Toward fully integrated mouse phenotype data

Hiroshi Masuya[1] and Riichiro Mizoguchi[2]

[1] Technology and Development Unit for Knowledge Base of Mouse Phenotype
RIKEN Bioresouce center
Kouyadai 3-1-1, Tsukuba, Ibaraki, 305-0074, Japan
hmasuya@brc.riken.jp
http://www.brc.riken.go.jp/lab/bpmp/
[2] Department of Knowledge Systems
The Institute of Scientific and Industrial Research (ISIR), Osaka University
8-1 Mihogaoka, Ibaraki, Osaka, 567-0047, Japan

ABSTRACT. The multitude of murine genetic resources available worldwide, including inbred mouse lines, spontaneous and induced mutants, and ES lines ready for conditional gene targeting, represent a unique resource for the biomedical community. Recent international efforts are striving to complete the functional annotation the mouse genome, which will allow a systems-based view of mammalian biological networks and causes of disease. For this purpose, integration of the broad range of phenotype data has presented new challenges. The OBO ontologies and Minimum Information Standards have provided important frameworks for the integration and sharing of phenotype information. On the other hand, extension of these tools will need to allow advanced integration of phenotype data, namely, fully machine-processable integration of quantitative and qualitative data, advanced classification of phenotypes, and representation of the relationships between mouse phenotypes and human diseases. We propose that integration based on top-level ontology using YATO is a sophisticated answer to this issue.

1. Introduction

The mouse is an excellent model organism that facilitates the translation of genetics and genomics into clinical research. Mice have long been used as models of various human diseases, and numerous methods are available to experimentally manipulate the murine genome, including conditional gene targeting to inactivate specific genes in specific tissues and/or at specific developmental stages. Furthermore, the large collection of standard inbred strains and controlled animal rearing environments allow phenotypic observations to be confirmed, and environmental and genetic factors to be systematically changed to measure their effects. Following sequencing of the whole mouse genome [Mouse Genome Sequencing Consortium 2002], the International Knockout Mouse Consortium (IKMC) has been launched to generate mutations for every identified gene [Collins et al. 2007]. As a next step, systematic and comprehensive functional characterization of the generated mutants is desired to reveal the genetic foundation of biological processes and disease.

Mouse clinic systems, standardized and comprehensive phenotyping platforms to analyze individual mice, enable direct and highly accurate comparisons among the large number of generated mutants. Combining data from these independent systems and developing an open integrated database containing the mouse phenotype profiles from across academia and industry is likely to be beneficial to a range of biomedical communities studying mammalian and model organism biology [Gailus-Durner et al. 2005; Brown et al. 2005]. To coordinate worldwide efforts to functionally annotate the mouse genome, the International Mouse Phenotyping Consortium (IMPC), a cooperative network of mouse clinics, has been established. The ultimate goal of the IMPC is to develop a comprehensive database of genetic and molecular data that will illuminate biological networks in the mouse, particular those applicable to human biology and disease.

Informatics clearly represents an important component of the IMPC. For mouse genetics, the Mouse Genome Database (MGD) has been crucial as an integrated database containing detailed descriptions of mutant lines largely identified from published studies [Eppig et al. 2005]. The IMPC, however, requires integration based on raw experimental data. The Mouse Phenotype Database Integration Consortium (InterPhenome) has discussed such informatics issues regarding the sharing of phenotype data to integrate current and future mouse resources as much as possible, as well as to promote standardized phenotype descriptions using ontologies and file formats for phenotyping protocols and phenotype data sets [Mouse Phenotype Database Integration Consortium 2007].

In this paper, we review issues related to the integration of international mouse phenotype information and discuss future requirements to facilitate advanced integration.

2. Metadata to be integrated with mouse phenotype data

Raw data-based integration of phenotype information produced from large-scale phenotyping platforms requires standardized descriptions of the data and relevant metadata, such as phenotyping assay procedures. This will support unambiguous interpretation of results and reuse of the data. In this section, we summarize the data and metadata to be integrated.

