Proceedings of the 2006 Winter Simulation Conference
L. F. Perrone, F. P. Wieland, J. Liu, B. G. Lawson, D. M. Nicol, and R. M. Fujimoto, eds.

A METHOD FOR COMPARISON OF STANDARDIZED INFORMATION WITHIN SYSTEMS BIOLOGY

Lena Strömbäck

Dept. of Computer and Information Science
Linköpings Universitet
S-581 83, Linköping, Sweden

ABSTRACT

Standards and standardized data representation to allow efficient exchange of information is an important topic within systems biology. Within this area there is currently a rapid development of new standards as well as a need for import of datasets into various computer tools for further analysis. As the number of available standards within systems biology is large, tools for comparison and translation of standards are of high interest. In this paper we present a method for comparison of standards. We illustrate how the method works by providing an analysis of the three standards SBML, PSI MI and BioPAX. The analysis gives information on how similar the three standards are and it also gives pointers on how to build tools to aid a user in the analysis of a standard.

1 INTRODUCTION

The main goals for ystems biology is to understand the complex interaction between genes, proteins and other substances within living organisms. Two major research institutes (Hermjakob et al. 2004, Collins et al. 2003) states this as one major goal for future research. The latter (Collins et al. 2003) in particular mention, reuse of data and reusable software components as one major aim to reach these goals. To increase the possibilities for reuse and exchange standardized formats for representation of information are very important together with computer based tools for analysis and simulation large datasets.

One ongoing effort to reach this goal is the development of new standards for describing different aspects of proteins, interactions, pathways and experimental data. Many of those standards are implemented in XML or OWL (Strömbäck et al. 2006). This means that a user that wants to reuse existing datasets often needs to cope with several formats for expressing information. Since experimental data contains more and more detailed information it is important that researchers can access and reuse each other's data from results from single experiments to models for analysis and simulations in a transparent way. This is, in particular, important for simulation applications, where import of existing datasets into simulation software is important for combining larger pieces of knowledge.

In this paper we focus on three of the most well-known standards for representation of molecular interactions or cellular pathways, SBML, PSI MI and BioPAX. We then briefly describe our method (Strömbäck 2006) for analysis of XML-based standards and show how to use the method for capturing similarities and differences between the standards.

2 AVAILABLE STANDARDS WITHIN SYSTEMS BIOLOGY

There are currently many XML-based standards available within systems biology. (Strömbäck et al. 2006) These standards range from specific standards for describing experimental setup and results based on specific equipment to more general standards aiming at being standards for exchange of data.

In this work we focus on three standards for describing molecular interactions or signaling pathways, SBML (Hermjakob et al. 2004), PSI MI (Hucka et al. 2003), and BioPAX (BioPAX 2005). We have chosen these formats because they have been proposed as future standards and are currently under active development. For SBML and PSI MI there are already large datasets of data available, and for BioPAX there is ongoing work on providing larger datasets. Here, we give a short introduction to these standards; for a more extensive description see (Strömbäck and Lambrix 2005).

Systems Biology Markup Language (SBML) (Hucka et al. 2003) was created by the Systems Biology Workbench Development group. The scope of SBML is simulation models and it is currently supported by a large number of tools and databases. A brief example of an SBML model is given in Figure 1. As we can see, an SBML model contains a number of *compartments*, each of which is a description of the container or environment in which the reaction takes place. The substances or entities that take part in the reactions are represented as *species*. The interactions between molecules are represented as *reactions*, defined as

processes that change one or more of the species. *Reactants*, *products* and *modifiers* for reactions are specified by references to the relevant species.

