OWL: PAX of mind or the AX? Experiences of Using OWL in the Development of BioPAX.

Joanne S. Luciano[1]* and Robert D. Stevens[2]

[1] Department of Genetics, Harvard Medical School, 77 Avenue Louis Pasteur,Boston, MA 02115, USA
jluciano@genetics.med.harvard.edu
[2] School of Computer Science, Manchester University, Oxford Road, Manchester, M13 9PL, UK
robert.stevens@manchester.ac.uk

Abstract. This paper argues that several factors conspired to produce a less than optimal OWL version of BioPAX, describes what those factors were, and presents suggestions about a better way to proceed. The BioPAX workgroup agreed to implement BioPAX in OWL-DL and XML-Schema. OWL has a steep learning, and the difficulties at the time of its adoption by the BioPAX community were compounded by the lack of tutorials and examples, the lack of tools of any quality, and the general lack of experience. More generally, the BioPAX community found it hard to specify a coherent set of requirements (the fact that there were two camps, one in favor of OWL the other XML Schema, reflected this). The use of OWL was not seen as necessary by all members of the community and it required considerably more work initially than using existing known methods.

1 Introduction

This paper argues that several factors (technical, social, financial, ideological) conspired to produce a less than optimal OWL version of BioPAX, describes what those factors were, and presents suggestions about a better way to proceed. The BioPAX workgroup agreed to implement BioPAX in OWL-DL and XML-Schema. With respect to OWL there were number of challenges including the complexity of the language and its syntax, the open world assumption which was foreign people trained in databases, and the logical framework, which again most computer literate scientists have not been trained in. As a result, OWL has a steep learning, and the difficulties at the time of its adoption by the BioPAX community were compounded by the lack of tutorials and examples, the lack of tools of any quality, and the general lack of experience. More generally, the BioPAX community found it hard to specify a coherent set of requirements (the fact that there were two camps, one in favor of OWL the other XML Schema, reflected this). The use of OWL was not seen as necessary by all members of the community and it required considerably more work initially than using existing known methods. There has been a long tradition in bioinformatics of doing only just enough to do the job at hand and so the attitude was "why do more when a small increment already brings a bounty". Finally, we should note that it is human nature to resist the new or unknown, to doing more than is absolutely needed, to spending more time than absolutely necessary.

Thousands of biologists and biomedical researchers all over the world are tasked with the need to understand the inner-workings of biological cells. One way they describe these inner- workings is in terms of their cellular processes. These processes, called 'pathways', are generally one of four different types: metabolic pathways, signaling pathways, gene regulatory networks, and a variety of molecular interactions types that are also conceptualized as networks. Although it is understood that these different pathway types all interact within a living cell, individual pathway databases tend to capture only one of the four types. This of course limits their utility.

The need for a common representation, which the BioPAX initiative sought to address, exists because biological pathway conceptualizations have become the unifying conceptual framework for interpreting the cellular process data that will advance our understanding of biology and help us to achieve our international and interdisciplinary research goals and objectives. Biological pathways have become the place where we add

* We would like to thank Christopher Brewster for help preparing this paper.

our new findings when we discover a new interaction, as well as the places where we look for answers when we try to find a new therapeutic drug target. Much basic scientific research produces pathway data, including environmental research, energy research, genetic and clinical research, and virtually all of life science research today. At some point, in all of these areas, the question is asked, "What pathways are involved?" As this question is fundamental today, it is important to provide a mechanism for access and reuse to these data and thus enable it to have broad impact for science.

he major problem for researchers who use pathway databases has been that the representations of pathway data within these resources are not consistent or interchangeable. This is true both of the conceptual framework and of the details of encoding. For example, interactions in signaling pathways are described in terms of a cascade of interacting molecules (or molecular complexes) resulting in a change in some cellular process in response to some stimuli. Each step in the pathway involves a different molecule or molecular complex. Signaling pathways respond to environmental stimuli, either internal or external to the cell, and carry a message that causes (signals) a change in the cell's functioning. Contrast this with a metabolic pathway such as the glycolysis pathway, where one chemical, through a series of precise steps, is transformed into another chemical. In the glycolysis pathway, glucose is transformed into pyruvate. In signaling pathways, it is the activation or inhibition of a process, in metabolic pathways, the end product is a transformed chemical molecule.

