Summary of Discussions at the RIKEN/Interphenome/CASIMIR meeting on Data Exchange, 12-13 July 2009

Attendees: Kuniya Abe, Joe Banerjee, Kaoru Fukami, Michael Gruenberger, John Hancock, Norio Kobayashi, Hiroshi Masuya, Hugh Morgan, Kazuo Moriwaki, Yuichi Obata, Joel Richardson, Paul Schofield, Toshihiko Shiroishi, Toyoyuki Takada, Tetsuro Toyoda, Nora Tsao, Shigeharu Wakana, Atsushi Yoshiki.

Part 1 – Morning Session

1. The first point discussed was the appropriate licensing conditions for the exchange of data. RIKEN informed the meeting that the Japan Integration Database Project recommends the Creative Commons Attribution Share Alike licence. Paul Schofield summarised the discussions at a recent CASIMIR meeting in Rome at which it was recommended that the ideal license for data sharing was CC0 (http://wiki.creativecommons.org/CC0_FAQ). It was suggested that the Creative Commons Share Alike Non-Commercial license (http://creativecommons.org/licenses/by-nc-sa/3.0/) may be appropriate for the future agreement between CASIMIR and RIKEN. It is notable that EuroPhenome and MGI do not currently impose any explicit license and allow free use of all their data, which may be equivalent to CC0.

It was agreed that professional legal advice would be needed before any formal licensing framework could be agreed.

2. The importance of attribution of data presented in any unifying portal was pointed out and the value of some microattribution framework to allow measurement of the usage, and therefore value to the community, of data was agreed. RIKEN is now trialing some tagging technologies to launch microattribution.org in the future.

3. There then followed a discussion on the basic elements needed for shallow integration of phenotype data (at the level of summary descriptions of mouse lines). The following elements were regarded as needed:

- Genotype. This can be made up of four main elements:
    - Gene – to be identified by its MGI ID and/or symbol. This may be absent if no gene is available in a given line, e.g. unmapped ENU lines. It might be replaced by a genomic region or chromosome when these are known (pastial mapping)
    - Alleles. Both alleles should be identified so that it is explicit if one of them is wild type [NB – how does MGI nomenclature deal with this?]. If no gene is identified an MGI phenotypic allele designation can be issued.
    - Zygosity. This could be inferred from the allele composition
    - Genetic background. Identified by the standard strain name.
- Evidence code. This was to be an evidence code indicating how the annotation of a line was obtained. This set of codes would require further discussion but the following were suggested:

1. From a high throughput phenotyping experiment – automatically annotated
2. Annotated manually from literature
3. Submitted by the originators
4. Inferred by orthology/homology

- Phenotypic description. This should be framed using ontology terms. Abnormal phenotypes should always be summarised using MP (Mammalian Phenotype) ontology. These may be accompanied by EQ descriptions, which could also be used to describe normal phenotypic states. RIKEN raised the lack of "normal" phenotypes in MP and Joel agreed to raise this with the curators.
- Citation – for published phenotypic descriptions, in the form of a PMID [NB: should these be associated with individual phenotypic descriptions?]
- Data resource – where to see more details about the phenotypic description if these exist, e.g. raw data in EuroPhenome or at RIKEN. Could be in the form of a URL or URI.
- Animal Resource – where the line (ES cells, sperm, embryos, mice) can be obtained from. Again could be in the form of a URL or URI. There was a discussion about whether IMSR or local repository IDs could be supplied but there was uncertainty about whether IMSR IDs were readily accessible to or known by all the databases. Local IDs would have to be stable to allow end users of the data to access the correct strain.
- Data Source. Which database the data came from and who generated the original data. There was a discussion on whether UIDs are needed for databases and researchers. Databases might change names; researcher names can be ambiguous. It was agreed to start with text for these descriptions (which could be generated by the database providing the data) with the proviso that this might migrate to commonly acceptable IDs in future. This section can form the basis of any microattribution system that might be implemented. There is a need for users of data to cite the origins of data as a matter of course, and for Data Source tags to be indelibly linked to data so that this information cannot be lost.
- License type. It was discussed that the assignments of the same license level (e.g. CC by-nc-sa) to all the (shallow) datasets on the portal may maximize usability. However it was unclear to the group whether a license applied to a particular item of data or only to a collection. It might be possible to use a form of words such as "This item of data is part of a collection licensed under…" but legal advice is needed to clarify this point further.

5. The final item discussed in the morning session was the technology that should be used for shallow integration. RIKEN agreed to export their data using SOAP and XML but would then reimport data from the portal to provide an RDF channel. This RDF would be available to the community as an additional resource for integration.

Action points

1. Hugh and Michael to draw up an outline specification for the required XML for exchange of data at this level

2. All sites to implement availability of this data via SOAP web services. There would be a need to support searches on various items, initially phenotype and gene. There would also a need to specify a date range of data to be exported, e.g. all, since some set date, last month, last week etc. WSDL statements would need to be produced to describe these web services.
3. A concrete aim would be to have a demonstration of shallow integration ready for IMGC in November. This could be presented to the InterPhenome group and perhaps also at the main conference. Once a working system was established we would aim to produce a high-profile publication.

Part 2 – Afternoon Session

This was concerned with two main areas – the relationship between YATO and PATO, and how to proceed towards deep integration (the exchange of detailed phenotype data).

5. PATO-YATO relationship. YATO contains PATO terms but in a different structure subsumed under a top level ontology. The terms should be directly mappable. Phenotype descriptions in EuroPhenome using PATO are of the form E+Q. YATO-based descriptions are of the form EAV (Attribute separate from Value). Systems using PATO in EQ statements or YATO phenotype statements would therefore have their own, ontology-derived understanding of the meaning (relationships) of those terms.

Actions:
1. It was agreed to make the EQ annotations and the equivalent MP terms in EuroPhenome available to RIKEN
2. EuroPhenome would also investigate how easily YATO annotations could be converted to EQ.

6. Future work on Deep Integration. For full exchange of data a detailed specification of what is to be exchanged, including descriptions of control data, statistical analysis etc, will be needed. This would be proceeded with once shallow integration had been effectively tackled, perhaps starting at IMGC. This data exchange might follow a similar pattern to that for shallow exchange, e.g. RIKEN exporting data as XML and importing data into their RDF database, but this is still to be discussed fully. In addition, the necessity of several improvements to the existing ontologies is suggested to facilitate deep integration was discussed (e.g . MA).