The Proteomics Standards Initiative Molecular Interaction XML format (PSI MI) (Hermjakob et al. 2004) was developed by the Proteomics Standards Initiative. The scope of PSI MI is to describe experimental results and large datasets are available from a number of databases. An abbreviated example represented in PSI MI is shown in Figure 2. In PSI MI the *experimentList* describes experiments and links to publications where the interactions are verified. The pathway itself is described via the *interactorList*, which is a list of substances participating in the interaction, and the *interactionList*, a list of the actual interactions. For each *interaction* it is possible to set one or more names. The participating proteins are described by their names or by references to the *interactorList.* Note that, where the intention of SBML is to describe an actual interaction, i.e. that interacting substances produce some product, the purpose of PSI MI is to describe the result of an experiment, i.e. that there is some chemical interaction between the substances but roles of the substances in the interaction are not always known.

The *BioPAX Data Exchange* format is defined by the BioPAX working group (BioPAX 2005). The aim of this standard is to define a unified framework for sharing pathway information. BioPAX is different from PSI MI and SBML in that it uses OWL for implementation instead of pure XML. Due to this the standard is presented as a hierarchy of concepts, where all concepts can have different properties for further description of actual data. In SBML information is centered around substances, called *Physical Entities* and *Interactions.* For each of these main concepts a number of subclasses are defined allowing the user to define many types of substances, such as proteins and DNA, together with different kinds of interactions. Figure 3 gives an overview of the BioPAX hierarchy.

Even though the three formalisms have different scope there are many similarities between them. In this paper we further explore the similarities and differences.

3 THE RELATION BETWEEN XML AND OWL

As mentioned in the previous section, two of the standards we want to compare are implemented in XML and the third in OWL. Since we want to be able to compare standards on the semantic level, and thus avoid discussing differences due to differences between XML and OWL, we need some way of defining how concepts in one format can be compared to the other format. In principle the XML-structure defines a syntax to be used to represent data. This syntactic structure expresses the semantics of the data that we want to work with. The aim here is to find a translation that captures this semantics and translates it to the semantics expressed by an OWL implementation.

```
<model name="Example">
<listOfCompartments>
 <compartment name="Mithocondrial Matrix"
 id="MM"/>
</listOfCompartments>

<listOfSpecies>
 <species name="Succinate"
 compartment="MM" id="Succinate" />
 <species name="Fumarate" compartment="MM"
 id="Fumarate" />
 <species name=”Succinate dehydrogenase"
 compartment="MM" id="Succdeh" />
</listOfSpecies>

<listOfReactions>
 <reaction name="Succinate dehydrogenas
 catalysis" id="R1">
 <listOfReactants>
 <speciesReference species="Succinate" />
 </listOfReactants>
 <listOfProducts>
 <speciesReference species="Fumarate" />
 </listOfProducts>
 <listOfModifiers>
 <modifierSpeciesReference
 species="Succdeh" />
 <modifierSpeciesReference species="S4" />
 </listOfModifiers>
 </reaction>
</listOfReactions>
</model>
```

Figure 1: Example of data in SBML

```
<entry>
<interactorList>
 <Interactor id="Succinate>
 <names>
 <shortLabel>Succinate</shortLabel>
 <fullName>Succinate</fullName>
 </names>
 </Interactor>
 ….
</interactorList>
<interactionList>
 <interaction>
 <names>
 <shortLabel> Succinate dehydrogenas
 catalysis </shortLabel>
 <fullName>Interaction between ....
 </fullName>
 </names>
 <participantList>
 <Participant>
 <proteinInteractorRef ref="Succinate"/>
 <biologicalrole>neutral</role>
 </proteinParticipant>
 <proteinParticipant>
 <proteinInteractorRef ref="Fumarate"/>
 <role>neutral</role>
 </proteinParticipant>
 <proteinParticipant>
 <proteinInteractorRef ref="Succdeh"/>
 <role>neutral</role>
 </proteinParticipant>
 </participantList>
 </interaction>
</interactionList>
```

