Toward fully integration of mouse phenotype information

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. Tremendous worldwide genetic resources of mice represent a unique resource for the biomedical community, including inbred mice, spontaneous and induced mutants, and a large resource of conditional gene targeting ready ES-cell lines. Recently, international efforts are ongoing toward the completion of functional annotation the mouse whole genome, which furnishing systems view of mammalian biological networks and causes of disease. In this situation, integrations of broad-ranged phenotype data have been presented new challenges for biomedical field. The OBO ontologies and Minimun Information Standards have provided important frameworks of integration and sharing of phenotype information. On the other hand, extension of these tools will be need to utilize advanced integration of phenotype data, namely to represent fully machine-process-able integration of quantitative and qualitative data, advanced classification of phenotype and representation of relationship between mouse phenotype and human disease. We propose that top-level ontology based integration using Mizoguchi's ontology is one of the most sophisticated answer for this issue.

1 Introduction

The mouse is one of the excellent model organisms to hold the translational advantages in genetics and genomics into wider developments in clinical research. As a disease model animal, it has long history of study, and there exist various technologies for the experimental manipulation of its genome such as conditional gene targeting to inactivate specific genes at the specific tissues and developmental stages. Furthermore, the large collection of standardized classical inbred strains and controlled animal rearing environments provides the ability to confirm phenotypic observations as well as systematically change environmental factors and genetic input to measure effects under defined conditions. Following the determination of the whole genome sequence of mice [Mouse Genome Sequencing Consortium 2002], the International Knockout Mouse Consortium (IKMC) has been launched to generate mutations for every gene in the mouse genome [Collins et al. 2007]. As the next step to reveal the

genetic foundation of biological processes and disease, systematic and comprehensive functional characterization of generated mutants is desired.

The mouse clinic system, a standardized and comprehensive phenotyping platform to analyze mouse individuals enables direct and higher accuracy comparison among the large number of mutants generated. It may be a great contribution for biomedical communities, if these independent systems combine their data and develop open integrated data base of the profile of mouse phenotyping with diverse groups, across academia and industry, in mammalian and model organism biology [Gailus-Durner et al. 2005], [Brown et al. 2005]. To coordinate a worldwide effort to functionally annotate the mouse genome, the International Mouse Phenotyping Consortium (IMPC), a worldwide cooperative network of mouse clinics, is being established. The ultimate goal of IMPC is to develop a comprehensive database of the outcomes of molecular interventions that will reveal a framework for biological networks in the mouse on which human biology and disease networks.

Informatics represents clearly an important part in IMPC activity. In mouse genetics field, Mouse Genome Database (MGD) has been crucial role to give integrative database which contains detailed descriptions of mutant lines largely identified from published literature [Eppig et al. 2005]. However, IMPC activity requires experimental raw-data based integration. Mouse Phenotype Database Integration Consortium (InterPhenome) had been discussed about such informatics issues on phenotype data sharing to aim integrating as far as possible the current and future mouse phenotype resources, and promoting 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 [Mouse Phenotype Database Integration Consortium 2007].

In this paper, we review international issues to integrate mouse phenotype information and discuss future requirements to materialize advanced integration.

2 Meta data to be integrated with mouse phenotype data

To enable raw-data based integration of phenotype information produced from large-scale phenotyping platforms requires standardized descriptions of various kinds of phenotype data and relevant metadata to support the unambiguous interpretation and reuse of the data such as procedure of phenotyping assays. 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 individual animal, anatomical part or biological process, that are termed as experimental parameter or traits. In the integration of phenotype data, it is indispensable to share the identities what parameters to be measured. On the other hand, values of these parameters must be distinguished from parameters

themselves as experimental results that are roughly classified into quantitative data and qualitative data. In general discussion in InterPhenome consortium, it is complied that quantitative values must be converted into qualitative values in the process to integrate phenotype data produced from different phenotyping platforms.

