Metabolomics Standards Initiative (MSI)

Group E: Data Exchange - Roadmap

Sponsor: Metabolomics Society [http://msi-workgroups.sourceforge.net/]
Reference: http://msi-workgroups.sourceforge.net/exchange-format/roadmap/
Version: 1.5
Date: 2006/03/20 12:07:33
Authors: Nigel Hardy
Chris Taylor

Table of Contents

1. Status of this document

1. This document was first drafted by Nigel Hardy as a proposal for the management and planning of this workgroup's activity. It has been reviewed and extended by Chris Taylor. It is intended as a basis for discussion and will likely continue to evolve.

2. Brief

2. *"This group will define data exchange formats and produce a schema for such operations that cover all aspects of the metadata, the analytical data (both spectroscopic and chromatographic) and the data analysis."* (Taken from "Draft proposal" of 2005-08-31")
3. ***Commentary*.** Existing general schemas such as ArMet [1]and FuGE [2] will be included in this group's considerations. Proposals from the 'MetaboMeetings' will also be factored in, such as the re-use of appropriate schemata such as mzData [3] and the cross-domain collaborative development of formats such as spML [4]. This group will coordinate with the other MSI working groups to receive reporting requirements and to validate draft formats and reference implementations. Databases for reference compound spectra, or links to them, will also be included.

3. Intended deliverables

4. A specification of the data for exchange ("the schema").
5. A reference implementation("the format"), intended for general application. Whatever the nature of the ultimate product, it must reflect the scope of "the schema" in the sense of 4. (above). Note that "the format" is likely to consist of a suite of well decoupled modules.
6. Documentation of the standard. It is not clear how the initiative as a whole will document its output, but whatever is agreed this working group must conform to this so as not to duplicate work. In particular, the metabolomics ontology is expected to provide definitions, controlled value lists and an explanatory framework for users of the format.

Note

A matter for discussion with other groups.

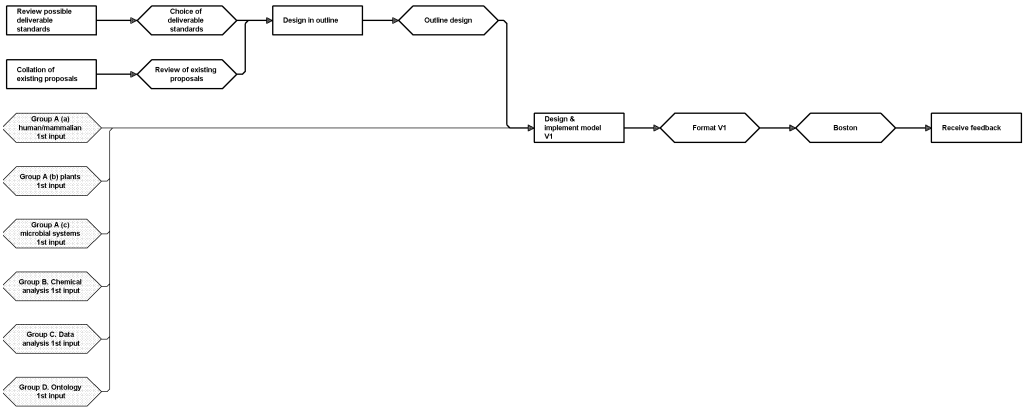
7. Working demonstration applications; including but not limited to:
   a. Datasets marked up in the XML, viewable on the web via suitable XSLT, CSS or similar.
   b. A simple data entry system to create the markup.

4. Relationships to other working groups

3. **Group A. Biological sample context.** It is assumed that this group will produce a report listing the necessary data items, constraints on their values and co-occurences and perhaps groupings of data items.
4. **Group B. Chemical analysis.** It is assumed that this group will produce a report listing the necessary data items, constraints on their values and co-occurences and perhaps groupings of data items.
5. **Group C. Data analysis.** It is assumed that this group will produce a report listing the necessary data items, constraints on their values and co-occurences and perhaps groupings of data items.
6. **Group D. Ontology.** It is assumed that this group will produce an ontology in machine readable format. This will need to accomodate (at least) all terms and struturing specified by Groups A, B and C.
7. The work of Group E can therefore be seen as implementing the results of Groups A, B and C as structured by the work of Group D. Group E output can therefore only be properly validated once these groups' products are available, though the whole process is likely to benefit greatly from iteration and as such should proceed in parallel.