Figure 2: Example of Data in PSI MI

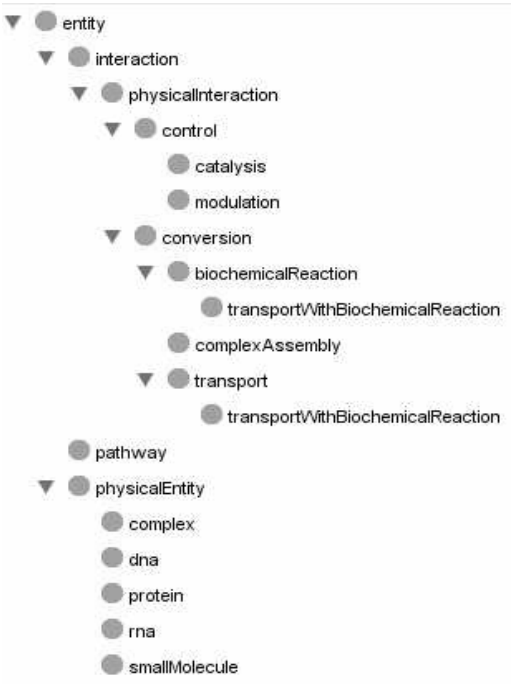

Figure 3: The BioPAX Concept Hierarchy

The structure of PSI MI and SBML are described by XML-schemas. There are large differences between XML-schema and OWL. One is the semantic scope of OWL versus the syntactic description of XML. Another important difference is that XML only uses substructure relation to express information, while OWL has many kinds of constructions, i.e. *class-subclass*, *datatypeproperties* and *objectproperties*. This means that there are several possible translations between the two formats, depending on what kind of features that are important to capture. In this work it has been important to capture the semantics of each of the standards as closely as possible. We have therefore chosen to follow the approach of lifting XML-schema to OWL (Ferdinand *et al.* 2004). Table 1 gives a summary of which constructions within XML and OWL that are considered equivalent.

With this translation we get a clear correspondence between each concept in an XML Schema and corresponding concepts in OWL. We exemplify it by a small excerpt of the XML schema:

```
<xsd:element name=sbml type=Sbml/>
  <xsd:complexType name=Sbml>
    <xsd:extensionbase=?SBase?>
      <xsd:sequence>
        <xsd:element name=model type=Model/>
      </xsd:sequence>
    </xsd:extension>
  </xsd:complexType>
```

This example will by our definition be translated to the following two OWL constructs:

```
<owl:class ID= Sbml>
 <rdfs:subclassOf rdf:about=Sbase/>
</owl:class>

<owl:objectProperty name=model>
 <rdf:range source=Model>
 <rdf:domain resource=Sbml>
</owl:objectProperty>
```

The example also demonstrates that the type hierarchy defined for the XML-schema is naturally translated into the OWL ISA hierarchy.

Table 1: Overview XML mapping to OWL

XML schema concept	OWL concept
Named complexType	Class
Unnamed complex type	Class
Non-leaf element	ObjectProperty
Leaf element	DatatypeProperty
Attributes	DatatypeProperty

4 A METHOD FOR COMPARING STANDARDS

In previous work (Strömbäck 2006) we provide a categorization for judging the similarity of different standards within the same domain. This categorization describes how semantic concepts are expressed by the structure provided in the standard and, thus, how this differs between different standards. The categorization can be used as a basis for creation of domain specific tools, but also as a tool for judging how well automatic methods for aligning standards would work within a domain.

In general, working with this kind of standards, there are not one single semantic domain agreed by the community. Our categorization assumes that the syntax behind a standard describes the semantic concepts for this standard and these concepts are what we want to find and compare. The categorization contains two dimension that describes how similar standards are. The first dimension concerns what semantic concepts are available within a standard and how they are expressed. The second concerns what information that is available in the standard that can help us in judging semantic similarity.

We base our categorization on the two dimensions and for each of the dimensions we present a number of categories that describe the similarities between standards. Analyzing standards based on these categories give information on how difficult it would be to merge information in standards into one system.

Dimension S: This dimension gives a description of the common semantic concepts of the standards and how these are expressed:

S.1 Find all concepts common for the standards.

