

International Phenome Database Integration Workshop
Barcelona, 25 February 2006
DISCUSSION SUMMARY

Initial presentations:

MGI are currently starting to build an infrastructure/repository for data from sunsetting mutagenesis projects.

RIKEN have or are developing a number of things that may be useful to this group including:

- A structured terminology of experiments
- An SOP page
- Genetic Mapping Markup Language
- Visualization tools

There is a need for a means of phenotype data exchange (XML schemata) and consistent nomenclature.

Tools and structures should be as simple as possible to ensure uptake.

A: Pilot study to link 3 existing databases – MGI, MPD & EuroPhenome

Build simple portal

A portal could initially operate at a very simple level i.e. as a collection of links. We would subsequently aim to develop this to allow joint querying of databases and information about database status. It was noted that Jax are setting up web services to their data.

Define questions to determine queries

“Find me a mouse next door”

Largely through the current IMSR system, possibly extended to include more small labs.

“Find me everything about my phenotype of interest”

MGI provides indirect access to information about mouse lines showing a particular phenotype including their locations (via IMSR). A possible way of linking to other relevant material would initially be via the top level of the MP (Phenoslim), with some terms removed and possibly some split (pathology to be added? others?). SOPs could be annotated with these top level terms so that SOPs dealing with a particular system could be readily searched, with the potential to link to baseline data. Similar linking would be possible to categories of phenotypes identified in mutagenesis projects (RIKEN, Jax, ...).

Harwell to take lead

B: Subgroups were established to discuss two main areas: ontologies and data exchange. The aims of the groups were to define short-term and medium-term objectives and to define a group of people who would take their areas forward in the future.

1. *Ontology subgroup* to compare existing ontologies and ontologies being developed – harness individual expertise

Initial discussions concerned what was required, e.g. a husbandry ontology, environment ontology. There is a need to define what we require the ontology to do. It was therefore decided to use the high level of MP (phenoslim) and to test PATO to identify how useful it is, and also to get an update of whats happening with PATO from George and Michael.

Short-term aim:

Annotate EMPress SOPs with phenoslim. Phenome Database to be asked to do the same.

Medium-term aims:

- 1: Testing PATO and MP – **Kirsty Lee** to do this
- 2: Look at Assay Ontology
- 3: Define what ontologies are required

Working group core

Additional people would be representatives from PATO [NB: George Gkoutos was invited to the meeting but could not attend].

2. *Data exchange subgroup*

Phenotype databases should be interconnected. We discussed possible interactions between the various databases and decided on something simple to start:

- 1) Update: Each database should generate a simple RSS feed to tell the world its current status, when it was updated and how much data has changed
- 2) We should have an Interface which will allow access to the database via computer programs. This interface should be used by the prototype page developed at Harwell.

We need a vocabulary for the type of database.

We need a list of potential query language components to get data from the database

We have to know the data type that will be returned from the call.

We have to know if it is a primary database with "authority" or a secondary database.

First we want to keep it simple and again would like to display the current status of the database.

An additional important area is to define a minimum requirement for the description of a phenotyping experiment. This should consist of an assay description (SOP) and additional metadata. (NB: some of this has already been defined in the course of EUMORPHIA data acquisition, but more may be needed, particularly information on environmental conditions, handling, feed etc..) Harwell and MPD have already discussed harmonizing their SOP formats, which can be built into Harwell's draft SOPML. However, other sites also hold SOPs

and it will be important to bring their experience into the same framework. We will therefore circulate the results of the Harwell-MPD discussions to the rest of the group for further comment and development.

Harwell MPD to discuss

A further development could be an “assay ontology” - structures of this kind are being developed at RIKEN and Harwell. This overlaps significantly with a phenotype ontology as currently defined but serves a useful purpose by grouping together assays with a similar purpose. The Harwell version currently includes EUMORPHIA and MPD SOPs.

C: Future meetings. It was agreed to hold a further meeting in September 2006. In between the groups would schedule telephone conferences as appropriate to help drive progress forward.

D: The group would aim to produce a position statement, possibly for publication (in as high profile a journal as possible).

JMH to coordinate

E: Other general points

- The use of universal IDs (GUIs) would greatly facilitate searching for all information on a mouse line. Jax already issues identifiers but some lines may be represented in public databases but not have IDs. Some mechanism may be needed to extend the system to lines of this type. A single SOP database could similarly issue IDs for experimental methods.
 - List of IDs (authoritative where possible) for other entities (e.g. probes) and how they relate to experimental flows. Including descriptions.
 - Remain cognisant of UID issues

NB: Judy Blake to talk about what we are doing to Mark Musen *et al* with a view to an application for funding.

Immediate Action Plan

Working groups to return their summaries to JMH within the next week. **JMH** to compile report and circulate to all participants.

Speakers to send their talks to **JMH**. These will be made available either by email or a web site.

Web site to be established (**Harwell**)

JMH to investigate holding a follow-up meeting at IMGC in September 2006.

Harwell to coordinate portal project. Initial step is to set up static web site (links to DBs) then develop this to address questions suggested by the working group.

Harwell to label their SOPs with PhenoSlim (presumably using only a subset) to allow them to be linked to phenotype queries at Jax.

RIKEN also to label their data with a subset of PhenoSlim.

Harwell/MPD to discuss SOP composition and enhance SOPML schema accordingly. This will then be circulated to the group to see if any further additions are needed. This should lead into a discussion of Minimum Information for description of a Phenotyping Experiment, to be developed at the next meeting of the group.

MGI: To encourage submissions to IMSR from small labs, and encourage use of MGI strain identifiers in publications

Harwell to look at George Gkoutos' "assay ontology", specifically to see if it disambiguates individual assays within an SOP but also with a view to eventual release after consideration by the whole group. Also to consider its relationship to the RIKEN experiment hierarchy.

Edinburgh (Kirsty/Duncan) to develop tests for PATO and PhenoSlim.