in a similar way, this time using the ‘Tissue Damage’ term as it is found in the NCI (National Cancer Institute) vocabulary.

The relationship being represented in this example is simply ‘decreases’, and the shorthand way to write that in BEL is just ‘-|’. If the relationship had been ‘increases’, it would have been abbreviated ‘->’. Both of those relationship types are examples of causal relationships. In this example, we are stating that the presence of some amount of corticosteroid causes there to be less tissue damage.

BEL can express many other functions as well. Besides many different kinds of biological processes (such as diseases), BEL can also represent modifications to molecules (eg, phosphorylation of proteins); transformations (including movement from one physical location to another); and activities (kinase activity, for example). BEL can also represent a rich set of relationships, including correlative relationships, genomic relationships and other relationships that can describe concepts such as biomarkers, or even define a hierarchical arrangement of concepts to form a taxonomy.

In addition to assertions like this, BEL can also capture each statement's context. That is, if the statement above were only reported to be the case in vascular tissue, then that metadata could also be recorded to ensure that the knowledge is interpreted correctly in subsequent analyses. Context can include other concepts as well, such as species, disease state, etc. Such metadata is critical for biological understanding, but because it can be present in a very large variety of forms it is very difficult to capture in a bespoke relational database schema.

The object of a BEL statement may be used as the subject of one or more additional BEL statements and *vice versa*, and so many related BEL statements may be assembled into a thematic graph knowledge base, perhaps containing information relevant to a specific disease. This process of assembly may include such refinements as the addition of inferred statements, or the unification of multiple entity names (**Figure 3**).

The graph structure of the knowledge base readily supports if-then kinds of reasoning, in particular because of the semantic relationships represented by the links between entities. As an example, consider that one triple in the knowledge base might say that Protein A increases the expression of Gene B. If you can then associate an empirical measure of the expression level of Gene B (perhaps from a microarray experiment) with the graph node for Gene B, then you can use simple graph traversal techniques to begin exploring what might be happening with Protein A that might corroborate the idea that Protein A was doing something interesting both in terms of other entities it might be affecting and other entities that might be affecting it. This is straightforward, qualitative reasoning of the kind that human researchers apply when they interpret experimental data.

In this way, BEL can be used to overcome the limitations of traditional relational databases. Importantly, because of its standardisation, expressivity and readability, BEL can finally be the *lingua franca* for life sciences, the common language that will help us solve the Tower of Babel problem we have been fighting for so many years. When organisations begin to represent their life sciences information in BEL format, they will be able not only to make better, faster use of external information, but they will also be able to greatly simplify their own internal knowledge environments. As more and more organisations adopt this standard way of communicating life sciences knowledge, opportunities for sharing and collaborating will present themselves all across the information 'landscape', from content providers to pharmaceutical companies.

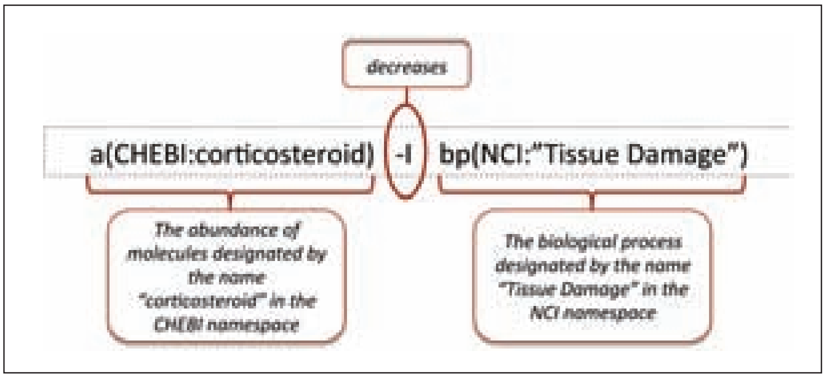
decreases
a(CHEBI:corticosteroid)
-I
bp(NCI:"Tissue Damage")
The abundance of molecules designated by the name "corticosteroid" in the CHEBI namespace
The biological process designated by the name "Tissue Damage" in the NCI namespace

Figure 2
Anatomy of a simple BEL statement

The OpenBEL Consortium

The BEL language standard has been released to the life sciences community by Selventa as an important part of OpenBEL. OpenBEL also includes the BEL Tools, a set of BEL-based applications and APIs that work together to support knowledge acquisition and management, knowledge publishing in the form of graphs called KAMs (Knowledge Assembly Models), and knowledge use through visualisation and other means. The BEL Tools are all available without cost under open source licensing at http://openbel.org/.