At the encoding or syntax level, the representation of a single reaction can be very different. In one database, HumanCyc, the term β-`D-glucose-6-phosphate` is used while in another database, KEGG, the term β-`D-Glucose-6P` is used. It is clear that we are referring to the same molecule, i.e. the same real world class of instances, but the vocabulary label used to name these instances differs and while this difference is insignificant for a human reader, it is significant for computational processing. The syntax in KEGG is XML and a biochemical reaction is defined as an `XML ELEMENT`. In KEGG `reaction` elements have two components, a `substrate` and a `product`. The substrate and product elements each have one required attribute, a `name` which is a KEGG identifier. In HumanCyc, a biochemical reaction is represented as an `ENZYMATIC-REACTION` in OCELOT [1]. Ocelot is a frame based system which uses slots `LEFT` and `RIGHT` to represent the reactions participants. At the time BioPAX began, HumanCyc was encoded only in OCELOT syntax which is an ASCII flat-file format.

The consequence of these different conceptualizations and encoding standards was that it was impossible to construct federated queries across databases. It meant that either individual queries were needed for each database or individual parsers needed to be written. Researchers embarking on a new wet-lab investigation need to consult a multitude of databases in widely different syntaxes. This meant that there was a growing burden to the researcher in making effective preparation before continuing their investigations.

2 The process by which BioPAX came about

A group of researchers within the molecular biology community recognized the need for these data to be united in a common language and format and set out to create a standard representation for biological pathway data that would enable that goal. The group called themselves and their representation BioPAX. The starting point was to use the existing pathway data that resides in pathway databases. [3] A small team was assembled and an organizational structure was put in place. Biweekly conference calls were held and face-to-face meeting venues rotated amongst the workgroup members to distribute travel inconvenience and costs.

2.1 The Representation Language

Biological processes are complex and there was concern about what language would be suitable to represent these complexities. The most prominent exchange format then (and probably still now) were the World

[3] When the project began, there were several dozen such databases and this has grown over the years to over 200 now.

Wide Web Consortium (W3C) standards, eXtensible Markup Language (XML), and XML Schema. XML is a standardized method for marking up text (or data) in an organized way, allowing that text (or data) to be exchanged between any application that understand the XML standard. XML Schema is a W3C XML language for describing and constraining the content of XML documents. At the same time, another W3C standard, OWL, was emerging as a standard for representation of ontologies. Ontologies are formal ways to describe the relationships among the terms and attributes and can represent knowledge about a domain. In addition to describing and constraining the content, they describe and constrain how the content elements relate to each other. The workgroup decided that a representation using a more semantically rich language was a better and more appropriate choice. In addition to ontologies being machine readable, they enabled a conceptually higher level of data characterization and processing that included detailed expressiveness in the relationships defined and logical inference across those relationships.

The workgroup discussed their concerns and impact associated with the risks of each decision. OWL was not yet a standard. What if it were not adopted by the W3C? Would the benefit OWL-DL computability outweigh the cost of OWL-Full expressivity? There was a plethora of tools for XML and a paucity of tools for OWL, how much impact would this have on the project? Several within the group knew XML, none knew OWL.

