ICBO: International Conference on Biomedical Ontology
July 28-30, 2011 · Buffalo, NY, USA

An Advanced Strategy for Integration of Biological Measurement Data

Hiroshi Masuya[1], Georgios V Gkoutos[2], Nobuhiko Tanaka[1], Kazunori Waki[1], Yoshihiro Okuda[3], Tatsuya Kushida[3], Norio Kobayashi[4], Koji Doi[4], Kouji Kozaki[5], Robert Hoehndorf[2], Shigeharu Wakana[1], Tetsuro Toyoda[4], Riichiro Mizoguchi[5]

[1]RIKEN BioResource Center, Tsukuba, Japan; [2]Department of Genetics, University of Cambridge, UK; [3]NalaPro Technologies, Inc, Tokyo, Japan; [4]RIKEN BASE, Yokohama Japan; [5]Department of Knowledge Systems, ISIR, Osaka University, Ibaraki, Japan

Abstract. Aiming at integration of measurement data across various biological experiments, we investigated a methodology for expanding the Phenotypic Quality Ontology (PATO), commonly used for descriptions of biological phenotypes, based on the YAMATO top-level ontology. The mapping of ontology terms from PATO to the YAMATO framework brings several benefits, including: introduction of a classification of quality values to represent measurement scales; distinction of different contexts in which comparisons of ordinal values are made; and establishment of interoperability of quality description formalisms based on different top-level ontologies. In this study, we propose an ontological basis for integrating cross-species and cross-experimental biological measurement data.

Keywords: Interoperability, top-level ontology, quality, phenotype

1 Introduction

A phenotype is an observable and measurable quality of a biological entity. Phenotypes represent a broad range of variations in measured qualities along dimensions such as morphology, development, biochemical or physiological properties, behavior, and so on. For a better understanding of living organisms at the systemic level, it is essential to integrate phenotypic information at all levels and along all such dimensions. This requires the development of a sophisticated informatics infrastructure for the description, exchange and integration of phenotypic data. The ontological formalization of the description of phenotypic qualities is a core issue in the development of such an infrastructure.

To realize a high degree of freedom of data representation, phenotypes are often described by means of representations that employ ontology terms [1]. The Descriptive Ontology for Linguistic, Cognitive Engineering (DOLCE) [2], in contrast, uses the <Entity, Attribute, Quality value> (EAV) formalism, as in <John, height, 180 cm>). The Generalized Architecture for Languages, Encyclopedias and Nomenclatures (GALEN) [3] employs the <Entity, Property, Quality value> (EPV) formalism, as in <John, height, tall>. The EAV triple is used also in [4], which employs Minsky's frame-based knowledge representation, the *de facto* standard in the field of artificial intelligence studies. The EAV triple distinguishes between the quality and quality value [2].

Within the Open Biomedical Ontology (OBO) community, the Phenotype Quality Ontology (PATO) provides a practical basis for vocabulary and semantics for the description of phenotype information across species [5]. PATO follows the Basic Formal Ontology (BFO) framework, [6] which recommends a variant of the <Entity, Property> formalism (EP), as for example in <John, height>, <John, above average height>, <John, 220 cm height>. This yields what are called "entity quality" (EQ) annotations of experimental parameters and parameter values; the EQ is an equivalent of the EP formalism. In contrast to the EAV and EPV approaches, this implies a single hierarchy of qualities, rather than a double hierarchy of both qualities and values.

PATO has contributed greatly to the development of a practical basis for the

qualitative and quantitative description of biological phenotypes. However, a number of problems still remain to be resolved, including the problem of classification of quality values, of measurements made in different contexts, and of variations in the descriptions of quality-related information created by separate research communities.

There are various efforts to improve ontologies of measurement and of qualities [7-9]. In this study, we attempted to expand the PATO ontology to ensure a more advanced framework within the YAMATA (Yet Another More Advanced Top-Level Ontology) framework [10]. Our goal is to integrate quality descriptions deriving from experimental studies in biomedicine through the development of a reference ontology named "PATO2YAMATO".