Selventa is in the process of forming an independent, non-profit home for OpenBEL, which will be operational by Q4 of 2012. The OpenBEL Consortium will be responsible for driving further global adoption of OpenBEL, guiding future enhancement and expansion of the BEL language standard, helping to increase the availability of BEL-formatted content and BEL-compatible tools, and prioritising BEL Tools development efforts within the Consortium. The BEL Consortium will also be responsible for providing BEL validation tools and other means by which to

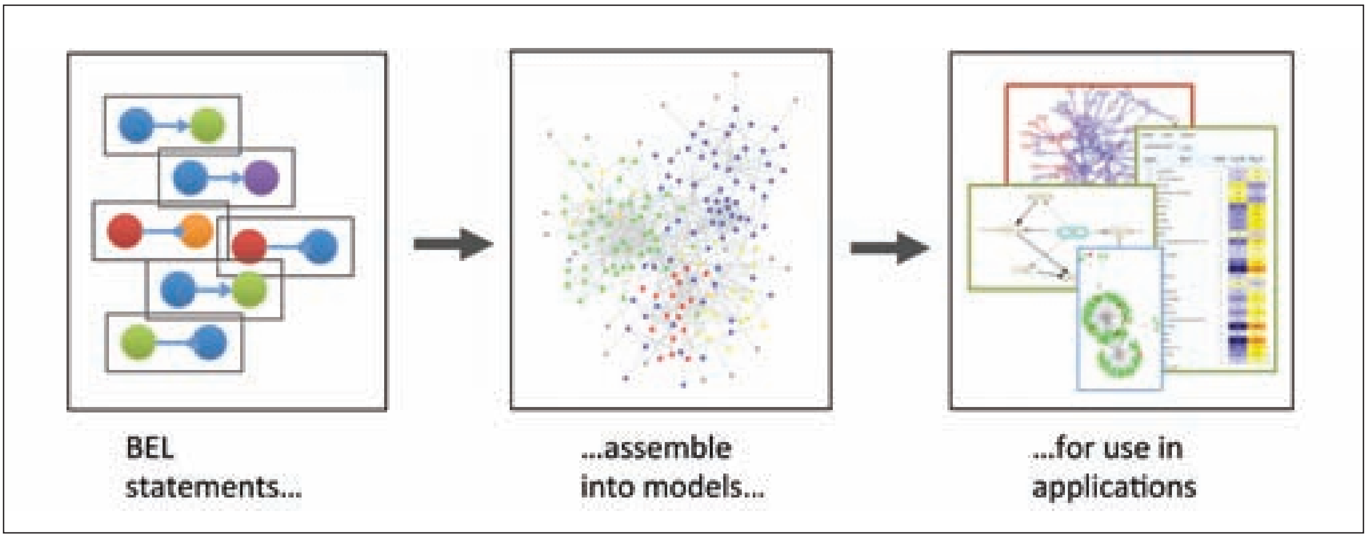

Figure 3
BEL statements can be assembled into graph-based models, which can then be put to many uses

ensure a maximally-valuable cohesive standard for the OpenBEL community at large. The Consortium will also provide training opportunities worldwide.

OpenBEL already has a strong community around it. Like any standards-based initiative, OpenBEL will benefit from further participation. There are a number of ways for readers to get involved.

The OpenBEL portal at http://openbel.org/ is a great place to start, and there you will find a link to https://github.com/OpenBEL, the GitHub site where OpenBEL's documentation and open source software tools are made available.

The OpenBEL Wiki, at http://wiki.openbel.org/display/home/Home, is a public forum for contributions of many kinds, including suggestions for v2.0 of BEL; there you will also find training materials for OpenBEL.

You can visit the OpenBEL Community on LinkedIn, and you can also follow OpenBEL on Twitter at @openbel. DDW

__Ted Slater__ is a Chief Technology Officer, OpenBEL Consortium at Selventa, Inc and an Executive Director at Broad Reach Strategic Advising LLC, based in the Boston area. He is an expert in the application of knowledge- and semantics-based methods to pharmaceutical R&D. He holds an MA in Molecular Biology from the University of California at Riverside and an MS in Computer Science from New Mexico State University.

__Dr Diane H. Song__ is a Director of Market Research at Selventa. Dr Song has more than 15 years of oncology and metabolic disorders-related research experience. Prior to joining Selventa, she was a faculty member at Boston Medical Center, where her research was recognised through numerous grant awards and speaking opportunities. She holds a BS in Chemistry from the University of Michigan and a PhD in Chemistry from Boston University.