S.3 For each concept how does it correspond to concepts in other standards. Same concept, sub concept or is it

instantiated in several places with different conditions.

S.2 For each of the concept pairs check how they occur in comparison to each other, i.e. side by side or as sub concepts of each other. Compare between the standards.

S.4 Check whether any of the interesting concepts occur within fields free for the user to extend.

Here, category S.1 specifies the semantic concepts that we want to find in our standards and thus gives the prerequisites on whether a match between them is at all possible. Category S.2 gives more information about how well the semantic concepts as such match between standards. Here, the specification of reaction or interaction is a good example where information under *participantList* in PSI MI occurs within one of *listofReactants*, *listofProducts* or *listofModifiers* in SBML. This category gives information on semantic similarity between the concepts.

Category S.3 gives information about the overall semantic, for instance, in this domain an *interactor* is normally defined in a list parallel to the list of *interactions*, while the participants of an interaction are always a substructure of the *interactor*. If the variation differs between different concepts this is a sign on differences in the overall semantic model of the standard. Finally category S.4 is important since many standards include free slots, where the user can add further information.

As a help to find common concepts, there is currently ongoing research on alignment tools. The second dimension describes which information in a standard that is available to find a match between standards. This categorization is modified from a categorization (Lambrix and Tan 2005) of matching algorithms for ontology merging.

Dimension I: This dimension gives a description of available information for matching concepts in a standard.

I.1 Linguistic or string matching information.

I.2 Auxiliary external information for finding synonyms, such as Wordnet or a domain specific ontology of concepts.

I.3 Information regarding the XML structure or other information in the XML or OWL specification.

I.4 Information from examples of instantiated data.

Note here that the first three categories only make use of information available within the XML schema or OWL hierarchy definition while the last one also uses information from sample data files.

5 DIMENSION S ANALYSIS

We now demonstrate how the analysis works by giving a semantic analysis on the S-dimension. This will also give information on how complex the connections between the information in the different standards are.

For category S.1 we need to find a match between common concepts in all three standards. We have done this manually by analyzing the standards and also available information about conversion between PSI MI and BioPAX (BioPAX 2006). We have concentrated our analysis on six main concepts, the general level, representation of interacting substances, interactions, pathways, compartments and experiments and an overview of the results is shown in Table 2. Where relevant, we show which are the main concepts and how these are connected in the ISA hierarchy. For all concepts we also show which structure is used for describing this entity. Due to the scope and space of this paper we have concentrated on the semantically most important concepts and therefore many concepts of more detailed and administrative nature have been omitted. We have also concentrated the descriptions on showing similarities, that is in principle we only show concepts that are common for at least two of the hierarchies, unless the concepts are important for understanding the main structure of the standard.

Each row in the table corresponds to a semantic entity. This means that when there are concepts on the same row for more than one of the standards these concepts correspond to the same semantic entity and should be aligned by an alignment system. Please note that in the information for the type hierarchy indentation corresponds to an ISA-relation. This means that, for instance, on the general level we have aligned BioPAX's *Interaction* with SBML's *Reaction*, while we on the more specific level *Interaction* show that it actually corresponds to one of the subclasses, *Conversion* or *Control* of BioPAX's *Interaction*.

For the structure parts there are two important things to comment. First, according to our translation, BioPAX concepts corresponds to *Datatypeproperties* or *Objecttypeproperties* in OWL and for SBML and PSI MI they correspond to XML-structure. Something that is not clear from the presentation in the table is that for BioPAX the available properties are dependent on the subtype. Thus, for instance *Sequence* is only available for *DNA*, *RNA* and *protein*. For interaction the properties *controller* and *controlled* are only available for control interactions while left and right are available for conversion interactions. In this case, as indicated by the indentation, all these properties are represented as subtypes of the *property participant*.

For category S.2 we want to know how well the semantic concepts match between the standards. Here there are in principle four important differences.