2.2 Experimental conditions

Experimental results are generally strongly influenced by its condition such as detailed procedures of phenotying assay and animal housing. These procedures include broad ranged information, namely, handling of animal, regents, equipments and consumables to be used. In addition, a baseline data of each procedure represents quite important properties of background of data production. In mice, inbred strains provide repeatable baseline data because individuals of each strain have identical genetic backgrounds [Masuya et al. 2007]. It is reported that baseline data using multiple inbred strain provide better accuracy for comparison of different platforms [Tucci et al. 2006], [Wahlsten et al. 2003].

3 Tools and standards for integration of phenotype integration in mouse

Integration needs standardized description of broad kind of information. In this section, we outline tools playing essential roles for the ongoing processes of the data integration in mouse phenotype.

3.1 Ontologies

The Open Biomedical Ontologies (OBO) consortium, an open umbrella body for developers of ontologies in bioinformatics field obviously contributed for current annotation processes of various kinds of metadata in mouse phenotypes.

Mammalian Phenotype (MP). In the field of science literature based annotation of mouse phenotype, MGD succeeded to solve irrelevancy of searching system based on free-text description and enabled robust phenotypic annotations and querying capabilities for mouse phenotype data by the development of MP [Smith CL. et al. 2005]. For the annotation of raw-data, MP is also beneficial to indentify what detected phenotype is corresponds to common types used in mammalian genetics study.

Mouse Pathology (MPATH). MPATH ontology covers all currently known classes of lesion, with specific reference to the mouse. The inclusion of definitions and synonyms helps to clarify the often disparate set of terms used by pathologists trained in different

traditions which actually describe the same lesion. It incorporates the NIH Mouse Models of Human Cancer Consortium recommendations on haematopoietic 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 in any phenotypic change is occurred such as anatomical parts, embryological anatomy parts, cells, cellular components, chemical compound, biological processes, molecular interaction and 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 a ontology for the practical qualitative value of phenotype description. It classified various values with a basic framework 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 classification of units for the integration of quantitative value.

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

3.2 Library of cross-talks among ontologies

OBO Foundry, a coordinated reforming activity to promote integration of OBO ontologies has initiated to produce "cross-product" to represent logical definitions and cross-stalks of 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 has began to provide "post-coordinated" libraries to show definitions of pre-coordinated phenotype terms such as MP in terms of basic qualities defined in PATO and bearer biological entities by the methodology of E+Q annotation and cross-product strategy (http://bioontology.org/wiki/index.php/PATO:Pre_vs_Post_Coordinating). This approach represents relationship between mouse phenotype and disease (Gokoutos et al. personal communication).

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

InterPhenome consortium have identified three major priorities as requirements for standards for describing phenotyping procedures, data exchange technologies and phenotype ontologies [Mouse Phenotype Database Integration Consortium 2007]. These requirements gave rise to develop the minimal information standard which associated data formats such as XML schemas to allow the data to be reused and analyzed and interchange of data between public repositories. InterPhenome consortium is now under discussing draft version of MIMPP to be standardized (http://www.interphenome.org/ppxml/ppml_v1_3.html). The activity of MIMPP is cooperated with Minimum Information for Biological and Biomedical Investigations (MIBBI) consortium for the broad coordination in biomedical community [Taylor et al 2008].

4 Current issues toward advanced semantic integration

OBO ontology and Minimun Information Standards obviously contribute to the international efforts to integrate mouse phenotype information promoted by InterPhenome and IMPC consortiums to providing common vocabulary and data structure. However, we point out that some issues might be barrier to enable advanced semantic integration toward IMPC's ultimate goal: to develop a comprehensive database of the outcomes of molecular interventions in vivo – from genetic lesions to small molecules. Such a dataset will allow us to formulate a framework for biological networks in the mouse on which human biology and disease networks can be revealed.

4.1 Needs for advanced data model for anatomical parts

Current version of MA ontology provides mainly "part_of" links to represent spatial location of each tissue, but don't "is_a" link to represent logical definition. This problem will be solved in future version (Hayamizu et al, personal communication). Following extensions will be required to disclose relationships mouse phenotype and human disease. (1) Mapping of homologous organ between mouse and human. (2) Association of EMAP terms with detailed developmental and physiological events. (3) Identification of shared properties between organs or tissues such as morphological features and functional characteristics.