"The next topic we discussed was the choice of a syntax language for BioPAX. We saw four viable options: 1) use OWL to design the ontology, use XML Schema for data instances, distribute software that checks if the XML Schema is valid for a given ontology, 2) use OWL for everything, 3) option 1 with DAML+OIL instead of OWL, 4) use DAML+OIL for everything. Since OWL was deemed superior to DAML+OIL and would likely contain comparable tools in the near future, options 1 and 2 received the most attention. .... After discussing our options, we decided to choose option 2 above (using OWL for everything), but on a three-week trial basis (REF)." By the next meeting, the workgroup agreed to two parallel versions, one in OWL and another in XML Schema. It appears that there was confusion about syntax and semantics and the misapprehension one could translate between OWL and XML Schema in both directions without loss.

The workgroup struggled with these early decisions. Rogers describes the diffusion of technological innovations as involving a number of different types of people including early adopters, the late majority and laggards [2]. One of the fundamental challenges of the BioPAX consortium was that it included the whole range of such participants, from people who were making the innovations in the technology to people who might be characterized as laggards because of the need to support the existing databases. This presented a major sociological obstacle to easy progress especially in the early stages of development. Tus it was that they agreed to disagree and move forward with a proposal to do both, an OWL-DL version and an XML Schema version[4]. Readers interested in the topics and discussion details can consult the public minutes of the BioPAX consortium which are available at their website and wiki (www.biopax.org, biopaxwiki.org)

Ultimately, OWL-DL was developed and released, the XML Schema version was not. Or was it the other way around? There were several issues with the OWL-DL in the resultant BioPAX OWL (see [4–6]) and it could be argued that it is really an XML Schema version in OWL's clothing.

2.2 Reasons for choosing OWL

There were a number of further reasons to favor the use of OWL and specific versions OWL, in this context. In order to support future uses of pathway data, the BioPAX Workgroup included machine computability among its design principles[7]. This resulted in the choice of OWL-DL. OWL-DL enables full use of reasoners [8], which are software programs that perform inferences based on Description Logics (DL), a subset of first order logic [9]. Furthermore, OWL-DL enables sound and complete inferencing when used by these software reasoners [10, 11]. Reasoners can read an OWL file and based on the logical axioms of the OWL ontology, decide whether that set of axioms are logically consistent and, in addition, infer subsumption relationships that are not explicitly encoded in the ontology. For example, it can infer that a certain molecule is a protein. This is a significant advantage over the other representations in the management of knowledge [12].

[4] Minutes from Menlo Park meeting available at `http://www.biopax.org/Docs/2003-02-21_mtgmins.rtf` and also [3]

As OWL was soon to become a W3C recommendation for the standard web ontology language [13], OWL satisfied another BioPAX design goal, namely compatibility. BioPAX, wherever possible, would employ existing standards. In addition, and of great concern to the BioPAX Workgroup was the expressivity of the language. That is its capacity to represent complex relationships such as those found in biology. For example, in XML-Schema [14] or RDF [15] it is not possible to express that two classes are disjoint; that a molecule of DNA cannot be a molecule of RNA. It is not possible to express that two classes cannot contain any members in common; that instances that are of the class DNA are not and cannot be of the class RNA. In contrast, OWL has, for a Description Logic, a wide range of expressivity for describing constraints on class membership by instances [9, 12].

2.3 Mistakes in using OWL

Due to the participants relative inexperience with respect to OWL a number of mistakes were made early on. The OWL version of BioPAX exhibited the usual common errors and mistakes due in part to the huge learning curve and in part to the desire for the XML version, which is one of the common formats for OWL ontology exports. These included bad conceptualizations and poor use of of OWL.

– Bad Conceptualizations: There was confusion about what was being represented, biological processes or database records of biological processes. The Utility class, a concept used in Java, not in biology, was a top level node in the BioPAX ontology.
– Poor Understanding of OWL: The first release did not include disjoint axioms, there was confusion about domain and range, open world assumptions and implications, and much of the semantics were captured in the comments rather than in the ontology. What was said in the ontology, was not what was meant.

For further details concerning technical issues and errors in BioPAX, please consult Luciano and Stevens (2007).

3 Social Factors

There were a number of social factors which made it both possible to create BioPAX initially but also hindered its development at various stages.