One is the representation of names, where PSI MI allows more alternative names than the other formalisms. The other is the representation of experimental information, where PSI MI allows this to occur in more places than the other formalisms. The remaining two differences are again caused by the different uses of typing. First, the dif-

Table 2: Equivalent concepts in the standards

Concept	Semantic	BioPAX	PSI MI	SBML
General	Type hierarchy	OWL: Thing Entity Interaction Pathway Physical Entity UtilityClass Evidence	InteractionElementType (Interaction) InteractorElementType (Interactor) OpenCvType Compartment ExperimentType	SBase Reaction Species Compartment
	Main structure		Entryset Entry	SBML Model
Interacting substances	Type hierarchy	PhysicalEntity Complex DNA RNA Protein SmallMolecule	InteractorElementType (Interactor) Complex Protein-DNA-complex Protein complex …. Biopolymer Nucleic Acid DNA RNA …. Protein Peptide Gene Interaction Small Molecule Unknown Participant	Species
	Structure	PhysicalEntity Name, shortname xref Interactortype Organism Sequence Component Chemical formula	Interactor Id, names xref Interactortype compartment Organism Sequence	Species Id, name Compartment
Interaction	Type hierarchy	Interaction PhysicalInteraction Control Catalysis Modulation Conversion Biochemicalreaction ComplexAssembly Transport	InteractionElementType (Interaction)	Reaction
	Structure	Interaction Name, shortname xref Interaction_type Participants Controller Controlled Left Right Cofactor Evidence	Interaction imexID, id, names xref Interactiontype Participantlist Participant Biological role Experimental Role List of experiments Experiment	Reaction id, name ListofModifiers Modifier ListofReactants Reactant ListofProducts Product
Pathway	Structure	Pathway Evidence Organism Pathwaystep		Model
Compartment			Compartment	Compartment
Experiment		Evidence	Experiment	

ferent granularity of types causes, for instance, a need for the general concept *Interactor* in PSI MI to be matched with several more specific concepts in BioPAX. Secondly a *Reaction* within SBML has here been matched with both conversion and control. In reality this means that a reaction corresponds to either a conversion or a combination of a conversion and a control dependant on which of *Reactant*, *Product* and *Modifier* that are instantiated.

For category S.3 we need to analyze how these concepts occur in the overall structure. From the given table it is clear that the overall structure of the three standards are similar. If we consider PSI MI and SBML, there are differences concerning how they chose to realize their concepts in terms of XML-structure. Here SBML tend to use attributes whenever possible while PSI MI uses substructure. However, using our translation to OWL both these constructions are considered as data type properties.

One difference between the BioPAX and the other formalisms is that XML requires extra levels for forming lists. Which means that PSI MI and SBML in many places contain extra concepts for representing *ListofParticpants* or similar. One other main difference is the treatment of typing between PSI MI and BioPAX. In PSI MI different types of *interactions* and *interactiors* a represented by a the attributes *Interactiontype/Interactortype* as a substructure to *Interaction* and *Interactor*. This analysis on category S.3 tells us that the main semantic relations of the common concepts in the standards are represented in a similar way.

Finally, for category S.4, both PSI MI and SBML contain means for the user to add further structured information not specified by the standard. In principle these parts of the standards can be used for representation of any kind of information, and is often used by various databases to add information not specified by the standard. However, here we only note that this is a potential difficulty, but as this is very dependent on a particular user's instantiation of the standard we will not further discuss it here.

6 DIMENSION I ANALYSIS

To make an analysis for dimension I, we used two systems available for matching ontologies. The first system is SAMBO (Lambrix & Tan 2006) developed at Linköping University. This system is very good for our purposes since it provides several different matching algorithms where the user can choose and combine them by specifying weights and thresholds. This feature provides a very good possibility to make a comparison of the standards based on the different categories in the I dimension. The second system that were used was Prompt (Noy & Musen 2003). This system does not provide the possibility to combine matchers, but contains a synonym concept that was interesting to test in this setting.