5. Group Tasks

8. **Outline Pert Chart.**



9. The gross plan is:
   a. Review existing standards in preparation for input from other working groups
   b. Receive input and develop (from existing standards where possible) a first draft format
   c. Present the format at Boston
   d. Receive feedback (including via other working groups) and work towards a standard.
10. Important milestone documents in this process are:
    - **Choice of deliverable standard.** The deliverables of this workgroup will be "data exchange formats" and "a schema". The languages and technologies for these deliverables must be chosen.

      Progress to date. None. This is unlikely to be time-consuming or contentious. See Section 6.
    - **Review of existing proposals.** Existing standards of potential use should be cataloged, reviewed in this context and assessed for possible suitability. This will provide a basis for an initial outline design suitable for validation and enhancement in the light of detail from other working groups.

      Progress to date. Work has centred on MS and NMR data. Documents are lacking. The

full scope needs to be included. See Section 7.

- **Format V1.** The first versions of the deliverables.

Progress to date. Not relevant to start.

6. Choice of deliverable standard

The group is to deliver "a schema" and "a format" as defined in points 4 and 5. Practical decisions about the technologies (languages and possibly tools) to be used for these are yet to be made. The obvious solutions are UML for the schema amd XML for the format. A remaining issue is the XML schema language, with XMLSchema, RelaxNG and Schematron as initial candidates. XMLSchema has the greatest currency, but may lack power in some areas. Use of more than mechanism in the standard is a possibility, but uniformity makes for simplicity.

7. Review of existing proposals

Existing standards, which may be purposed either for data transport or as preferred referents, range from the general to the specific. Broad-scope models such as ArMet, FuGE and MAGE, which are focused on data transport, contrast with standards for identifying chemical moieties such as InChI or ChEBI. Previous work on content, such as SMRS and MIAMET, may offer some contribution to gross structure. The following is a working list of entities to consider. The group would welcome additional suggestions.

Data exchange formats/models	**Reference 'knowledgebases'**	**Reporting specifications**
FuGE [2]	InCHI [5]	SMRS [6]
MAGE [7]	ChEBI [8]	MIAMET [9]
ArMet [1]	LipidMaps [10]	MIAPE [11]
mzData ???		MIAME [12]
analysisXML [13]		MIGS [14]
spML [4]		MIRIAM [15]
CCPN [16]		ArMet [1]
LSID [17]		CDISC/SEND [18]
CDISC/SEND [18]		NMR-STAR [19]
NMR-STAR [19]		

References

[1] http://www.armet.org/

[2] http://fuge.sourceforge.net/

[3] http://psidev.sourceforge.net/ms/#mzdata

[4] http://psidev.sourceforge.net/gps/#psiFormats

[5] http://www.iupac.org/inchi/

[6] http://smrsgroup.sourceforge.net/

[7] http://www.mged.org/Workgroups/MAGE/mage.html

[8] http://www.ebi.ac.uk/chebi/

[9] Bino, R. J. and Hall, R. D. and Fiehn, O. and Kopka, J. and Saito, K. and Draper, J. and Nikolau, B. J. and Mendes, P. and Roessner-Tunali, U. and Beale, M. H. and Trethewey, R. N. and Lange, B. M. and Wurtele, E. S. and Sumner, L. W.. *Potential of metabolomics as a functional genomics tool*. *Trends In Plant Science*. 9. 9. 418-425. 2004.

[10] http://www.lipidmaps.org/

[11] http://psidev.sourceforge.net/gps/#miape

[12] http://www.mged.org/Workgroups/MIAME/miame.html

[13] http://psidev.sourceforge.net/ms/#analysisXML

[14] http://www.genomics.ceh.ac.uk/genomecatalogue/migs.php

[15] Novere, Nicolas Le and Finney, Andrew and Hucka, Michael and Bhalla, Upinder S. and Campagne, Fabien and Collado-Vides, Julio and Crampin, Edmund J. and Halstead, Matt and Klipp, Edda and Mendes, Pedro and Nielsen, Poul and Sauro, Herbert and Shapiro, Bruce and Snoep, Jacky L. and Spence, Hugh D. and Wanner, Barry L.. *Minimum information requested in the annotation of biochemical models (MIRIAM)*. 23. 12. 1509. 2005.

[16] http://www.ccpn.ac.uk/

[17] http://lsid.sourceforge.net/

[18] http://www.cdisc.org/models/send/v2/

[19] http://bmrb.protein.osaka-u.ac.jp/formats.html