# Integration of Mouse Phenome Data Resources

**The Mouse Phenotype Database Integration Consortium**[*]

## Abstract

Understanding the functions encoded in the mouse genome will be central to an understanding of the genetic basis of human disease. To achieve this it will be essential to be able to characterise the phenotypic consequences of variation and alterations in individual genes. Data on the phenotypes of mouse strains are currently held in a number of different forms (detailed descriptions of mouse lines, first line phenotyping data on novel mutations, data on the normal features of inbred lines, etc.) at many sites worldwide. For the most efficient use of these data sets, we have initiated a process to develop standards for the description of phenotypes (using ontologies), and file formats for the description of phenotyping protocols and phenotype data sets. This process is ongoing, and needs to be supported by the wider mouse genetics and phenotyping communities to succeed. We invite interested parties to contact us as we develop this process further.

## Introduction

With the advent of complete or nearly complete genome sequences for the major model organisms we are embarked upon a project to understand the roles of individual genes and to synthesise this knowledge into an understanding of the biological systems in which these genes participate (Brown et al. 2006). The mouse plays a central role in this project because of its status as the primary mammalian model organism and its close relationship to humans, which means that it is the best available model for many human diseases.

[*]The Mouse Phenotype Database Integration Consortium currently comprises: John M. Hancock[1], Niels C. Adams[2], Vassilis Aidinis[3], Andrew Blake[1], Judith A. Blake[4], Molly Bogue[4], Steve D.M. Brown[1], Elissa Chesler[5], Duncan Davidson[6], Christopher Duran[1], Janan T. Eppig[4], Valérie Gailus-Durner[7], Hilary Gates[1], Georgios V. Gkoutos[8], Simon Greenaway[1], Martin Hrabé de Angelis[7], George Kollias[3], Sophie Leblanc[9], Kirsty Lee[6], Christoph Lengger[7], Holger Maier[7], Ann-Marie Mallon[1], Hiroshi Masuya[10], David G. Melvin[2], Werner Müller[11,12], Helen Parkinson[13], Glenn Proctor[13], Eli Reuveni[14], Paul Schofield[15], Aadya Shukla[16], Cynthia Smith[4], Tetsuro Toyoda[10], Laurent Vasseur[9], Shigeharu Wakana[10], Alison Walling[17], Jacqui White[2], Joe Wood[17], Michalis Zouberakis[3]

1. MRC Mammalian Genetics Unit, Harwell, Oxfordshire OX11 0RD, UK
2. Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK.
3. BSRC Fleming, 34 Fleming Street, 16672, Vari, Athens, Greece
4. The Jackson Laboratory, 600 Main Street, Bar Harbor, ME 04609, USA
5. Oak Ridge National Laboratory, Mammalian Genetics & Genomics Group, Biosciences Division, Oak Ridge, TN 37831-6445
6. MRC Human Genetics Unit, Edinburgh, EH4 2XU, UK
7. GSF – National Research Center for Environment and Health, Ingolstaedter Landstrasse 1, Institute of Experimental Genetics, D-85764, Neuherberg, Germany
8. Department of Genetics, University of Cambridge, Cambridge CB2 3EH, UK
9. Institut Clinique de la Souris, BP10142, 67404 Illkirch Cedex, France
10. RIKEN Genomic Sciences Center, 3-1-1 Koyadai, Tsukuba, Ibaraki, Japan
11. Helmholtz-Zentrum für Infektionsforschung GmbH, Inhoffenstraße 7, D-38124, Braunschweig, Germany
12. Faculty of Life Sciences, University of Manchester, Manchester, M13 9PT, UK
13. European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK
14. Mouse Biology Unit, EMBL Monterotondo Outstation, via Ramarini 32, 00016 Monterotondo, Rome, Italy
15. Department of Physiology, Development and Neuroscience, University of Cambridge, Downing Street, Cambridge CB2 3DY UK
16. Oxford University Computing Laboratory, Wolfson Building, Parks Road, Oxford, OX1 3QD, UK.
17. MRC Mary Lyon Centre, Harwell, Oxfordshire OX11 0RD, UK