The goal was to create a public resource for pathway data. People interested in this from the broader scientific community were invited to participate. A few had already been thinking about such an initiative. The present author was a co-organizer of the BioPathways Consortium, and found herself in a unique position to make a significant contribution. She had recently found out from the director of the Department of Energy that the DOE was prepared to support emerging and enabling technology. Thus it was that the DOE, the NSF, Dr. Sander (representing the pathway resource initiative) and a number of other researchers in pathways were brought together. The present author became a paid consultant in the role of community liaison and wrote the initial grant for the initial BioPAX DOE funds. The role was not one of a stakeholder in the traditional sense but rather the role was to make sure the community was involved, that the right people were at the table and their needs met (i.e. their data could be represented, or their tool could be used).

While there was a role as a public face in BioPAX, making presentations to spread the word and get people involved, most of the work was behind the scenes working individually with research groups around the world to inform them and invite their participation. At the bi-weekly meetings and face-to-face meetings, much of the task involved listening and, depending on what said or not said, working to keep things on track. The author and one other paid consultant met at coffee shops and worked from her home office to put in place the organizational structure and materials that were needed to support, manage, and document the effort. The effort was largely viewed as a success by the global community and many social logistic factors contributing to that success were presented at Bio-Ontologies in Glasgow [3].

Once the decision was made, the focus moved back to representational concerns of the pathway concepts. OWL was treated as an output format. There were no tools at the time that wrote OWL. The two ontology tools under consideration were Stanford's Protégé Editor and SRI's GKB Editor. We took a vote and the result was split (this time a more evenly). We went with the GKB Editor because SRI had already invested in it, didn't have to retrain and agreed to extend it to export in OWL format. The only serious effort towards understanding OWL, at the time, was in the Export function. Development was delayed and by the time we were to release, Protégé had just come out with it's OWL Alpha version and so we switched to make our release deadline. We switched back shortly thereafter, to GKB editor, had to reenter everything and then had to re-enter everything from the Protégé Alpha to the Protégé Beta version (these OWL formats were not compatible). The issue resolved itself when both tools supported OWL.

The community saw the value in BioPAX and while initially there were some sense of competition pathway models, a wider global community had formed and an increasing number agreed to be BioPAX compliant. The talk at Bio-Ontologies 2004 in Glasgow captured a lot of what went right with BioPAX[3]. In addition to interest from the Bio-Ontologies community, BioPAX also received support in the form of exposure from the semantic web community. Both Tim Berners-Lee and Carole Goble highlighted BioPAX in their plenary sessions, with Carole urging delegates to seek out BioPAX and help build tools.

4 What went right

The community outreach was a big part of the adoption of BioPAX and while it got a lot of details incorrect, BioPAX brought the wider community together by creating what effectively is an upper level ontology for pathways (a level of abstraction that can be used to bring different conceptual representation together).

5 Future Outlook

Many of the issues that plagued BioPAX are less of an issue now. For example, now there are multiple OWL syntaxes to choose from, a selection of tools, both open source and commercial, better support and methodologies for development and dubbing. However, tools are needed that support ontology developers to analyze the level and type of complexity of use cases and facilitate development in a staged process with increasing complexity at each stage. Tools that would support basic requirements first, controlled vocabularies, taxonomies, (XML data exchange) then interoperability (SBML/BioPAX) then on to richer semantics enabling integration, inference, and possibly integrated or in-line rules . Support (and perhaps encouragement) for coherent requirements are also important to prevent politics and bullying; two camps equals two solutions, which is not necessary when the only real difference is in the degree of implementation - that is data exchange at the XML level (often referred to BioPAX-DX, DX for data exchange on the biopax-discuss list) can be viewed as a 'different use case and therefor argument for a separate community' but it is, at the semantic level, a subset of the full intentions of BioPAX. The BioPAX community was always heterogeneous , stakeholders after all, have different interests and different goals vary. They come together in a community because each one has something to offer or has something to gain and can contribute a unique perspective. At one point BioPAX had this self image.