2.1. Phenotype data (parameters and parameter values)

Phenotyping assays are designed to measure or specify the quantity or quality of biological entities, such as the individual animal, some anatomical component, or a biological process, *i.e.* experimental parameters or biological traits. Two essential processes are involved in the integration of phenotype data: (1) identification of a biological entity or experimental parameter, such as “the tail length of the heterozygous mutant;” and (2) the values of the parameter (or data), which can be quantitative or qualitative (“10 cm” or “long”). In addition, integration of phenotype data from various institutes may require a distinction between the “true value” of the examined

entity and the experimental data, which may be influenced markedly by such institute-specific experimental conditions as assay methods or animal housing procedures.

2.2. Experimental procedures and baseline data

Developing a common format for descriptions of phenotyping procedures is a major issue for the InterPhenome consortium. Procedure information includes a broad range of topics, including the handling of animals, reagents, equipment, and materials to be used. In addition, baseline data for each procedure includes important information about how the data was obtained. Inbred mouse strains can provide reproducible baseline data because individuals of each strain have identical genetic backgrounds [Masuya et al. 2007]. It has been reported that baseline data from multiple inbred strains provide better accuracy when comparing different platforms [Tucci et al. 2006; Wahlsten et al. 2003].

3. Tools and standards for mouse phenotype integration

Integration requires standardized descriptions of various types of information. In this section, we outline tools that play essential roles during the ongoing integration of mouse phenotype data.

3.1. Ontologies

The Open Biomedical Ontologies (OBO) consortium, an open umbrella organization for developers of bioinformatic ontologies, has markedly contributed to the current annotation processes used for various types of metadata associated with mouse phenotypes.

Mammalian Phenotype (MP). To incorporate published literature, MP was developed by the MGD to allow robust annotation of mouse phenotypes and querying capabilities of mouse phenotype data using a search system based on free-text descriptions [Smith CL. et al. 2005]. For annotating raw data, MP is also beneficial when determining whether a detected phenotype corresponds to other common phenotypes used in mammalian genetics studies.

Mouse Pathology (MPATH). MPATH ontology covers all currently known classes of lesion, specifically in mice. Inclusion of definitions and synonyms helps to clarify the often disparate set of terms used by pathologists with different training backgrounds to describe the same lesion type. MPATH incorporates the NIH Mouse Models from the Human Cancer Consortium recommendations on hematopoietic neoplasms [Schofield et al. 2004].

Mouse Adult Gross Anatomy (MA), Mouse Gross Anatomy and Development (EMAP), Cell Type (CL), Gene Ontology (GO), Chemical Entity of Biological Interest (ChEBI), and Biological Pathway Exchange (BioPAX). MA, EMAP. CL, GO, ChEBI, and BioPAX represent biological entities affected by any phenotypic change in anatomic components, embryologic anatomic components, cells, cellular components, chemical compounds, biological processes, molecular interactions, and signaling pathways [Baldock et al. 2003; Hayamizu et al. 2005; Bard et al. 2005; Gene Ontology Consortium 2000; Degtyarenko et al. 2008; Jiang et al. 2005].

Phenotypic Quality (PATO) and Unit Ontology (UO). PATO is an ontology that provides practical qualitative values for phenotype description. It classifies various values using a basic framework, such as "qualitative value is_a parameter." Typically, it is used for "entity plus quality" (E+Q) annotation of experimental parameters and parameter values [Gkoutos et al. 2004; Gkoutos et al. 2005]. UO represents units classified for the integration of quantitative values.

Ontology of Scientific Experiments (EXPO) and Experiment ACTions (EXACT). EXPO and EXACT are ontologies that serve as the basis of a method for representing biological laboratory protocols to enable the publication of protocols with increased clarity. EXACT includes several different and important top-level concepts, such as process, objects, proposition, and quality. These concepts function to describe the experimental activities [Soldatova and King 2006; Soldatova et al. 2008].