To understand the roles of individual genes, we need to be able to characterise the phenotypic consequences of mutating (via ENU or other mutagens), knocking out or otherwise modifying individual genes and of natural variation in these genes. This is giving rise to the idea that large scale phenotyping centres need to be established, alongside the experimental resources to generate mutations and knockouts in all mouse genes (reviewed by Brown et al. 2006). There need to be public data resources which collect together phenotypic data on both mutant mice and wild-type inbred strains to allow quantification of natural trait variation and whether a given observation in a given mutant line deviates significantly from expectation given the genetic background upon which the mutation was analysed. In order to compare results obtained by different centres, collections of well-characterised and reproducible protocols for mouse phenotyping, such as the EMPReSS resource (http://empress.har.mrc.ac.uk) developed by the EU-funded EUMORPHIA consortium (Brown et al. 2005), need to be in place. These data resources should be openly available via the world-wide web and should be linked to other genomic and functional genomic resources to allow a thorough understanding of the phenotypes, how they were measured, and deeper analysis of molecular processes that underlie any given phenotype. It should also be possible to make the data seamlessly accessible through web interfaces, allowing joint mining and analysis of the data. This requires the establishment of well-structured, curated, open source and appropriately funded databases and portals to provide this information to the mouse community.

A characteristic problem of biological databases is the emergence of different databases containing similar but not identical data at different sites internationally. The classic example of this and its eventual resolution are the GenBank, EMBL and DDBJ sequence databases which eventually developed a data sharing model whereby all three databases effectively merged (Brunak et al. 2002). What follows are the conclusions of two discussion meetings held in Barcelona on 25th February 2006 and Munich on 9th September 2006, which initiated a process of integrating as far as possible the current (and future) mouse phenotype resources.

## Current Resources

For the discussion of current resources we distinguish three types of data - data characterising a wide range of phenotypes in established mutant mice compared with their normal controls, data collected as phenotypic screens to discover new mutations produced via mutagenesis (organised either as formal databases or presented via web sites), and data characterising normal phenotypic parameters across inbred strains (Figure 1). These distinctions are somewhat artificial, reflecting our inability to experimentally measure all phenotypes in all mice, both technically and practically, and our need to analyse data from the perspective of mutations and normality.

### Characterising phenotypes in mutant mice compared to normal controls

The Mouse Genome Database (MGD, http://www.informatics.jax.org) aims to integrate data on all phenotypic mutations known in the mouse (Eppig et al. 2005; Blake et al. 2006). Data are gathered from the scientific literature and direct submissions from individual researchers and mutagenesis centres. Data are annotated with Mammalian Phenotype (MP) Ontology

terms (Smith et al. 2005) to enable integrated searches for phenotypes across all mouse mutations. Key to the context of the phenotype is the genotype, comprising the allelic composition of mutations carried by a particular mouse cohort and the genetic background on which the phenotypes were analysed. In addition, MGD has recently made available phenotypic data on a set of knockout mice created and characterised by Deltagen Incorporated and Lexicon Genetics Incorporated that are being deposited in public repositories for research use as part of the NIH-funded mouse repatriation process (see http://www.nih.gov/science/models/mouse/deltagenlexicon/factsheet.html). These mutants are also integrated into MGD and annotated with MP terms to make them searchable in the context of all other known mouse mutants. As of September, 2006, over 66,100 annotations to MP terms had been curated to over 17,500 genotypes.