It should be remembered that it was a parallel goal to develop the XML Schema version of BioPAX . This fell from the official agenda but has not been forgotten. The community had originally agreed to support both formats and it is evident from the user community (as expressed by the mailing list) that there is a considerable user base for the XML schema version. Furthermore there is no indication that this community is diminishing.

6 Conclusion and Lessons Learned

From this experience and the other lesson's learned, we can see the need to assess the complexity of the use case, the needs of the community and the capability of the language (and its limitations) as well as tools available and proceed in a more constructive manner, one in which subsequent levels of complexity are supported by a sound foundation.

Then we put forth some notions on levels of complexity and making an argument for evaluation at (at least) the levels of correct, complete/comprehensive (needed for the task) and utility/effectiveness.

References

1. Karp, P.D., Paley, S.: Integrated access to metabolic and genomic data. J Comput Biol **3**(1) (1996) 191–212
2. Rogers, E.M.: Diffusion of Innovations. Fifth edition edn. Free Press, New York, N.Y. (2003)
3. Joanne Luciano, J.: Biopax: Data exchange ontology for biological pathway databases. In Stevens, R., McEntire, R., Lord, P., James.A.Butler, eds.: Proceedings of the Seventh Annual Bio-ontologies Meeting. (2004)
4. Luciano, J.S., Stevens, R.D.: e-science and biological pathway semantics. BMC Bioinformatics **8 Suppl 3** (2007) S3
5. Ruttenberg, A., Clark, T., Bug, W., Samwald, M., Bodenreider, O., Chen, H., Doherty, D., Forsberg, K., Gao, Y., Kashyap, V., Kinoshita, J., Luciano, J., Marshall, M.S., Ogbuji, C., Rees, J., Stephens, S., Wong, G.T., Wu, E., Zaccagnini, D., Hongsermeier, T., Neumann, E., Herman, I., Cheung, K.H.: Advancing translational research with the semantic web. BMC Bioinformatics **8 Suppl 3** (2007) S2
6. Ruttenberg, A., Rees, J.A., Luciano, J.S.: Using owl dl for the exchange of biological pathway information. In Grau, B.C., Horrocks, I., Parsia, B., Patel-Schneider, P., eds.: OWL: Experiences and Directions (OWLED 2005). (2005)
7. BioPAX workgroup: BioPAX – Biological Pathways Exchange Language Level 2, Version 1.0 Documentation. Posted on the BioPAX.org website (December 2005)
8. Bechhofer, S., van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D.L., Patel-Schneider, P.F., Stein, L.A.: OWL Web Ontology Language reference. W3C Recommendation (10 February 2004) Available at `http://www.w3.org/TR/owl-ref/`.
9. Horrocks, I., Patel-Schneider, P.: Reducing OWL entailment to description logic satisfiability. J. of Web Semantics **1**(4) (2004) 345–357
10. Horrocks, I.: The fact system
11. Sirin, E., Parsia, B.: Pellet: An owl dl reasoner. In Haarslev, V., Möller, R., eds.: Description Logics. Volume 104 of CEUR Workshop Proceedings., CEUR-WS.org (2004)
12. Stevens, R., Aranguren, M.E., Wolstencroft, K., Sattler, U., Drummond, N., Horridge, M., Rector, A.: Managing owl's limitations in modelling biological knowledge. International Journal of Human-Comptuter Studies (2007) forthcoming.
13. McGuinness, D.L., van Harmelen, F.: Owl web ontology language overview. Technical report, W3C - World Wide Web Consortium (January 2004)
14. Fallside, D.C., Walmsley, P.: Xml schema part 0: Primer. Technical report, W3C (2004)
15. Lassila, O., Swick, R.: Resource description framework (RDF). model and syntax specification. Technical report, W3C (1999) W3C Recommendation. http://www.w3.org/TR/REC-rdf-syntax.