

# **The BioPAX Data Exchange Format v0.5 (draft release)**

Tuesday, September 23, 2003

The BioPAX Data Exchange Format v0.5 (draft release) .....	1
1. Summary .....	1
2. Purpose of this document .....	1
3. Background .....	1
4. BioPAX Mission Statement and Goals .....	2
4. How to Participate .....	3
4. FAQ .....	4
3. Glossary .....	5

## **1. Summary**

BioPAX (<http://biopax.org>) is a new community-based initiative to address the growing need for a unified framework for sharing pathway information. Several groups are participating in the BioPAX effort to develop a data exchange format that will both allow communication between existing pathway databases and facilitate deposition of data into a common public pathway information repository.

## **2. Purpose of this document**

This document introduces BioPAX, focusing on design philosophy, goals and organization. The actual BioPAX data exchange format (DEF) documentation is available in a separate file.

## **3. Background**

In recent years, the number of pathway databases has been increasing. This is due largely to the emergence of a variety of novel experimental methods for generating pathway data on a large scale. It is difficult to gather pathway information from these varied sources for analysis.

A standard exchange format for pathway data would allow pathway databases to provide their data in a common form, thus significantly reducing the amount of time and energy spent by the bioinformatics community on data integration. BioPAX (<http://www.biopax.org>) is a new community-based effort to develop a technical recommendation for a biological pathways data exchange format. The format is being designed to represent data from diverse sources such as BIND, BioCyc, KEGG and WIT.

In designing BioPAX, we endeavor to balance the many different representational needs of the biological pathways community by remaining flexible and extensible. This 0.5 version draft ontology is being made available to get input from the biological pathway community on the proposed format to help make sure that it is extensible and flexible enough to allow the addition of detailed information in the future and that it provides a simple format for the exchange of data. This framework will be extended to include more detail via a leveled

approach similar to that used by SBML. Encapsulation, compatibility and computability are also being emphasized in the design, as is the use of existing standards, such as the Gene Ontology, when available. There is also an opportunity to have the BioPAX standard completely compatible with the Proteomics Standards Initiative Molecular Interaction (PSI-MI) standard for describing protein-protein interactions (<http://psidev.sourceforge.net>).

This initial implementation of BioPAX is available in GKB, Protégé, OWL and XML Schema versions. The projected users of BioPAX are bioinformatics researchers, from beginner to advanced. The development process, the format and all documentation is open for comment and feedback and participation of all interested parties is encouraged.

For the purposes of BioPAX, a biological pathway is a network of biological relationships. This definition is general enough to encompass metabolic pathways, signal transduction pathways, gene regulatory pathways, genetic interaction pathways, pathways through word relationships found in papers using text mining, computationally predicted pathways and pathways of cellular interactions. This definition does not currently aim to cover pathways in biological systems that are higher order than the cell, such as physiological pathways. This defines the scope of BioPAX to be all pathways relating to cellular and molecular biology, although initially, BioPAX will focus on metabolic, signal transduction and gene regulatory pathways, as most existing data is encompassed by one of these three types.

#### **4. BioPAX Mission Statement and Goals**

**Mission Statement:** BioPAX will develop a data exchange format for biological pathways that is flexible, extensible, optionally encapsulated, compatible and computable.

**Flexible:** Different preferred representations of pathway data can be described using BioPAX. Rationale: Increased flexibility lowers the barrier to acceptance, as the many existing pathway knowledge representations are compatible with BioPAX. We would also like to be able to represent the different types of data that are variously regarded in the community as being part of biological pathways. An issue of semantic mapping between different representations still exists and this will have to be dealt with when integrating data.

**Extensible:** Specific classes of data, such as controlled vocabularies or lists of classes, in BioPAX have been marked as extensible to allow addition of new types of data in the future. Rationale: The requirements of a leveled approach to data specification design require a general framework to be built at the start that can hopefully handle the various use cases that will arise in the future. This avoids technical barriers to extending a specification when trying to enable description of new data types that would break the existing specification.