### Data from phenotypic screens to discover new mutations

A number of major phenotyping centres present information via their web sites about mutants discovered during their mutagenesis screens. Examples include the Harwell (http://www.mgu.har.mrc.ac.uk/mutagenesis/access/), Baylor College of Medicine (http://www.mouse-genome.bcm.tmc.edu/ENU/MutagenesisProj.asp), Oak Ridge National Laboratory (http://bio.lsd.ornl.gov/mgg/resources.html) and RIKEN (http://www.gsc.riken.go.jp/Mouse/) resources (see http://www.informatics.jax.org/mgihome/other/phenoallele_commun_resource.shtml for a fuller listing). Summary information on mutant lines is primarily pre-publication data or data on lines that are made available to the mouse community for further experimentation. This summary information is typically provided in the form of a free text description of the main features of the mutant phenotype. This information, while useful to researchers browsing a particular web site, could be made more useful for computational integration through community adoption of standard vocabularies to describe phenotypes (see section on Ontologies, below). MGD has integrated many of these mutations into their phenotypic data and provided MP annotations to facilitate searching and computational analysis of these mutations. Potentially these mutagenesis centre databases also contain underlying phenotype data on individual mice as well as summary information, although this is not usually made available.

### The Phenotypic Characteristics of Inbred Mouse Lines

Databases containing characteristics of inbred strains of mice are becoming an increasingly important resource for mouse geneticists. These databases serve two major purposes. First, they provide baseline data for the characterisation of mutation effects. Well established and robust estimates of trait values are critical to successful detection of extreme alterations. Second, they allow comparison and genetic correlation of complex traits across diverse populations. Currently there are several significant resources which contain data characterising normal parameters among inbred mouse strains. Examples of this growing body of resources include: the Mouse Phenome Database (Grubb et al. 2004) based at the Jackson Laboratory, MuTrack at the Oak Ridge National Laboratory (Baker et al. 2004), the EuroPhenome database (Mallon et al, unpublished) based at MRC Harwell, PhenoSITE at RIKEN, and GeneNetwork at the University of Tennessee (Chesler et al. 2004).

The Mouse Phenome Database (http://www.jax.org/phenome) is the database of the Mouse Phenome Project (Bogue and Grubb 2004), which aims to gather quantitative phenotype data on a large set (up to 40) of standard inbred strains. The aim of collecting data from a large number of strains is to provide broad coverage as a community resource and to allow mining of the data for correlations between phenotypic measures across strains. A feature of the database is that each data collection is associated with a protocol which describes how the data were generated. The project also provides online analysis tools to allow identification of correlations within its data set.

GeneNetwork (http://www.genenetwork.org), encompassing WebQTL, is a database of genotypes and complex phenotypes ranging from gene expression to behaviour in standard inbred strains, and six panels of mouse recombinant inbred strains including the two largest sets (BXD and LXS) of approximately 80 strains each. Rat and Arabidopsis populations are also represented. Approximately 1500 phenotypes spanning the 25 year history of these strains are incorporated in this public resource, many of which were retrieved from the literature. All phenotypes are integrated with an analytic engine for basic statistics, multivariate and genetic analysis (Chesler et al. 2004). Phenotype records in this database reference the publications from which they are drawn. Integration to other phenotype resources is a key step in enhancing the usefulness of this resource.

Data currently in EuroPhenome (http://www.europhenome.org) result from applying the Standardised Operating Procedures (SOPs) making up EMPReSS (Brown et al. 2005; Green et al. 2005) to four inbred mouse strains. EMPReSS is a set of standardised and validated SOPs for large-scale mouse phenotyping. EuroPhenome data are validated across a number of phenotyping laboratories. EuroPhenome also contains data from experiments which produce qualitative as well as quantitative data and is essentially protocol-centred. It is planned to build on EuroPhenome to include phenotyping data on knockout mouse lines produced by the EUCOMM project (http://www.eucomm.org/) during the EUMODIC programme (http://www.eumodic.org).

The MuTrack system (Baker et al. 2004) (https://www2.tnmouse.org/mutrack/stats/Statistics.php) was developed for the Tennessee Mouse Genome Consortium's effort in the NIH Neuromutagenesis Program (Goldowitz et al. 2004), and it is still in use today for a variety of studies of complex phenotypes. The database contains trait data for several hundred phenotypes including common inbreds, consomics, 80 BXD recombinant inbreds, hybrids, and over 60,0000 mutagenised mice including ENU mutants and several knockout lines. SOPs are employed for phenotypic data acquisition. This publicly accessible database is an excellent example of one that can be made significantly more valuable to the community with a standard in place for the reporting of these protocols.