For our tests we used the string matcher provided by SAMBO to test how many concepts that could be matched based on category I.1. For these tests we used three thresholds, 0,6, 0,5 and 0,4 to decide if a match were relevant or not. For category I.2 we used two SAMBO matchers, one using the UMLS (National Library of Medicine), and a second allowing to look for synonyms via Wordnet. In this tests we also compare with Prompt's possibility to find synonyms. Category I.3 was analysed by SAMBO's structural matcher. This matcher allows us to test whether the structure of the standard can be helpful to provide a match. This means that two concepts are judged as a match if they occur in similar surroundings in the standards.

The results of our tests are presented in Table 3. For each combination of two standards and matchers we present three figures. The top one (C) represents the number of correct suggestions by the matcher. The middle, (R) the number of relevant matches. A match is judged as relevant if it is helpful for the user, i.e. it is close in the structure or ISA hierarchy. The last figure for each match (W) represents erroneous and confusing matches.

Table 3: Suggested Matches by Different Matching Algorithms

		String 0.6	String 0.5	String 0.4	UMLS 0.6	Wordnet 0.5	Prompt	Structure
SBML BioPAX	C	2	2	2	2	2	3	2
	R	0	0	1	0	0	0	0
	W	0	2	2	0	2	7	0
SBML PSI MI	C	11	11	13	11	11	12	16
	R	1	1	7	1	1	2	15
	W	0	2	2	0	2	4	3
PSI MI BioPAX	C	17	17	34	17	20	19	17
	R	0	11	17	0	10	3	0
	W	0	9	33	0	9	13	0

The first conclusion that can be drawn from the table is that PSI MI and BioPAX are the two standards where we find most interesting matches. This is true, also in the sense that these two standards are the ones with most similarity, both in terms of semantic concepts and naming conventions. We can also conclude that the pair hardest to find any matches for are SBML and BioPAX. This is also true in the sense that there is a large difference in naming convention and structure between these two standards.

If we consider the different kind of matchers we can see that the string matcher (category I.1) finds reasonably many matches for PSI MI-BioPAX and SBML-PSI MI. We can see that lowering the threshold gives more interesting matches for both of these combinations. However, for PSI MI-BioPAX to the cost of many incorrect suggestions.

For category I.2, we can see that neither UML nor Wordnet significantly increases the number of correct matches compared to using only the string matches with

the same threshold. In principle the same matches are found here. Prompt's synonym matcher does, however find a small number of relevant new matches, this is though to the cost of a relative large number of incorrect matches. These results are relatively disappointing, however, none of the tested vocabularies were specific to systems biology and further tests need to be done on, for instance, available domain specific ontologies.

As explained above category I.3 was tested with a matcher guided by structural similarity. The matcher starts with similarity measures from some other approach, in this case we used the string matcher. Using these figures as a basis the matcher traverses the structure again and increases similarity values for substructures that have many children, parents or substructures in common. This matcher gave very interesting results for SBML-PSI MI, but no increase of suggestions for the other combinations. The reason for this is that the structural similarity is more clear between SBML and PSI MI. The current matcher does differ between relations, and therefore becomes disturbed by the ISA hierarchy in BioPAX. However, we find this approach promising and a separation of structural relations together with addition of cardinality constraints are improvements to test for the future.

In this setting we have concentrated on information in the XML-schema definition and therefore not tested matching based on category 1.4. There are however several indications that this kind of information could increase the number of interesting matches. One is the high amount of links to controlled vocabularies, which can be a useful source of information, another is that naming of entities within the area often follows strict conventions. This topic requires, however, further studies.