2 Practical Requirements of an Ontological Basis for Describing Biological Measurements

2.1 Fundamental Classification of Quality Values on the Basis of Measurement Scales

In the field of experimental biology, the results of measurements are described in terms of a variety of systems of "values". One of the most common classifications of values is "levels" or "scales" of measurement developed by Stanley S. Stevens [11]. This describes four different types of scales, namely "nominal (categorical)", "ordinal", "interval," and "ratio". This classification takes as its starting point the mathematical operation that is applied in analyses of measurement results. Therefore, such a classification is quite beneficial for data integration in the field of experimental biology. BFO provides a single node quality, and quality ontologies such as PATO are created by downward population from this single node (hence PATO is referred to as adopting a single hierarchy approach). DOLCE, in contrast, provides a bi-hierarchal system of classification for quality-related concepts involving both *quality-space* and *quale*, providing a classification of both the type of quality involved (in terms of *quality space*) and of the value or *quale*. Integration of qualitative and quantitative descriptions is realized by a combination of quality-space and quale in a single knowledge framework. As mentioned elsewhere [10], there seems to be some room for improvement in DOLCE in regard to the current running together or unclarity in its treatment of value, quality of real entity, and data about this quality.

2.3 Modeling of Context Dependencies of Ordinal Scale Values

In general, an ordinal value is used within a certain sort of phenotypic quality description when making comparison for example of *normal* versus *abnormal*, for instance in regard to some developmental or evolutionary change. The problem is that different contexts, here, can lead to different bases for comparison, illustrated by the fact that from *large ant*(x) and *small elephant*(y) we could never infer that x is larger than y. "Abnormally large" and "abnormally small", which are equivalent to "increased size" (PATO:0000586) and "decreased size" (PATO:0000587), refer to groupings of values based on deviation from the normal in specific biological populations such as species or strains. This yields a completely different view of classification from that which is obtained from a simple grouping into "small" and "large" (Fig 1).

On the other hand, when comparisons are based on deviations within distinct species are closely related to each other. For example, "abnormally largeness" of homologous skeletal elements in humans and in mice are often caused by mutations of homologous genes. This indicates that it might make sense to consider "abnormally large in ant" and "abnormally large in elephant" as subclasses of a single class "abnormally large". There are analogous interrelationships between classifications of biological species and of experimental population, and so on. To establish advanced integration of biological measurements, it is essential to provide an explicit and consistent way to document such interrelationships.

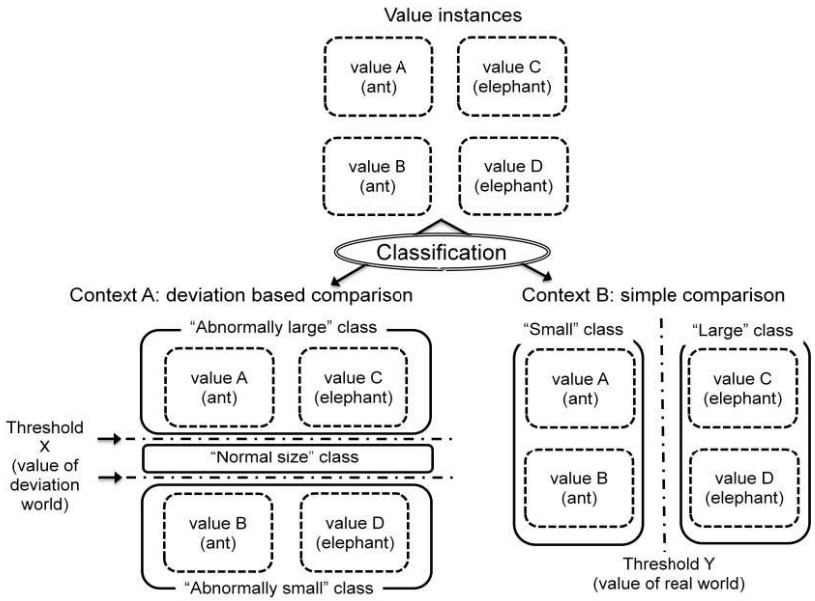

Figure 1. The problem of "large ant and small elephant".
Rounded squares with broken lines and solid lines represent instances and classes of quality values respectively.