PhenoSITE (http://www.gsc.riken.go.jp/Mouse/phenotype/top.htm) provides baseline phenotype data for three inbred strains and their F1 hybrids. Data were generated by analyses using a comprehensive phenotyping platform developed in the mouse mutagenesis program in RIKEN GSC. SOPs of the phenotyping platform are also posted on the website. PhenoSITE also contains phenotype annotation of ENU-induced mutant mouse strains

generated in RIKEN GSC. The annotation is based on multiple ontologies such as MP and mouse adult gross anatomy.

### Access to Mutant and Inbred Strains

In addition to accessing data about mutant and inbred strains, the strains themselves must be physically accessible to the research community for further experimentation. The International Mouse Strain Resource (IMSR, http://www.imsr.org) is a searchable online database of mouse strains and stocks available worldwide, including inbred, mutant, and genetically engineered mice (Strivens and Eppig 2004). Here repository sites and consortia, as well as individual laboratories, that distribute mouse resources as live stock, cryopreserved embryos or gametes, or ES cell lines can list their available holdings. All major public repository sites contribute their listings to IMSR, which currently contains listings from 16 repositories and repository consortia, comprising 24 repository sites in the U.S., Canada, Europe, Japan, and Australia.

These major repositories have recently formed an international organization, the Federation of International Mouse Resources (FIMRE, http://www.fimre.org) with the goals of coordinating repository centres to meet research demand for genetically defined mice and ES cell lines, establishing consistent high quality animal health standards, providing genetic verification and quality control for mouse resources, and providing training to enhance utilisation of cryopreserved resources (FIMRe Board Of Directors 2006).

## Essential Components for Integration of Mouse Phenotype Databases

We have identified three main areas which need to be addressed to enable and support the integration of mouse phenome resources internationally. These are:

- Data description standards (ontologies and vocabularies). The need to store phenotype data in a human-comprehensible and computationally-accessible ontological structure drove the development of the MP (Smith et al. 2005). The need to capture individual data measurements on individual mice has given rise to the EAV (Entity+Attribute+Value) approach (Gkoutos et al. 2005) and its derivative, the EQ (Entity+Quality) approach (http://www.bioontology.org/wiki/index.php/PATO:Main_Page), making use of PATO (the Phenotypic Quality Ontology). Although these systems represent different perspectives on the description of phenotype information, cross-referencing of terms between these ontologies is a goal. In addition there is a need to standardise on other vocabularies that provide supporting data for phenotypic information and to identify any new ontologies that may be required.

- Phenotyping Protocols. Several websites, including those for the Mouse Phenome Database, EMPReSS, Mutrack and PhenoSITE, make phenotyping protocols available. There is a need for developing standard vocabularies for naming protocols and the common data elements within them to foster global understanding of methods and provide a single framework allowing protocols to be searched and shared across sites and used in annotation of phenotype data.

- Data exchange technologies. It will be necessary to develop a common data format for exchange of phenotype data which should be linked to information on protocols used to obtain the data. This will allow data to be exchanged between databases and analysis tools to import data from the different databases and carry out analysis over this wider data set.

## Ontologies

Ontologies are widely used to represent genomic and functional genomic information (Bodenreider and Stevens 2006). A confounding factor for phenotype data is the evolution of ontologies that are of different character, yet not orthogonal (*e.g*., Gkoutos et al. 2005; Smith et al. 2005). For example, different types of knowledge representation require different levels of granularity: in some cases summary information is adequate, in others a more detailed approach is required. It is important that studies to evaluate currently available ontologies be carried out in collaboration with major centres for ontological research such as NCBO (National Center for Biomedical Ontology), NCOR (National Center for Ontological Research), ECOR (European Centre for Ontological Research) and others. It will also be important to study ontologies for traits not covered by either of these approaches and it remains to be established whether all the necessary vocabularies and ontologies needed to represent phenotype information are currently available (for example to describe housing and handling conditions, or welfare status, which can affect the results of phenotyping experiments). In the medium to short term the community will need to investigate means of cross-referencing MP and EAV/EQ-based ontological descriptions. In the longer term, these approaches may converge to produce a unitary phenotype ontology. Finally, as the protocol used is a critical factor in determining the results obtained, there is a need to investigate the utility of linking protocols or protocol types into an assay vocabulary (Gkoutos et al. 2005) which provides information on the relatedness of different protocols from the perspective of the phenotypic attributes they measure.