7 RELATED WORK

Comparison of XML-standards is similar to the topic of semantic schema matching. Within this area there is a lot of work done for databases in general. Surveys are given in (Doan and Halevy 2004, Rahm and Bernstein 2001). The latter of these presents a categorization of approaches to schema matching. They distinguish between six categories of schema matchers; instance vs. schema matching; element vs. structure matching; language vs. constraint matching; cardinality of matching; to what extent is the matcher dependent on auxiliary information; and matchers using a combination of methods for obtaining the best results.

Similar work is done in the area of categorization of strategies used by systems for aligning ontologies (Lambrix and Tan 2005). Their categorization includes the following kind of matchers; linguistic-based; structure-based; constraint-based; instance-based; and auxiliary-based.

The presented method for comparing XML standards presented in this work is inspired by these two previous categorizations. Our aim of comparing standards meant that, in particular, the S-dimension had to be further developed from the previous categorizations.

In this work we have used SAMBO (Lambrix and Tan 2006). Another interesting tool is COMA (Do and Rahm 2002). Both these tools provide a possibility to combine different kinds of matchers and similar methods could be used for achieving automatic analyses of standards.

Another related and active research field is methods for comparison of XML-structure to be used both for querying and matching of XML documents (Amer-Yahia et al. 2005, De Meo et al. 2003, Melnik et al. 2002, Shen & Wang 2003, Smiljanic et al. 2005). All these matchers rely on various methods of relaxing the XML structure, but use string similarity can be used for identifying common concepts. In this work we have instead put the focus on identifying semantic concepts and from this describe variation in structure. An interesting continuation of this work would be to evaluate how well these proposed methods cover the variations detected in standardized applications.

8 CONCLUSION AND FUTURE WORK

In this paper we present a method for analysis and comparison of XML-based standards within systems biology. The method consists of two dimensions, one for comparing the semantic concepts and one for finding information for automatic match between the standards. We apply this method on three standards within systems biology, SBML, PSI MI and BioPAX.

Our results show that even though these standards express information on the same domain, there is a large difference in information content in them. It also shows that a combination of string and structural information is an interesting choice for providing matches. The analysis suggest that a pure syntactic approach to automatic translation would not be sufficient and that semantic understanding of the standards is important for providing matching and translations between standards.

There are several interesting lines of future work based on this article. One is to make an extended investigation and also include other standards in the analysis. A second is to investigate possibilities for a more fine-grained structural approach to matching and also to test matches based on domain specific matchers. The third topic is to investigate how this analysis can form the basis for tools capable of automatically adapt and work with more than one standard. A final goal are semi-automatic approaches, where the user can interact with the matching process to achieve the most accurate match for each concept.

ACKNOWLEDGMENTS

We acknowledge the financial support from the Center for Industrial Information Technology. I am also grateful to

Patrick Lambrix for comments on this work, He Tan for support on the SAMBO system, and Dagmar Köhn for her work on translation from XML to OWL.

REFERENCES

Abello, A., X. Palol, and M. S. Hacid. 2005. On the Midpoint of a Set of XML Documents. *Proc. DEXA 2005*, Copenhagen Denmark.

Amer-Yahia, S., N. Koudas, A. Marian, D. Srivastava, and D. Toman. 2005. Structure and Content Scoring for XML. *Proceedings of the 31st Internationa Conference on Very Large Databases.* Trondheim, Norway, pp 361-372.

BioPAX working group. 2005. BioPAX – Biological Pathways Exchange Language. Level 2, Version 1.0 Documentation. Available via <http://www.biopax.org> [Accessed May 15, 2006]

BioPAX working group. 2006. PSI-MI Conversion. Available via <http://www.biopaxwiki.org/cgi-bin/moin.cgi/PSI-MI_Conversion> [Accessed May 15, 2006]

Collins, F.S., E. D. Green, A. E. Guttmacher, and M. S. Guyer. 2003. A vision for the future of genomics research: A blueprint for the genomic era. *Nature 422: 835-847.*

De Meo, P., G. Qittrone, G. Terracita, and D. Ursino. 2003. Almost Automatic and Semantic Integration of XML Schemas at Various Severity Levels. R. Meersman, Z. Tari, and D. C. Schmidt. (eds.): *CooPIS/DOA/ODBASE 2003*, LNCS 2888.