4.2 Needs for ontological framework to represent quality and quantity

The advanced integration of phenotype data requires integrative representation of both of qualitative and quantitative values to enables automatic conversion from quantitative value to qualitative value with reference to threshold value. It is also required accurate distinction of parameters to be measured with parameter values as experimental results, and to describe changes of value for a specific attribute of same biological entity during the time course, aging or development. Furthermore, phenotypic evaluations are often represented by the combination of parameter value and its degree, i.e. “severe short tail” for the attribute of “morphology”. Current ontological structure of PATO that do not distinguish parameter with value is not sufficient to ensure machine interpretable semantics. Our suggestion is to remodel the structure of PATO to provide classification of quality-related concepts as described in chapter 5.

4.3 Data framework of more broad-ranged concepts

MIMPP will provide data framework of phenotyping experimental data and description of procedures on XML schema. However, fully integration of phenotype-related information ranging molecular level to biological function level requires further universal data model for broad-ranged items. Indeed the post-coordination libraries of OBO ontologies will help to ensure semantic links between deferent data, more formalized method may be required. Web ontology language (OWL) or OWL compatible language may be one of the candidates for standardized language to develop universal data model to represent semantics [Mizoguchi 2004A].

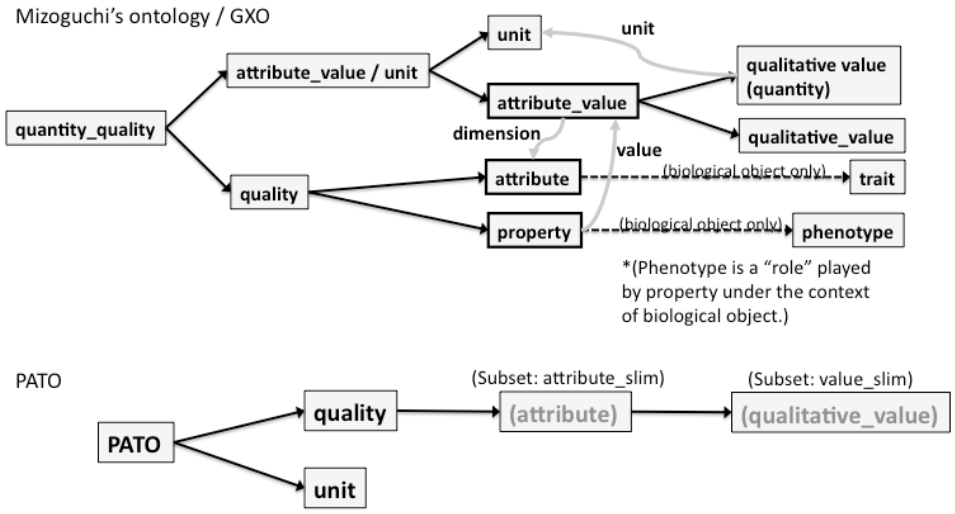

Figure 1. Comparison of quality-related concept in Mizoguchi’s ontology /GXO and PATO.