## Protocols and Minimum Information for a Phenotyping Experiment

As well as EMPReSS, MPD, Mutrack and PhenoSITE, a number of other web sites also host protocol collections. Protocols are central to the acquisition and comparison of phenotyping data and can potentially be used both in data acquisition software as a direct means of specifying the information to be reported from any given experiments, and as a means of specifying formally the units of measurement and reasonable ranges of the data collected. During the design of the EMPReSS database (Green et al. 2005) a basic XML schema was developed that allowed the consistent description of SOPs developed during the EUMORPHIA project. We propose to take this as the basis for the development of a more comprehensive XML schema that will allow the representation of all the information needed to describe a phenotyping protocol. A natural offshoot of this process is to consider what is the minimum set of information needed to describe a phenotyping experiment, by analogy with the MIAME criteria developed for microarray data sets (Brazma et al. 2001). As well as the protocol used, it is clear that variables such as mouse strain (or genetic composition), mutation type, gene mutated (where known), housing conditions (possibly including history), feeding regime and handling conditions will need to be recorded. We have set up a working group to develop these general ideas into a more formal framework.

### Data Exchange

An essential requirement for exchange of phenotype data will be a standardised means (such as an XML schema) of describing phenotyping data. Current precursors of such a schema are PhenoXML (http://reaper.lbl.gov/phenote/pheno-xml.rnc) and a schema being developed for transfer of phenotypic screen data to the EuroPhenome database by the EUMODIC consortium (http://www.eumodic.org/). A number of the established data resources containing mouse phenotype data contain different types of data held in different data structures. Our strategy is to facilitate the establishment of portals to make access to these various resources as seamless as possible. With this in mind, we are establishing an experimental web site (http://www.interphenome.org) that will initially provide links to individual sites providing access to mouse phenotype data. We will then start to implement a phased process of improving the integration of these data sources.

Potentially, the IMSR web site presents an accessible route through which to access phenotype data in the way we discuss here. However there are other possibilities - for example RIKEN's search engine MusBanks (http://omicspace.riken.jp/MusBanks/), which overcomes the differences in phenotype description frameworks at different sites by directly searching the web pages of the original phenotype databases. MusBanks also inferentially connects arbitrary phenotypic keywords with the resources via text mining of MEDLINE, so as to suggest potentially undiscovered phenotypes remaining to be measured. Most likely, principles developed during the process outlined in this paper will be usable by any number of sites wishing to access and analyse mouse phenotype data. Access to this information need not be restricted to conventional interfaces - for example it is possible to imagine interfaces similar to the visual interface used by the EMAP digital atlas of the embryonic mouse (http://genex.hgu.mrc.ac.uk/Atlas/intro.html). Another possibility is to present data in the form of an "ideal mouse", which would summarise state of the art knowledge on individual inbred lines derived using data extracted from various databases (Figure 2).

## Conclusions

The aim of linking phenotype to genotype in the laboratory mouse will only be achieved by a worldwide effort of mutagenesis, quantitative trait locus detection and inbred strain profiling. These efforts provide us with converging insights into the role of the genome in trait variation, but the convergence only occurs if data can be combined and compared. It is therefore essential that information on mouse phenotypes is made available in an integrated manner to the mouse community internationally. With the advent of large-scale projects in these areas, it is vital that the mouse informatics community moves towards this goal as quickly as possible. We have initiated this process and aim to continue it with regular meetings to be held over the next few years with the aim of delivering an integrated portal to mouse phenotype data. The only way such an initiative can succeed is by engaging as many members of the mouse community involved in these sorts of experiments as possible. We have established a web and wiki site (http://www.interphenome.org) to act as a central coordinating site for this project and we welcome input from members of the mouse community we do not currently represent.