3.2. Library of cross-talk among ontologies

OBO Foundry, a coordinated reforming activity to promote the integration of OBO ontologies, has initiated efforts to produce "cross-product;" this process identifies logical definitions and cross-talk for terms in existing OBO ontologies, spurring the development of the OBO Relation Ontology (RO) [Smith B. et al. 2005; Smith B. et al. 2007]. PATO developers have begun to provide "post-coordinated" libraries to show definitions of pre-coordinated phenotype terms, such as MP, based on basic qualities defined in PATO and biological entities using E+Q annotation and cross-product strategy (http://bioontology.org/wiki/index.php/PATO:Pre_vs_Post_Coordinating). This approach seeks to elucidate relationships between mouse phenotypes and diseases (Gkoutos et al. personal communication).

3.3. Minimum Information to describe a Mouse Phenotype Procedure (MIMPP)

The InterPhenome consortium has identified three major priorities as requirements for standardized descriptions of phenotyping procedures, data exchange technology, and phenotype ontologies [Mouse Phenotype Database Integration Consortium 2007]. These requirements gave rise to the minimal information standard, which integrated

such data formats as XML schemas to allow the data to be reused and analyzed as well as interchanged between public repositories. The InterPhenome consortium is now discussing a draft version of MIMPP to be standardized (http://mibbi.sourceforge.net/projects/MIMPP/). MIMPP is functioning cooperatively with the Minimum Information for Biological and Biomedical Investigations (MIBBI) consortium to facilitate broad coordination in the biomedical community [Taylor et al. 2008].

4. Current issues in advanced semantic integration and our proposals

OBO ontology and Minimum Information Standards have obviously contributed to the international efforts to integrate mouse phenotype information. The necessary steps, including development of common vocabulary and data structure, have been promoted by the InterPhenome and IMPC consortiums. Of note, however, some barriers may hinder advanced semantic integration and the IMPC's ultimate goal—a comprehensive database of the outcomes of *in vivo* molecular interventions, including genetic lesions and small molecule interactions. The following proposals are derived from ontological studies based on the Yet Another Top-level Ontology (YATO); the latest top-level ontology includes multiple improvements from the existing top-level ontologies, specifically in the areas of Quality description, Representation, and Process/Event [Mizoguchi 2004] [Mizoguchi 2009].

4.1. The need for advanced data modeling of anatomic components

The current version of MA ontology provides mainly "part_of" links to represent the spatial location within tissues, but does not include an "is_a" link to provide a logical definition. This problem will be solved in a future version [Hayamizu et al., personal communication]. The following extensions will be required to disclose relationships between mouse phenotypes and human diseases: (1) mapping of homologous mouse and human organs; (2) association of EMAP terms with detailed developmental and physiologic events; and (3) identification of shared properties between organs or tissues, such as morphological features and functional characteristics.

4.2. The needs for an ontological framework to represent quality and quantity: Advanced description of Phenotypic Quality

Quality description is a core principle for the integration of phenotype information. In the present situation, PATO plays an essential role to provide a practical basis for the integration of phenotype information across species. The current version of PATO is

arranged as a single-hierarchy model of "Quality," which was modified from a previous two-hierarchy model composed of "Attribute" and "Value"; this represents the equitable quality type and quality value mentioned above. This modification was made after lengthy discussion, with the current model avoiding the redundancy of the EAV (Entity + Attribute + Value) annotation with no apparent information loss compared with old versions (http://bioontology.org/wiki/index.php/PATO:About). We, however, propose another YATO-based phenotype description model with a number of advanced features.

Our model is arranged as two hierarchies, *Quality type* and *Quality value*, similar to the top level of the older version of PATO (Fig. 1); those subclasses, however, are arranged using a different philosophy developed from careful ontological examination. Additionally, our two hierarchies may not represent a revival of the redundancy of EAV annotation. Of note, a change of quality value cannot be described by a single-hierarchy model in which a specific quality (*e.g.* the length of the tail of the mouse X) is not distinguished from its value (say, 5 cm), because a change of quality means an alteration in the quality (say, 5 cm to 6 cm) for a quality value of the "identical" (dependent) entity from one to another (the length of the tail of the mouse X). On the other hand, in the two-hierarchy model of YATO, we can describe the change of quality, because it differentiates between *Quality type*, as the focused qualitative entity, and *Quality value*. This feature enables us to describe changes of qualities along the courses of time, such as growth, development, and the experimental time course.