**Encapsulation:** An entire pathway can optionally be encapsulated in a single BioPAX record. Rationale: Encapsulation makes a BioPAX record easier to use for the general user because they do not have to resolve multiple database identifiers used to e.g. point to proteins in SwissProt. Everything required to describe the pathway should optionally be available, including for example, chemical structure. This lowers the activation barrier for use of pathway data because it eliminates an otherwise necessary data integration step i.e. gathering required information from other databases. Encapsulation is optional. A user should be able to decide at download time

whether they want a fully encapsulated pathway or one that doesn't contain full description of chosen pathway components, just pointers to databases containing the full description.

**Compatible:** BioPAX will try to use existing standards for encoding biological pathway related information wherever possible.

Rationale: Using existing standards is practical because it avoids “re-inventing the wheel”, thus allowing the BioPAX group to focus on unsolved problems. Specifically, the Gene Ontology and related controlled vocabularies will be used where possible and we will try to make a version of BioPAX compatible with the PSI MI format.

**Computable:** BioPAX stores data in a computable manner

Rationale: Data exchange formats must store data in a computable manner in order for the stored data to be maximally useful to computational analysis. This means, for example, that values are strongly typed and that the class structure can be ‘understood’ by a machine, as it is in GKB, Protégé and OWL. The types of computability that can be thought of is everything from simply reading the file in to making logical inferences based on the data.

#### **4. How to Participate**

You can help develop BioPAX: Participate, promote and provide feedback, provide data.

BioPAX participation is currently volunteer and members have typically paid their own travel expenses. The US Department of Energy (DOE) has provided some funding for holding meetings and more funding may become available to hold more general meetings in the future.

**Participate** in BioPAX meetings (currently requires an invitation as we are trying to keep the core group small to avoid endless discussions)

**Participate** in BioPAX subgroup meetings. Subgroups may be proposed to extend the BioPAX specification and address specific use cases. For example, we currently have small “small molecule”, “molecular state” and “BioPAX example” subgroups.

**Promote** BioPAX to your colleagues.

**Provide feedback** via BioPAX discussions mailing lists on BioPAX activities and documentation

**Provide data** in the BioPAX format.

Details are available on the [www.biopax.org](http://www.biopax.org) web page.

#### **Organizational Structure of BioPAX**

BioPAX is a small core group of volunteers advancing the standard. The small group enables discussions to stay short. One possible problem with this is lack of representation, but we would like to address this issue by actively reaching out to people in the community for feedback. We would also like to address this by creating special interest subgroups to address specific use cases.

BioPAX has bi-weekly conference calls within the core group, currently about 10-12 people. Minutes for these are on the [www.biopax.org](http://www.biopax.org) website. Monthly to bi-monthly face to face meetings occur with the core group. Minutes for these are on the website. Core group

participants pay their own travel expenses for these meetings. Special interest subgroups are defined and may be proposed to address specific issues and to extend the ontology. For example, we currently have “small molecule”, “state” and “BioPAX example” subgroups.

#### **4. FAQ**

##### **What are the advantages of building a common data exchange format (DEF)?**

There are three main advantages for a common DEF:

1. **Common good:** A common DEF reduces the total amount of work the community must do to integrate data from disparate sources. The types of work that are reduced are custom format translation tools and software application import and export features. Instead of having to create translation tools for all combinations of sources, only translation tools to and from the DEF need to be created. This reduces the number of translations from quadratic scaling to linear scaling with the number of data sources.
2. **Promotes collaboration:** A common DEF promotes collaborations between groups using the DEF in their software as it is easier for these groups to communicate via the DEF.
3. **Accessibility:** Reduces the barrier to entry in the field by making biological pathway data more accessible, which should stimulate research to some degree.