## Acknowledgments

## Literature

1. Baker EJ, Galloway L, Jackson B, Schmoyer D, Snoddy J. MuTrack: a genome analysis system for large-scale mutagenesis in the mouse. BMC Bioinformatics. 2004; 5:11. [PubMed: 15018655]

2. Blake JA, Eppig JT, Bult CJ, Kadin JA, Richardson JE. The Mouse Genome Database (MGD): updates and enhancements. Nucleic Acids Res. 2006; 34:D562–567. [PubMed: 16381933]

3. Bodenreider O, Stevens R. Bio-ontologies: current trends and future directions. Brief Bioinform. 2006; 7:256–274. [PubMed: 16899495]

4. Bogue MA, Grubb SC. The Mouse Phenome Project. Genetica. 2004; 122:71–74. [PubMed: 15619963]

5. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, et al. Minimum information about a microarray experiment (MIAME) - towards standards for microarray data. Nat Genet. 2001; 29:365–371. [PubMed: 11726920]

6. Brown SDM, Chambon P, Hrabé de Angelis M, EUMORPHIA_Consortium. EMPReSS: standardized phenotype screens for functional annotation of the mouse genome. Nat Genet. 2005; 37:1155. [PubMed: 16254554]

7. Brown SDM, Hancock JM, Gates H. Understanding mammalian genetic systems: the challenge of phenotyping in the mouse. PLoS Genetics. 2006; 2:e118. [PubMed: 16933996]

8. Brunak S, Danchin A, Hattori M, Nakamura H, Shinozaki K, et al. Nucleotide Sequence Database policies. Science. 2002; 298:1333. [PubMed: 12436968]

9. Chesler EJ, Lu L, Wang J, Williams RW, Manly KF. WebQTL: rapid exploratory analysis of gene expression and genetic networks for brain and behavior. Nat Neurosci. 2004; 7:485–486. [PubMed: 15114364]

10. Eppig JT, Bult CJ, Kadin JA, Richardson JE, Blake JA. The Mouse Genome Database (MGD): from genes to mice - a community resource for mouse biology. Nucleic Acids Res. 2005; 33:D471–D475. [PubMed: 15608240]

11. FIMRe Board Of Directors. FIMRe: Federation of International Mouse Resources: Global networking of resource centers. Mamm. Genome. 2006; 17:363–364. [PubMed: 16688526]

12. Gkoutos GV, Green ECJ, Mallon A-M, Hancock JM, Davidson D. Using ontologies to describe mouse phenotypes. Genome Biol. 2005; 6:R8. [PubMed: 15642100]

13. Goldowitz D, Frankel WN, Takahashi JS, Holtz-Vitaterna M, Bult C, et al. Large-scale mutagenesis of the mouse to understand the genetic bases of nervous system structure and function. Brain Res Mol Brain Res. 2004; 132:105–115. [PubMed: 15582151]

14. Green ECJ, Gkoutos GV, Lad HV, Blake A, Weekes J, et al. EMPReSS: European Mouse Phenotyping Resource for Standardised Screens. Bioinformatics. 2005; 21:2930–2931. [PubMed: 15827082]

15. Grubb SC, Churchill GA, Bogue MA. A collaborative database of inbred mouse strain characteristics. Bioinformatics. 2004; 20:2857–2859. [PubMed: 15130929]

16. Smith CL, Goldsmith CA, Eppig JT. The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. Genome Biol. 2005; 6:R7. [PubMed: 15642099]