2.3 Modeling of a Datum as an Informational Entity

Empirical measurements are always approximations that do not accurately represent what is being measured. In other words, the "true value", a quality (value) that is independent of any act of measurement, is clearly distinct from the "measured value" (or "datum"), a representation or description of the quality obtained through such an act. As pointed out already in [10], it is important for the integration of biological measurements to define a "datum describing a quality" as an informational entity that is related with the quality itself by some relation of aboutness.

While the acknowledgement of two entities – a datum to describe the quality in addition to the quality itself – seems to imply redundancy, this move is essential for any strategy leading to the integration of biological measurement data. To see why, consider the following example of a case of measurement of body weight (Wn) of a mouse performed using different procedures (A and B) along a time course (Tn). If all measured values are regarded as true values, the changes of the mouse's body weight would be recorded as:

w1(a, t1) w2(b, t2) w3(a, t3) w4(b, t4)

and so on (see Fig. 2A). In the case in hand, however, this will likely lead to a wrong interpretation of the changes in the body weight. A more standard interpretation would consider the two series of results, wi(a, ti) and wj(b, tj) independently, and then form an estimate of the true values of the mouse's weight at successive times, as in (Fig. 2B).

Within the bio-ontology community, phenotype annotations are usually recorded following the EQ or EAV formalisms. These formalisms are convenient to represent qualities in the biological measurements. In order to deal with symbolically formalized information items, an ontology of representation have been proposed in [12]. The Information Artifact Ontology (IAO) [13] also defines measurement data and descriptions as information entities to be related to the real-world in biomedical investigation. The sophisticated modeling and classification of "formalisms" to be used in the informational items and the facilitating the interoperability between these formalisms are then a further issue that must be addressed for the purposes of integration of biological measurement data.

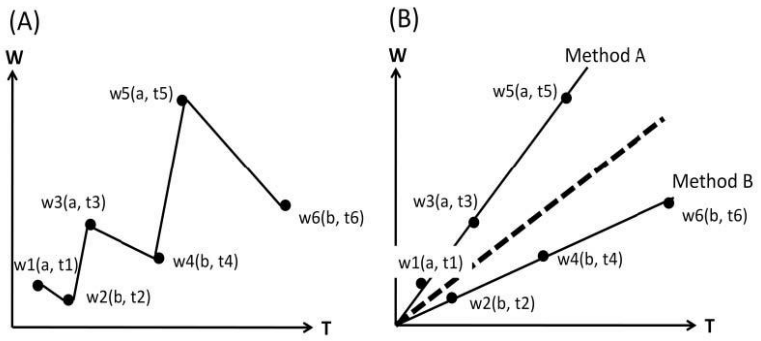

Figure 2. Distinction between "true value" and "measured value" is essential for proper interpretation of data obtained from different protocols.

3 An Ontology to Enhance PATO through Biological Measurements

To allow PATO to meet the above requirements, we have developed an ontology which incorporates PATO terms within the YAMATO upper ontology framework. YAMATO has the following characteristics [11]:

a) Quality-related concepts (dependent continuant entities) are divided into: "Property", "Generic quality" and "Quality value". "Quality" in BFO is identical to "Property" in YAMATO.

b) "Quality value" is classified in the same way as in classification in scales of measurement.

c) The context dependency of "Ordinal value" is represented by using the category of Role.

d) Multiple kinds of informational entities are represented symbolically. For the quality description, representations with both EAV-triple and EP-double formalisms are defined.

The ontology is developed via manual mapping. The basic working principles are as follows:

1) Assumption of context where PATO terms are used in deviation-based comparisons in mice, for example mutant vs. wild-type comparisons in mice.

2) Import of PATO terms such as "Property" into YAMATO.