In addition, a number of detailed descriptions can be used with the YATO model. (1) Phenotype evaluations with nominal, ordinal, and rational scales (each of these is a subclass of *Quality value*) are fully integrated using the two-hierarchy model. (2) For modeling relatively complicated systems, such as fertility [PATO: 0000274], female fertility [PATO: 0000277], and male fertility [PATO: 0000279], single-hierarchy model is difficult to understand, in which "Attribute" and "Value" terms are intricately arranged. On the other hand, two hierarchies model of YATO can represent more understandable schema, where each *Quality type* and *Quality value* hierarchy and the relationship between them by the "value of" links, (3) "*Property*", a specified *Quality type* obtained from the abstract *Quality value* enables the systematic representation of a more detailed quality description derived from a specific *Quality value*, i.e. "severely shortened tail" as an extension of "shortened tail" ("refer to" link in Fig. 1) with "severe" ("value" link in Fig. 1). The *Property* model in YATO represents practical integration of different *Quality values*, namely, short/long and mild/severe, into the quality description. Moreover, annotation of the phenotype using *Property*, which is subclass of *Quality type*, seems to be a proper ontological representation, because the essence of the pheno-"type" may be a Quality "type" to represent a classification (or a type) of observable characteristics, corresponding to the genotype.

4.3. Ontological framework for more broad-ranging concepts

MIMPP provides a guideline for phenotype experimental data and description of procedures, including the types of metadata required for data exchange as observed with

the XML schema. On the other hand, EXACT and EXPO appear to successfully allow ontology-based formalization of knowledge about scientific experimental design, methodology, and results representation. These approaches serve as the middle-level tier to bridge top-level ontologies, such as Basic Formal Ontology (BFO) [Grenon and Smith 2004] and Suggested Upper Merged Ontology (SUMO) [Niles and Pease 2001], with domain knowledge [Soldatova and King 2006; Soldatova et al. 2008]. Full integration of phenotype-related information ranging from the molecular level to biologic functions requires additional universal data models for a wide range of items, however. In particular, practical ontological modeling of concepts in genetics may be of primary importance; examples include "genomic sequence," a string of symbols representing nucleotides; "gene and allele," a representative and variant form of a genome fragment as a carrier of the minimum unit of genetic information that encodes functional RNA; "locus," a location in the genome identified using various experimental methods; "genotype," the genetic composition of a specific gene; and "genetic background," a summation of the genetic composition of an individual. These concepts will bridge the gap between phenotypic descriptions, molecules, and biological processes to augment the model of molecular genetics.

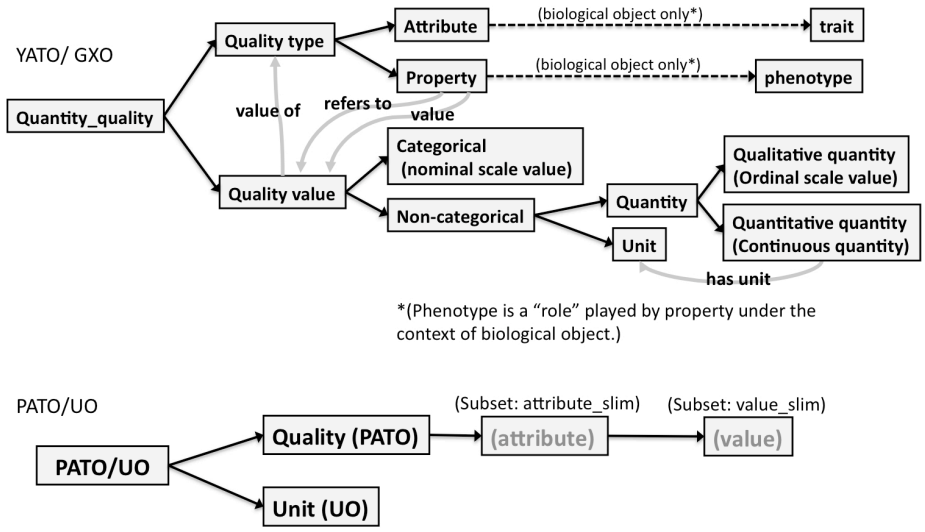

Figure 1. Comparison of the quality-related concepts in YATO/GXO and PATO/UO.

