

APO-SYS: Recommended Data Exchange Formats, Ontologies and minimum Reporting Guidelines

Philip Jones, EMBL-EBI
 Christophe Roos, Medicel
 2008-04b

Table of Content

1	Introduction.....	2
2	Pathway Models.....	2
..2.1	Exchange format.....	2
..2.2	Minimum Reporting Guidelines.....	3
3	Microarray Data.....	3
..3.1	Microarray Data Exchange Formats.....	3
..3.2	Minimum Reporting Guidelines.....	3
..3.3	Ontologies and Controlled Vocabularies.....	3
4	Mass Spectrometry based Proteomics Data.....	3
..4.1	Proteomics Data Exchange Formats.....	3
..4.2	Minimum Reporting Guidelines.....	4
..4.3	Ontologies and Controlled Vocabularies.....	4
5	Molecular Interactions.....	4
..5.1	Data Exchange Formats.....	4
..5.2	Minimum Reporting Guidelines.....	4
..5.3	Ontologies and Controlled Vocabularies.....	4
6	Other data from high throughput screenings.....	5
..6.1	Data Exchange Formats.....	5
..6.2	Minimum Reporting Guidelines.....	5
..6.3	Ontologies and Controlled Vocabularies.....	5

1 Introduction

This document describes recommended solutions for data exchange and data reporting in the APO-SYS project. Typically, data exchange means import of measurement or other observation data to the Mediceal Integrator platform for analysis or data submission to the public databases hosted by EBI and others. These solutions for data exchange aim, as far as possible, to apply the use of public standards for each specific domain whilst at the same time complementing the plans and existing technologies being used in each work package.

The three categories of information in each section include:

1. suggested data exchange formats;
2. relevant minimum reporting guidelines;
3. relevant ontologies and controlled vocabularies.

Regarding minimum reporting guidelines, a very useful resource for searching for appropriate guidelines is the MIBBI project: <http://mibbi.sourceforge.net/> that aims to provide a central portal for biological and biomedical minimum reporting guidelines.

2 Pathway Models

..2.1 Exchange format

In **WP6** the BioPAX format is named as the format of choice for the integration of pathway maps and models from the partners. The BioPAX format (<http://www.biopax.org/>) is an OWL based format for biological pathway data. BioPAX is documented here: <http://www.biopax.org/release/biopax-level2-documentation.pdf>. The BioPAX format comprises pathway information (but not kinetic models) and molecular interactions, based upon the PSI-MI (molecular interaction) format¹ described below. BioPAX has recently split into two efforts, BioPAX-DX and BioPAX-OBO. It is of note that this WP6 suggests that models generated using Systems Biology Markup Language (SBML) should be converted to BioPAX to allow the models to be integrated in a single repository. It may perhaps be useful to consider maintaining SBML models in their original format, as any kinetic data will be lost in this conversion as discussed below.

SBML is currently supported by over 120 different software systems². The BioPAX homepage³ on the other hand currently lists a total of three: two implementations of BioPAX level 1 and one implementation of BioPAX level 2.

SBML has a different scope to BioPAX, allowing reaction kinetics to be modelled. BioPAX does not provide this support, although it does provide a more rich semantic model of reactions. It is stated in the WP6 'Objectives' section, regarding integration of models from the partners that "Depending on the source of the model, they may come under different native formats, the most common being...SBML and BioPAX." Perhaps the decision to encode the integrated models in BioPAX will therefore result in some information loss.

It is stated in WP6 that "Once published, the models will be submitted to the Reactome database for worldwide dissemination". (Final sentence of the 'Objectives' section).

Reactome is not a submission database in this sense and does not consume BioPAX models (although it does allow data export in this format). The BioModels database at the EBI does support submission of systems biology models (using either SBML or CellML formats) so this may be a more appropriate target for placing the models in the public arena.

¹ BioPAX is defined using OWL and PSI-MI is defined using XML schema, so BioPAX is not directly compatible with PSI-MI without some transformation.

² <http://sbml.org/index.psp>

³ <http://www.biopax.org/>

SBML is documented in detail on the SBML web site: <http://sbml.org>. This includes links to the schema, documentation of the format and various tools that can be used to manipulate SBML files.

The Medical Integrator has its own internal representation for pathway maps and models. It supports for example an integrated name-space for pathway nodes, hierarchical topology (different locations) and reaction kinetics. Medical is developing exchange mechanisms for SBML and BioPax import and export.

..2.2 Minimum Reporting Guidelines

Please see <http://www.ebi.ac.uk/compneur-srv/miriam/> for documentation of "Minimal Information Requested In the Annotation of biochemical Models" (MIRIAM).

3 Microarray Data

..3.1 Microarray Data Exchange Formats

The MGED Society (<http://www.mged.org/Workgroups/MAGE/mage.html>) currently recommends the use of MAGE-TAB (<http://www.mged.org/mage-tab/>) as the 'best practice approach' to reporting microarray experiment results (measurement data and meta-data covering experimental design and biological context). MAGE-TAB makes use of a series of simple flat file formats (tab separated or spreadsheet formats) that are used together to define a complete experiment. Also available from MGED is the MAGE-ML XML data exchange standard (<http://www.mged.org/Workgroups/MAGE/mage-ml.html>).