3) Definition of Generic qualities from an imported attribute slim subset of PATO terms.

4) Definitions of context-independent Quality values for nominal scale values.

5) Definitions of context-dependent Quality values for ordinal scale values.

According to this workflow, we have currently manually mapped about 1000 PATO terms into the YAMATO framework using the Hozo Ontology Editor [14], where roles are clearly visualized in a human-understandable GUI. The ontology file of PATO2YAMATO is posted on our website [15].

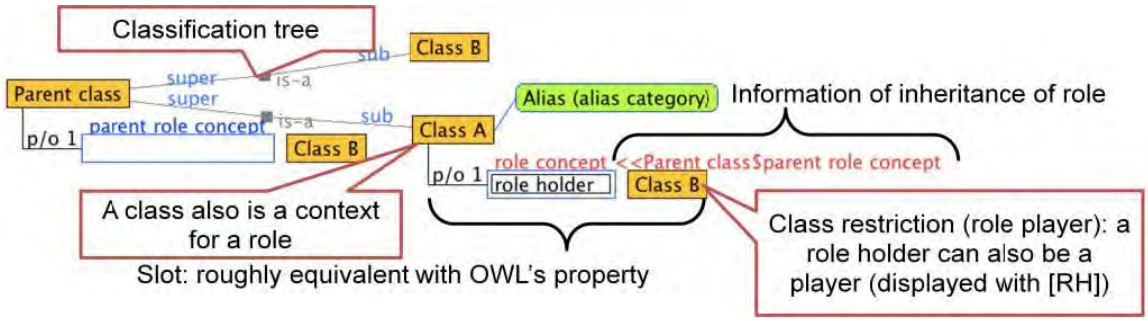

Figure 3. A legend of representation of ontology in this study (drawn by Hozo Ontology Editor).

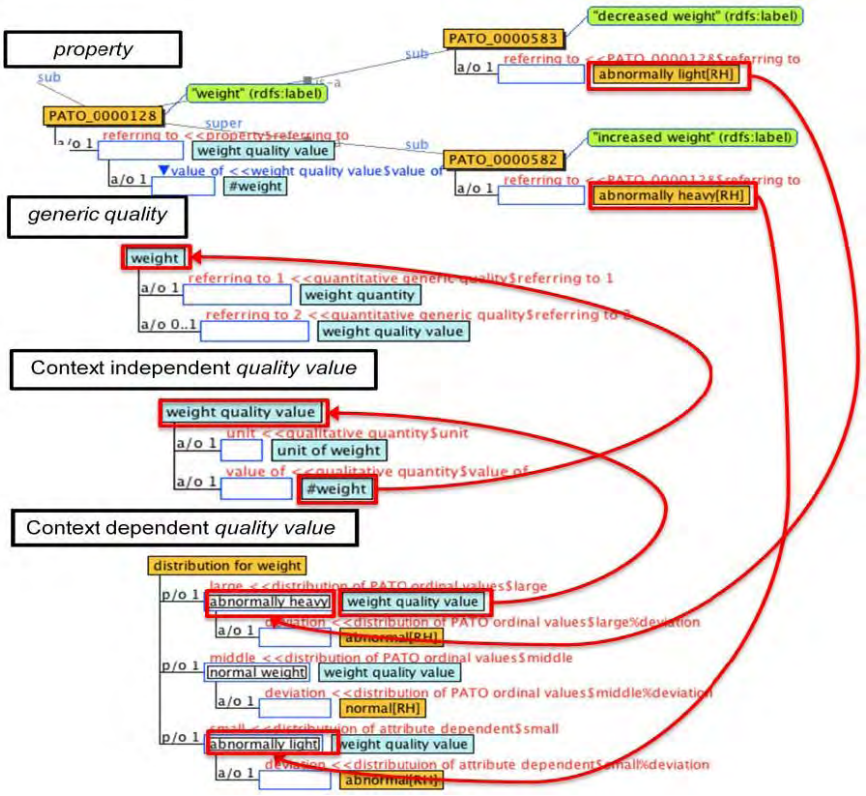

Figure 4. Example of the inter-relationships between imported original terms and newly defined terms in PATO2YAMATO. Orange and blue boxes represent terms to be incorporated into PATO2YAMATO and YAMATO respectively.