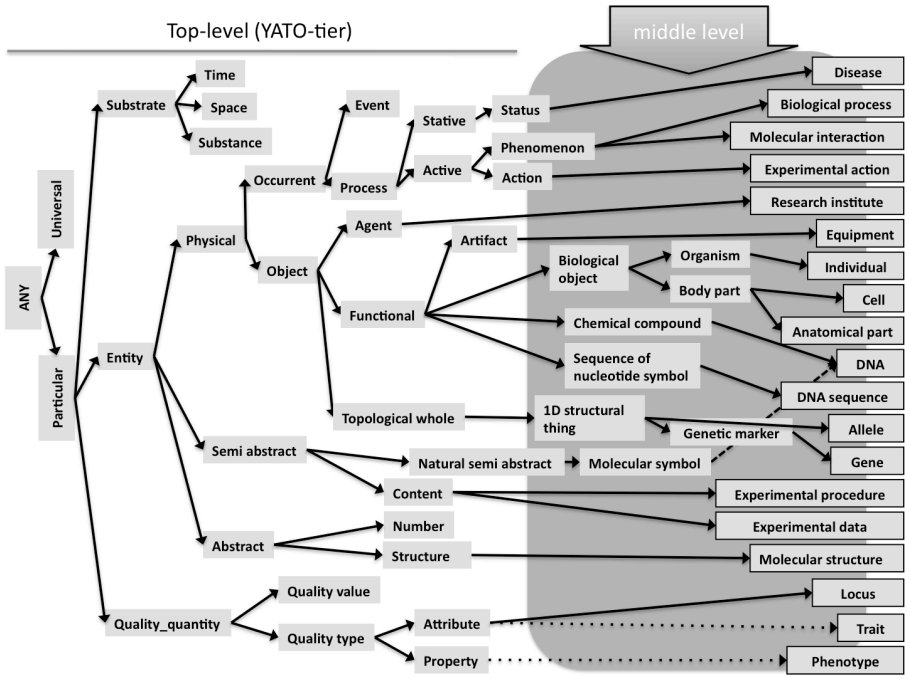

Figure 2. Simplified schema of GXO (in progress).

5. Toward the development of a top-level ontology-based integration of experimental genetics

Top-level ontology-based integration is a methodology that attempts to depict the world and integrate a broad range of concepts using a general data model. This methodology promotes the sharing of information within and between subject areas, reducing both duplication and the loss of knowledge [Mizoguchi 2003; Soldatova and King 2006]. To establish top-level ontology-based integration of mouse phenotype information, we are now developing Genetics Ontology (GXO) as a middle-level ontology to bridge YATO with the biological domains (Fig. 2). GXO will incorporate trait and phenotype descriptions, as mentioned above, as well as representation of experimental design, results, and genetics, including proper and explicit definitions of genetic concepts and modeling in which a gene is a design plan for one or multiple gene products written using a molecular symbol; this process is analogous with the model of artificial representation in YATO. Development of this semantic framework will facilitate the development of domain ontologies and integrated databases in the biomedical field to provide one of the guidelines for the formalization of common data structure.

References

[Baldock et al. 2003] Baldock RA, Bard JB, Burger A, Burton N, Christiansen J, Feng G, Hill B, Houghton D, Kaufman M, Rao J, Sharpe J, Ross A, Stevenson P, Venkataraman S, Waterhouse A, Yang Y, Davidson DR., EMAP and EMAGE: a framework for understanding spatially organized data., *Neuroinformatics*. 1(2003), pp.309-325

[Bard et al. 2005] Bard J, Rhee SY, Ashburner M., An ontology for cell types, *Genome Biol*. 6(2005), pp. R21.

[Brown et al. 2005] Brown S.D.M., P. Chambon and M. Hrabe de Angelis; Eumorphia Consortium, EMPReSS: standardized phenotype screens for functional annotation of the mouse genome, *Nat. Genet*. 37(2005), pp.1155.