The Medical Integrator has a more fine-grained representation for the meta-data than MAGE-TAB. The Medical Experiment application, a LIMS-like application can be used for manually capturing the meta-data together with the measurement. The Medical Workflow application, an application for designing and enacting data manipulation experiments can be used for more automated import of large amounts of data. Import of measurements to the Medical Integrator supports tab separated or spreadsheet format tables (including MAGE-TAB) for reporting microarray experiment meta-data, provided they contain at least what is required by the MAGE-TAB format. The value tables (such as the CEL files in the case of Affymetrix data) are imported as such. Medical will provide tools for MAGE-TAB export, allowing for MIAME compliant datasets to be submitted to ArrayExpress at the EBI.

..3.2 Minimum Reporting Guidelines

These are defined in the well established "Minimum Information About a Microarray Experiment" (MIAME). See <http://www.mged.org/Workgroups/MIAME/miame.html> for further details.

..3.3 Ontologies and Controlled Vocabularies

MGED provide an ontology designed to provide standard terms to annotate microarray experiments. Full documentation can be found here: <http://mged.sourceforge.net/ontologies/index.php>. Medical Integrator implements some of these ontologies and more can be made available when necessary.

4 Mass Spectrometry based Proteomics Data

Public formats, guidelines and ontologies for proteomics data are defined by the HUPO Proteomics Standards Initiative (PSI).

..4.1 Proteomics Data Exchange Formats

Mass spectrometry data: mass spectra and processed peak lists. The currently available format from HUPO-PSI for mass spectrometer instrument output are described here: <http://psidev.info/index.php?q=node/>

80. The format currently implemented by the majority of mass spectrometer instrument vendors is **mzData 1.05**, documentation of which can be found at the hyperlink above. The mzData 1.05 XML schema can be found here: <http://www.psidev.info/docstore/mzdata.xsd> and is documented here: http://www.psidev.info/files/mzData_1.05_spec.doc.

HUPO PSI are currently in the process of finalizing the replacement for mzData, which will be called mzML. Full documentation of this format is given here: <http://psidev.info/index.php?q=node/257>.

Mass spectrometry search engine output formats:

HUPO-PSI is still in the process of finalizing its first format for search engine output, which is currently called analysisXML. Details of the development of this format can be found here: <http://psidev.info/index.php?q=node/105>. The procedures for data import to the Mediceal Integrator will need to be further specified. In general terms, it will be analogous to microarray data import (section ..3.1 above). Mediceal will develop tools for export of the data following the evolving recommendations referred to above.

..4.2 Minimum Reporting Guidelines

HUPO-PSI has developed "The Minimum Information About a Proteomics Experiment" (MIAPE). This is a series of documents under the umbrella of a single parent document, that provides reporting guidelines for different aspects of proteomics. MIAPE is documented, with links to the individual guidelines at: <http://www.psidev.info/index.php?q=node/91>. The Mediceal Experiment and Workflow applications (see ..3.1) can be used for manual and automatic data import respectively.

..4.3 Ontologies and Controlled Vocabularies

Both mzData and mzML make use of the PSI ontology for mass spectrometry, which is documented at http://psidev.info/index.php?q=wiki/Mass_Spectrometry and is available in OBO format from <http://psidev.sourceforge.net/ms/xml/mzdata/psi-ms-cv-latest.obo>. (This version is intended for use with mzData 1.05. A new version is being published to support mzML.). This ontology will be made available on the Mediceal Integrator platform.

5 Molecular Interactions

..5.1 Data Exchange Formats

PSI offers a very mature standard for the exchange of molecular interaction data: PSI-MI, currently at version 2.5. This format is described and documented at: <http://psidev.info/index.php?q=node/31>. The procedures for data import to the Mediceal Integrator will need to be further specified. In general terms, it will be analogous to microarray data import (section ..3.1 above). Mediceal will develop tools for export of the data following the evolving recommendations referred to above.

..5.2 Minimum Reporting Guidelines

The Molecular Interactions community has developed a set of guidelines "The Minimum Information about a Molecular Interaction eXperiment" (MIMIx). At the time of writing, this document is still under review via Nature Biotechnology. Brief documentation and a link to the guidelines in their current form can be found here: <http://psidev.info/index.php?q=node/103>. The Mediceal Experiment and Workflow applications (see ..3.1) can be used for manual and automatic data import respectively.

..5.3 Ontologies and Controlled Vocabularies

To complement the PSI-MI format, a mature ontology for molecular interactions is available that is also documented and linked from <http://psidev.info/index.php?q=node/31>. This ontology will be made available on the Mediceal Integrator platform.

6 Other data from high throughput screenings

..6.1 Data Exchange Formats

Other measurement technologies will be applied in APO-SYS than those listed above. A major category includes cell screening experiments with large molecule libraries (e.g. siRNA), tissue microarray (TMA) and other imaging data. These technologies are not covered by the recommendations listed so far. Therefore, the partners need to pay attention to evolving best practice recommendations, so that they can be applied in APO-SYS. The procedures for other data import to the Medice1 Integrator will need to be further specified. In general terms, it will be analogous to microarray data import (section ..3.1 above).

..6.2 Minimum Reporting Guidelines

The appropriate guidelines need to be identified when they become available. At the time of writing, there are no recommendations for example for siRNA screens. Nevertheless, the partners of APO-SYS need to internally agree on some minimum reporting recommendations. The Medice1 Experiment and Workflow applications (see ..3.1) can be used for manual and automatic data import respectively.

..6.3 Ontologies and Controlled Vocabularies

It is probable that the controlled vocabularies necessary to APO-SYS are already available at <http://obofoundry.org/>. The relevant ontologies will be made available on the Medice1 Integrator platform as required.