##### **How does BioPAX deal with the choice of representations in different databases?**

BioPAX does not currently directly address the semantic mapping issue that affects data sharing of biological pathways. Different groups use different representations sometime for the same types of data. Generally, the representation that is used is optimized for the use cases of that group, thus having everyone switch to a common representation is not a practical solution. BioPAX, then, must be flexible enough to support the different representations used in the field and the use of the specification in this respect is up to the user. For example, a pathway could be represented using states of the entities that it is composed of (e.g. state transition model), recursively using a nested relationship structure (see examples) or using a classical enzyme-substrate-product reaction model for metabolic pathways. Certain different representations could be mapped to each other using a generic mapping technology (e.g. XSLT) that would allow third party users to choose their preferred representation. Unfortunately, this would likely not be able to solve all semantic mapping issues, but more complex software solutions might address this issue. The “common good” advantage of a DEF remains, since work is still being reduced.

##### **How are states represented in BioPAX?**

States are not currently represented in the v0.5 release of BioPAX. The state subgroup is readying a recommendation document that is periodically updated on the BioPAX website that contains details of a possible state representation to use for BioPAX.

##### **What implementation language has been chosen for BioPAX?**

An implementation language needs to be chosen to syntactically represent a BioPAX record. The OWL (very similar to DAML+OIL) and XML Schema languages have been evaluated. Each language has its own strengths and weaknesses. We would like to represent semantic

relationships for biological pathway data, which requires a powerful ontology language, such as OWL. We would also like to quickly build software around our exchange format e.g. databases, visualization tools. This requires an established schema language that has readily available software rapid application development tools, like XML Schema. Example tools that allow XML schema to be used quickly include automatic Java class generators (e.g. JAXB), XML schema visual editors (e.g. XML Spy) and automatic Java class to relational database mappers (such as OJB). The tradeoff between OWL and XML Schema is purely practical; OWL has the advantages of being able to describe a full ontology and a data exchange format at the same time, but it doesn't yet have established software libraries to work with it at a high level. Since OWL forms the basis of the highly touted semantic web, in the future it will very likely have the same tools available for it as are currently available for DAML-OIL.

XML Schema is a generally accepted data exchange format that provides some semantic constructs, such as hierarchy of fields/slots and complex types, cardinality on lists, data types, controlled vocabularies and constraints on those data types. It does not have constructs in the language for representing a class hierarchy.

OWL provides advanced semantic constructs including all that XML Schema has as well as more advanced constraints and the ability to describe classes. OWL can also differentiate between is-a, has-a and other relationships in the ontology, which helps logical inference engines automatically with more powerful queries.

While good arguments can be made for both OWL and XML Schema as the language for our data exchange format, a third option also exists in which we use both and provide tools to map between the two as much as possible.

We are also making BioPAX available in the GKB and Protégé ontology tools.

### **Extensibility and the leveled approach**

The current ontology is far sighted in that it contains many classes that may not be chosen for inclusion in level 1. Some of these classes are very light on detail, like a sketch of the class. Because we would like to develop BioPAX via a leveled approach similar to SBML (<http://www.sbw-sbml.org/index.html>) and PSI MI (<http://psidev.sourceforge.net/>), we need to think about what types of data we want to represent in the future so that we can add more detail later without rendering previous versions incompatible as much as possible. Thus, certain classes are included in this version of BioPAX for discussion and to try to avoid making initial design decisions that are not extensible in the future (avoid "painting ourselves into a corner").

### **3. Glossary**

Some of the following definitions are only for the context of BioPAX and may not be general.

**Biological pathway:** A network of biological relationships.

**Data exchange format:** Any data format, usually electronic, used to exchange data.

**GKB:** Generic Knowledge Base Editor. <http://www.ai.sri.com/~gkb/>

**Ontology:** A system for describing knowledge, a conceptualization of a domain of interest usually made up of any or all of the following: concepts, relations, attributes, constraints, objects, values.

**OWL:** Web ontology language. <http://www.w3.org/TR/owl-features/>

**Protégé:** Protégé ontology and knowledge base editor. <http://protege.stanford.edu/>

**XSLT:** Extensible Stylesheet Transformations language. Allows conversion from one XML format to another. <http://www.w3.org/TR/xslt>