[Degtyarenko et al. 2008] Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcántara, R., Darsow, M., Guedj, M. and Ashburner, M., ChEBI: a database and ontology for chemical entities of biological interest., *Nucleic Acids Res*. 36(2008), pp. D344–D350.

[Collins et al. 2007] Collins, F.S., Rossant, J. and Wurst, W., A mouse for all reasons, *Cell* 128(2007), pp. 9-13.

[Eppig et al. 2005] Eppig JT, Bult CJ, Kadin JA, Richardson JE and Blake JA, The Mouse Genome Database (MGD): from genes to mice - a community resource for mouse biology. *Nucleic Acids Res* 33(2005), pp. D471-D475

[Gailus-Durner et al. 2005] Gailus-Durner V., Fuchs, H., Becker, L., Bolle, I., Bielmeier, M., Calzada-Wack, J., Elvert, R., Erhardt, N., Dalke, C., Franz, T.J., Grundner-Culemann, E., Hammelbacher, S., Hölter, S., Hölzlwimmer, G., Horsch, M., Javaheri, A., Kalaydjiev, S., Klempt, M., Kling, E., Kunder, S., Lengger, C., Lisse, T., Mijalski, T., Naton, B., Pedersen, V., Prehn, C., Przemeck, G., Racz, I., Reinhard, C., Reitmeir, P., Schneider, I., Schrewe, A., Steinkamp, R., Zybill, C., Adamski, J., Beckers, J., Behrendt, H., Favor, J., Graw, J., Heldmaier, G., Höfler, H., Ivandic, B., Katus, H., Kirchhof. P., Klingenspor, M., Klopstock, T., Lengeling, A., Müller, W., Ohl, F., Ollert, M., Quintanilla-Martinez, L., Schmidt, J., Schulz, H., Wolf, E., Wurst, W., Zimmer, A., Busch, D.H., and Hrabé de Angelis, M., Introducing the German Mouse Clinic: Open access platform for standardized phenotyping, *Nat Methods* 2(2005), pp. 403-404

[Gene Ontology Consortium 2000] The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nature Genet*. 25(2000), pp. 25-29.

[Grenon and Smith 2004] Grenon,P. and Smith,B., *SNAP and SPAN: towards dynamic spatial ontology*., Spat. Cogn. Comput., 4(2004), pp. 69–103.

[Gkoutos et al. 2004] Gkoutos GV, Green EC, Mallon AM, Blake A, Greenaway S, Hancock JM, Davidson D., Ontologies for the description of mouse phenotypes., *Comp Funct Genomics*. 5(2004), pp. 545-551.

[Gkoutos et al. 2005] Gkoutos GV, Green ECJ, Mallon A-M, Hancock JM and Davidson D, Using ontologies to describe mouse phénotypes, *Genome Biol* 6(2005), pp. R8

[Hayamizu et al. 2005] Hayamizu TF, Mangan M, Corradi JP, Kadin JA, Ringwald M., The Adult Mouse Anatomical Dictionary: a tool for annotating and integrating data, *Genome Biol*. 6(2005), pp.R29.

[Jiang et al. 2005] Jiang K, Nash C., Ontology-based aggregation of biological pathway datasets, *Conf Proc IEEE Eng Med Biol Soc*. 7(2005), pp. 7742-7745.

[Masuya et al. 2007] Masuya H, Yoshikawa S, Heida N, Toyoda T, Wakana S, Shiroishi T., Phenosite: a web database integrating the mouse phenotyping platform and the experimental procedures in mice, *J Bioinform Comput Biol*. 5(2007), pp. 1173-1191.

[Mizoguchi 2003] Mizoguchi, R., Tutorial on ontological engineering - Part 1: Introduction to Ontological Engineering , *New Generation Computing*, 21(2003, pp.365-384.

[Mizoguchi 2004] Mizoguchi, R., Tutorial on ontological engineering - Part 3: Advanced course of ontological engineering, *New Generation Computing*, 22(2004), pp. 193-220.