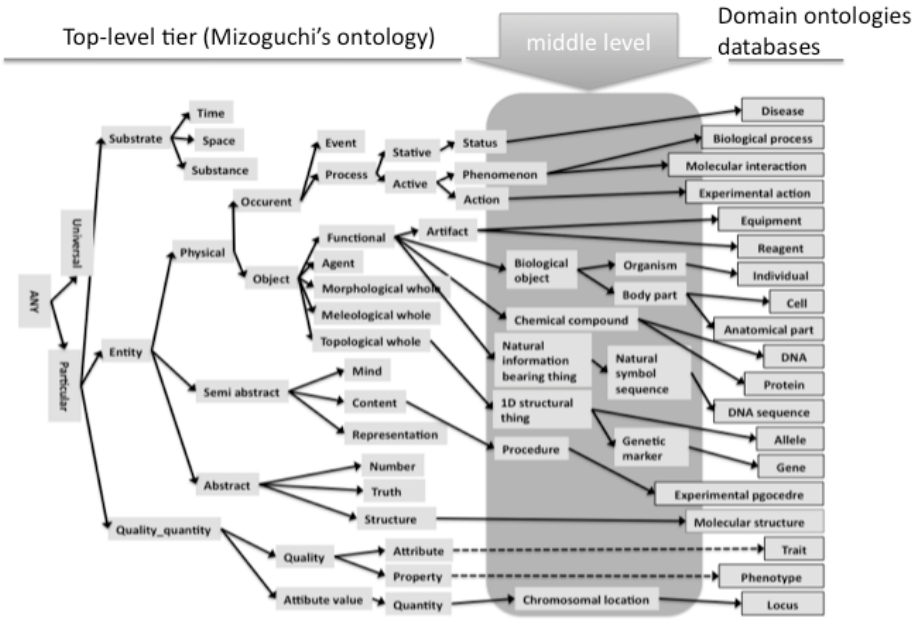

Figure 2. Simplified schema of GXO (development in progress).

5 A proposal of the Top-level ontology based integration of experimental genetics world

The top-level ontology based integration is one of the methodologies to ideally express the world in a general data model to deal with broad-ranged concepts. This methodology would promote the sharing of results within and between subjects, reducing both the duplication and loss of knowledge [Mizoguchi 2003], [Soldatova and King 2006]. It is also an essential step in formalizing knowledge of science, and so fully exploiting computer reasoning in science. In the example of EXACT and EXPO, it seems successfully to reveal data model formalizing knowledge about scientific experimental design, methodology, and results representation. They served as the middle-level tier to bridge top-level ontology 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]. For the fully integration of mouse phenotype information, it seems essential to develop a middle-level ontology to represent generic logic of experimental genetics world.

Now we propose a top-level ontology based integration using Mizoguchi's ontology that is classified into a common Aristotelian and descriptive ontology like as BFO and

Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) [Gangemi et al. 2002]. Mizoguchi's top-level ontology has the number of features based on sophisticated ontological considerations. (1) It is based on detailed investigation about quality-related concepts such as "attribute", a dimension or parameter for a quality-related concept, "attribute-value", a magnitude or multitude for an attribute (this is identical to "quale" of DOLCE), and "property", an abstracted concept of an attribute value representing the degree of quality. This feature is essential for integration of experimental procedure, experimental result and intrinsic quality of biological entities, that includes definitive representation of the parameters to be measured and parameter values to describe phenotypes (Fig. 1). (2) The advanced role theory embedded in this ontology. For the integration of broad-ranged information, it is often needed to model the interaction between general concepts with context dependent concepts. For example, "trait" in genetics means an attribute under the context of description for biological object (shape of a seed, coat color or weight of individual mouse). Similarly, a phenotype is a property under the biological object. Mizoguchi's role theory clearly explains this kind of context dependency as: an entity that is played by another entity in a context [Mizoguchi 2004B], [Mizoguchi et al. 2007]. (3) Advanced theory for event and process is useful to describe developmental events essential for the description of biological information. (4) Consideration for representation is able to divert to the detailed description of the flow of genetic information coded by one-dimensional sequence of molecular entities.

We are now developing genetics ontology (GXO) as a middle-level ontology based on Mizoguchi's ontology to represent genetics which is one of fundamental set of logics to explain heredity and the variation of inherited characteristics in biological study (Fig. 2). This semantic framework will be applicable as one of the prescriptions for developments of domain ontology and integrated databases in bio-medical field.

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

[Gangemi et al. 2002] Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., Schneider, L., Sweetening Ontologies with DOLCE, *Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web, 13th International Conference*, (2002), pp. 166-181

[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 2004A] Mizoguchi, R., Tutorial on ontological engineering - Part 2: Ontology development, tools and languages, *New Generation Computing*, 22(2004), pp.61-96.

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

[Mizoguchi et al. 2007] Mizoguchi R., Sunagawa E., Kozaki K. and Kitamura Y., A Model of Roles within an Ontology Development Tool: Hozo, *J. of Applied Ontology*, 2(2007), pp.159-179.

[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.