Do, H. H. and E. Rahm. 2002. COMA - A system for flexible combination of schema matching approaches. *VLDB 2002*, Hong Kong, China.

Doan, A. and A. Y. Halevy. 2005. Semantic integration research in the database community: A brief aurvey. AI *Magazine, Special Issue on Semantic Integration, Spring 2005*.

Ferdinand, M., C. Zirpins, and D. Trastour. 2004. Lifting XML-schema to OWL. *Proceedings of the International Conference on Web Engineering (ICWE 2004)*, pp 354-358. Springer Heidelberg.

Hermjakob, H., L. Montecchi-Palazzi, and G. Bader. 2004. The HUPO PSI's Molecular Interaction format - a community standard for the representation of protein interaction data. *Nature Biotechnology 22(2):177-183.*

Hucka, M., A. Finney, and H. M. Sauro. 2003. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics 19(4):524-531.*

Lambrix, P. and H. Tan. 2005. A Framework for Aligning Ontologies. *F. Fages and S Soliman (Eds.): PPSWR 2005, LNCS 3703*, pp. 17-31.

Lambrix, P. and H. Tan. 2006. SAMBO - A System for Aligning and Merging Bio-Ontologies, *Journal of Web Semantics, special issue on Semantic Web for the Life Sciences.* [To appear].

Melnik, S., H., Garcia Molina H, and E. Rahm. 2002. Similarity Flooding: A versatile Graph Matching Algorithm and its Applications to Schema Matching. *Proceedings of the International Conference on Data Engineering*, 1063-6382/02. IEEE Computer Society.

National Library of Medicine, Unified Medical Language System. Available via <http://www.nlm.nih.gov/research/umls/> [Accessed May 26, 2006.]

Noy, N. F. and M. A. Musen. 2003. The PROMPT Suite: Interactive Tools For Ontology Merging And Mapping. *International Journal of Human-Computer Studies*. 62(5), pp. 578-596.

Rahm, E. and P. A. Bernstein. 2001. A survey of approaches to semantic schema matching. *The VLDB Journal 10:334-350.* DOI 10-1007/s007780100057.

Shen, Y. and B. Wang. 2003. Clustering Schemaless XML Documents. R. Meersman, Z. Tari, and D. C. Schmidt. (eds.) *CooPIS/DOA/ODBASE 2003*, LNCS 2888, pp. 767-784.

Smiljanic, M., M. Keulen, and W. Jonker. 2005. Formalizing the XML Schema Matching Problem as a Constraint Optimization Problem. *Proc. of the International Conference on Database and Expert Systems (DEXA 2005)*, Copenhagen, Denmark.

Strömbäck, L. 2006. A classification for comparing standardized XML data. *Proceedings of the International Conference on Databases and Expert Systems (DEXA 2006).* Krakow, Poland.

Strömbäck, L., D. Hall and P. Lambrix. 2006 A review of standards for data exchange within systems biology. *Proteomics*. [To appear.]

Strömbäck, L., and P. Lambrix. 2005. Representations of molecular pathways: An evaluation of SBML, PSI MI and BioPAX. *Bioinformatics, 21(24):4401-4407.*

AUTHOR BIOGRAPHIES

LENA STRÖMBÄCK is an Assistant professor at Linköpings Universitet. She has a solid background in working with databases and XML. She holds a Ph. D degree in computer science within natural language processing 1997. After her Ph. D., Lena worked at Nokia with research and development of products for the information society. This work included responsibility for European projects and work with future standards in XML. Her current research focuses on standards and tools for management of standards, mainly within the area of bioinformatics. Her e-mail address is <lestr@ida.liu.se> and her Web address is <http://www.ida.liu.se/~lestr>.