[Mizoguchi 2009] Mizoguchi R., Yet Another Top-lovel Ontology : YATO, this meeting (2009)

[Mouse Genome Sequencing Consortium 2002] Mouse Genome Sequencing Consortium, Initial sequencing and comparative analysis of the mouse génome, *Nature*. 420(2002), pp. 520-562.

[Mouse Phenotype Database Integration Consortium 2007] The Mouse Phenotype Database Integration Consortium, Integration of mouse phenome data resources, *Mammalian Genome* 18(2007), pp. 157-163.

[Niles and Pease 2001] Niles, I., and Pease, A., Towards a Standard Upper Ontology. In *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*, Ogunquit, Maine, 2001, pp. 17-19,

[Schofield et al. 2004] Schofield PN, Bard JB, Booth C, Boniver J, Covelli V, Delvenne P, Ellender M, Engstrom W, Goessner W, Gruenberger M, Hoefler H, Hopewell J, Mancuso M, Mothersill C, Potten CS, Quintanilla-Fend L, Rozell B, Sariola H, Sundberg JP, Ward A., Pathbase: a database of mutant mouse pathology, *Nucleic Acids Res*. 32(2004), pp. D512-D515.

[Smith CL. et al. 2005] Smith CL, Goldsmith CA and Eppig JT, The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information, *Genome Biol* 6(2005), pp. R7

[Smith B. et al. 2005] Smith B, Ceusters W, Klagges B, Kohler J, Kumar A, Lomax J, Mungall CJ, Neuhaus F, Rector A, Rosse C Relations in Biomedical Ontologies, *Genome Biology*, 6(2005), pp. R46

[Smith B. et al. 2007] Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ; OBI Consortium, Leontis N, Rocca-Serra P, Ruttenberg A, Sansone SA, Scheuermann RH, Shah N, Whetzel PL, Lewis S., The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration, *Nat Biotechnol*. 25(2007), pp. 1251-1255.

[Soldatova and King 2006] Soldatova LN, King RD., An ontology of scientific experiments., *J R Soc Interface*. 22(2006), pp. 795-803.

[Soldatova et al. 2008] Soldatova LN, Aubrey W, King RD, Clare A., The EXACT description of biomedical protocols., *Bioinformatics*. 24(2008), pp. i295-303.

[Taylor et al. 2008] Taylor CF, Field D, Sansone SA, Aerts J, Apweiler R, Ashburner M, Ball CA, Binz PA, Bogue M, Booth T, Brazma A, Brinkman RR, Michael Clark A, Deutsch EW, Fiehn O, Fostel J, Ghazal P, Gibson F, Gray T, Grimes G, Hancock JM, Hardy NW, Hermjakob H, Julian RK Jr, Kane M, Kettner C, Kinsinger C, Kolker E, Kuiper M, Le Novère N, Leebens-Mack J, Lewis SE, Lord P, Mallon AM, Marthandan N, Masuya H, McNally R, Mehrle A, Morrison N, Orchard S, Quackenbush J, Reecy JM, Robertson DG, Rocca-Serra P, Rodriguez H, Rosenfelder H, Santoyo-Lopez J, Scheuermann RH, Schober D, Smith B, Snape J, Stoeckert CJ Jr, Tipton K, Sterk P, Untergasser A, Vandesompele J, Wiemann S., Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project, *Nat Biotechnol*. 26(2008), pp. 889-896.

[Tucci et al. 2006] Tucci V, Lad HV, Parker A, Polley S, Brown SD and Nolan PM., Gene-environment interactions differentially affect mouse strain behavioral parameters, *Mamm Genome* 17(2006), pp. 1113-1120

[Wahlsten et al. 2003] Wahlsten D, Metten P, Phillips TJ, Boehm SL 2nd, Burkhart-Kasch S, Dorow J, Doerksen S, Downing C, Fogarty J, Rodd-Henricks K, Hen R, McKinnon CS, Merrill CM, Nolte C, Schalomon M, Schlumbohm JP, Sibert JR, Wenger CD, Dudek BC, Crabbe JC, Different data from different labs: lessons from studies of gene-environment interaction, *J Neurobiol*. 54(2003), pp. 283-311.