4 Summary of the Semantics of the Resultant Ontology

Figure 4 shows an example of the inter-relationships between imported original terms and newly defined terms. In the YAMATO framework, PATO:0000582 (increased weight) is defined as a *Property* (equivalent to BFO's *Quality*) that is philosophically a combination of a *Generic quality* (type of quality), *weight*, and a context-dependent *Quality value*: *Abnormally heavy*.

The context-independent value is defined as a class "*Weight quality value*". This class is instantiated in each specific context. In Figure 4, "*Increased weight*" is defined as being instantiated in the context, "*Distribution of weight*". This can be rephrased in the statement that "in the **distribution of weight**, some **weight quality values** playing **large**-roles thereby becomes role holders, **abnormally heavy**"

The classification of contexts for the ordinal values is illustrated in Figure 5. The context, "*Abnormally large, small and normal*" is the three-value comparison based on the deviations of abnormally large and abnormally small from normal values. This is clearly distinguished from the simple three-value comparison ("*Simple large, middle and small*" in Fig. 5). These values depend on the context of "*Distribution of ordinal values*", which is given by each species.

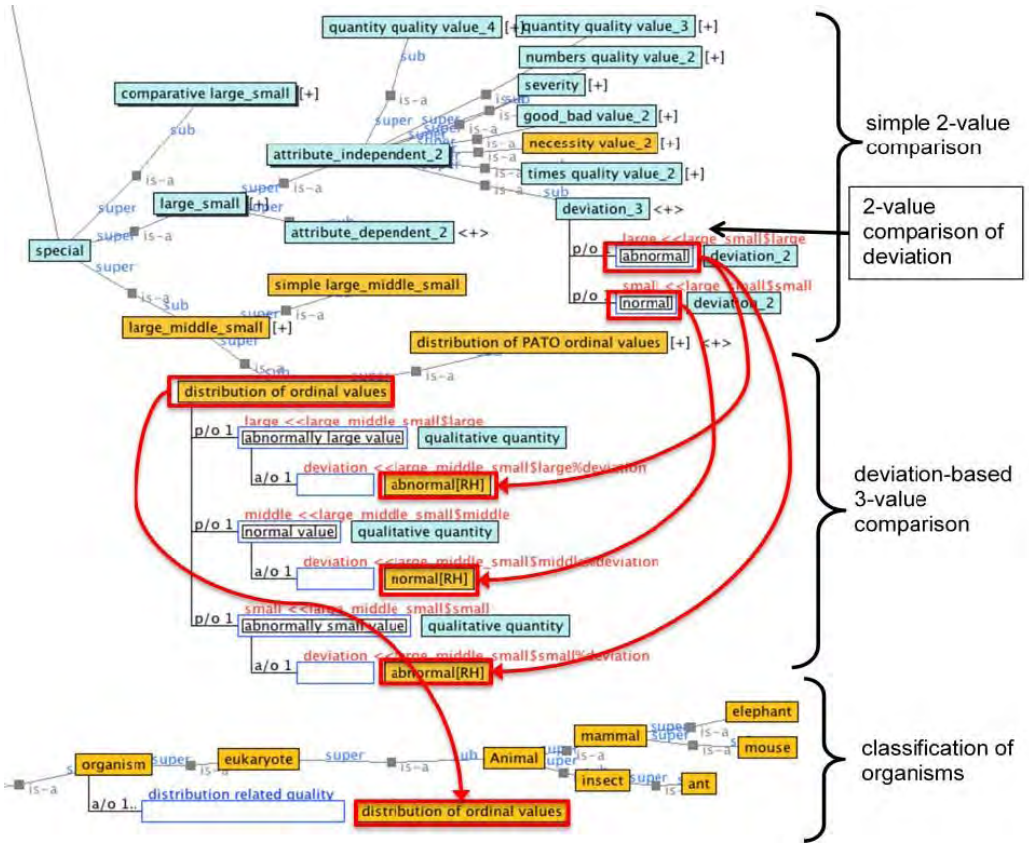

Figure 5. Classification of comparison contexts.
Values in the deviation-based 3-value comparison are related to those in the 2-value comparison of deviation (*deviation_3*).