17. Strivens M, Eppig JT. Visualizing the laboratory mouse: capturing phenotypic information. Genetica. 2004; 122:89–97. [PubMed: 15619965]
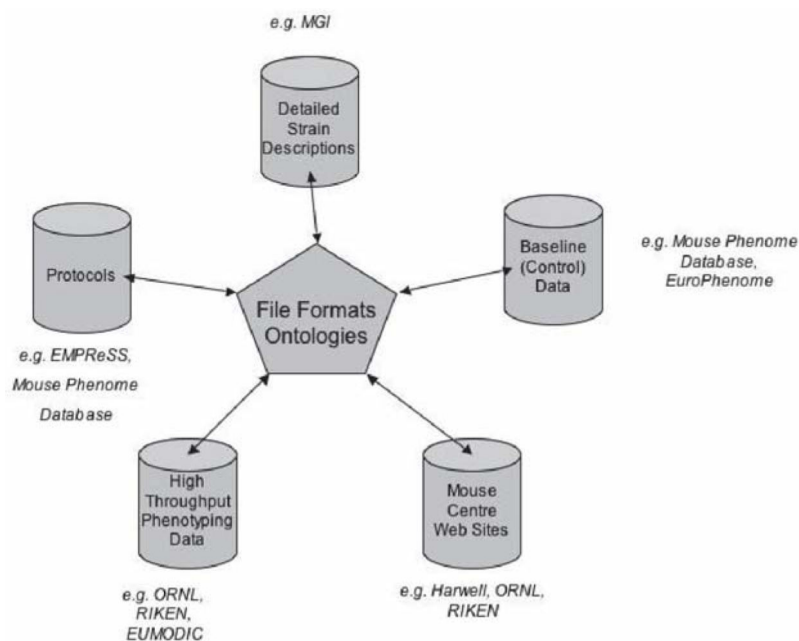
**Figure 1. Phenotype data resources.**

We can distinguish a variety of types of resource holding data on mouse phenotypes. These range from the detailed descriptions held by the Mouse Genome Informatics group through quantitative data on the normal phenotypes of inbred lines held at the Mouse Phenome Database and in Europhenome, through to descriptions of high throughput phenotyping studies such as EUMODIC and summary descriptions of individual mutant lines held at a number of sites. Knowledge of the protocol used to derive data is also essential. To unify this data we will require a common description vocabulary, in the form of ontologies, and common means of representing the data itself and the underlying protocols for transfer between databases and between databases and analysis applications.
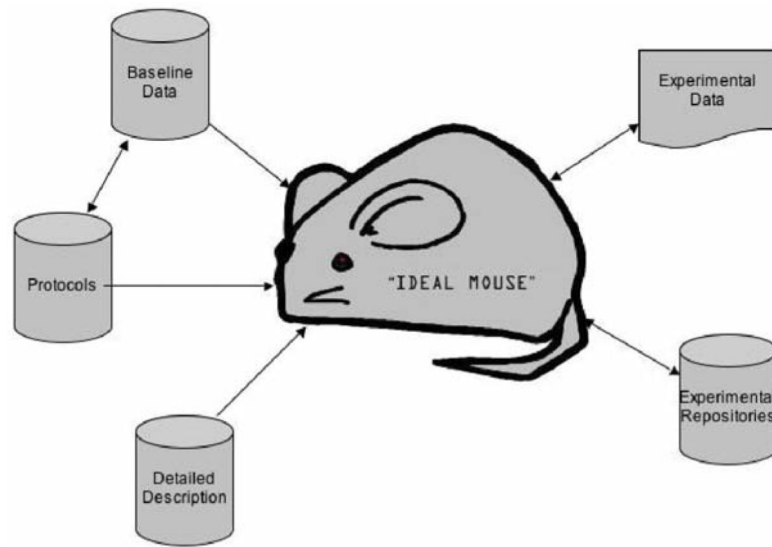
**Figure 2. The ideal mouse.**
For any given inbred line or mouse strain of defined genotype, data might be collected from various online sources to provide a summary of the phenotype of that line or strain in the form of a central tendency and a measure of variation. This would allow meta-analysis of phenotyping analyses carried out at different sites, and the integration of similar data relating to different genetic backgrounds, facilitating mining of genotype/phenotype datasets for novel patterns of association. By collecting data from different sites, information on the reproducibility of particular measurements could also be gathered. The information could be updated continuously to maintain up-to-date information