5 Preliminary Evaluations of the Reference Ontology

We worked out two kinds of preliminary evaluations of PATO2YAMATO. First we performed a simple experiment of converting phenotype descriptions created with the pre-compositional Mammalian Phenotype Ontology (MP) [16] into the EAV formalism. We used the OBO cross-product file with logical definitions for the terms of MP using PATO [17]. Using the perl script referring to PATO2YAMATO, we converted 1450 MP terms with EQ annotation into *Quality representation* in EAV triple form. Each triple refers to *Generic qualities* and *Quality values* in the YAMATO framework. The perl script file and conversion results are posted on our website [15].

We then examined the dynamic classification of ordinal values and their contexts. Using Hozo's function to allow inheritance of slots and class restriction, we visualized contexts of the values: "*Distribution of weight*" and "*Abnormally heavy*", in each organism. As a result, Hozo's reasoner generated *is_a*-hierarchies in both of these contexts and values analogous to a species tree (Fig 6A). We also examined the problem of "large ant and small elephant". We defined two classes of *weight quality values*, "*Weight of Ant A*" and "*Weight of elephant B*", and large-small relationships as "B is heavier than A". Context dependent forms (role holders) of these concepts revealed that "*Abnormally light in elephant*" is heavier than "*Abnormally heavy in ant*" (Fig 6B)

6 Discussion

In this study, we developed an ontology, PATO2YAMATO, to integrate phenotype descriptions residing in different structured comparison contexts. The ontology exhibits several advanced features. Thus it allows: 1) classification of quality values, in which scales

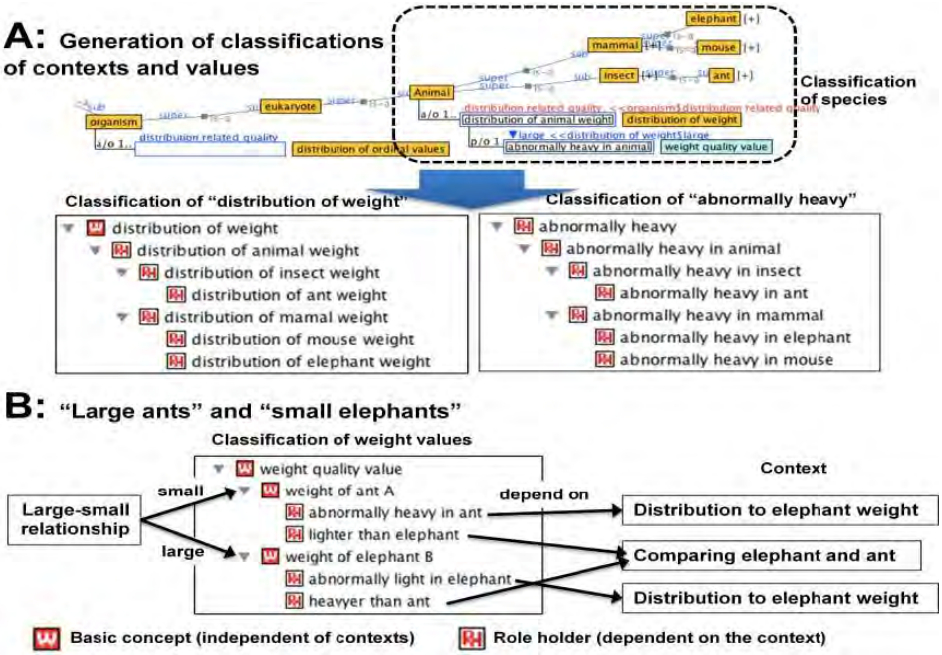

Figure 6. A: Generation of the inter-relationships among contexts and context dependent weight values. B: Representation of "small elephants are heavier than large ants".

of measurements are properly represented, 2) strict modeling of the context dependency of ordinal values, and 3) clear distinction between "true values" and "measured data" in which common formalisms are applied for quality descriptions.

It has been pointed out that EAV formalisms lead to some confusions so that those who are not familiar with ontologies often make incorrect triples, but there seem to be multiple solutions to such problems. First, YAMATO covers various formalisms for quality representation. For example, <liver, red, dark>, which is incorrect as EAV, can be processed using the EPV formalism. Second, some system can be developed to support the input of correct triples deriving from correct ontologies. Third, the conversion of existing EQ descriptions to EAV also helps to provide the advanced knowledge basis easily.

The results of our preliminary experiments suggest both: 1) that the ontology helps the automatic conversion from EQ to EAV, and 2) that deviation-based ordinal values created in contexts determined by distinct species can be generated automatically. For further evaluation, it will be necessary to perform a fully automatic generation of deviation-based ordinal values depending on multiple organism contexts, and to verify the inter-relationships generated. It has been reported that Hozo's Role model can be represented with the combination of Web Ontology Language (OWL) and Semantic Web Rule Language (SWRL) [14]. Verification using the OWL format would also help to establish broader interoperability.

We expect that the context dependency of ordinal values can be applied to the integration of biological measurements in general. Discrimination of experimental contexts is essential to the integration of experimental data derived from multiple groups of experimenters, as is shown by the case of large-scale mouse phenotyping programs such as that of the International Mouse Phenotyping Consortium (IMPC) [18], which seek a broad-based, systematic phenotyping of knockout mice through the world-wide cooperation of multiple research institutions. We have recently reported on a novel type of database system in which a top-level ontology is implemented for integration of the underlying knowledge and individual data records in existing databases to support advanced cross-database searches [19]. Such a methodology may also help the advanced integration of biological measurements.

One of the roles of ontology is to provide the basis of interoperability among different databases and data converters. This study

provides one of the basic procedures essential for a number of operations in the integration of biological measurement data to integrate the features of multiple top-level ontologies such as DOLCE and BFO. The Open Biomedical Ontologies (OBO) Foundry has coordinated the definition of scientific methods to be applied in ontological developments [20]. Toward forming a single, consistent, cumulatively expanding and algorithmically tractable whole, the OBO Foundry applied only BFO as the semantic framework. However, BFO itself has undergone and will continue to undergo modifications over time. Already for this reason, therefore, we believe that investigation with multiple top-level ontologies would facilitate not only YAMATO-based integration, but also the OBO foundry initiative to show the requirements and concrete examples of solutions.

Acknowledgments

With thanks to John Hancock, Paul Schofield, Michael Gruenberger, Toyoyuki Takada, Kuniya Abe, Ann-Marie Mallon, Chris Mungall, Shigeharu Wakana, Toshihiko Shiroishi, Yuichi Obata and InterPhenome members for meaningful discussion. This work is supported by the Management Expenses Grant for RIKEN BioResource Center, MEXT.

References

1. Aranguren M.E., Antezana E., Kuiper M., Stevens R.: Applying ontology design patterns in bio-ontologies, Proc. of 16th International Conference LNAI 5268, 7-16. (2008)
2. 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. 166-181. (2002)
3. Rector A.L., Gangemi A., Galeazzi E., Glowinski A.J., Rossi-Mori A.: The GALEN CORE Model Schemata for Anatomy: Towards a Re-usable Application-Independent Model of Medical Concepts. Twelfth International Congress of the European Federation for Medical Informatics, MIE-94, 229-233
4. Minsky M.: A Framework for Representing Knowledge, in: Patrick Henry Winston (ed.), The Psychology of Computer Vision. McGraw-Hill, New York, (1975).
5. Gkoutos G.V., Green E.C., Mallon A-M, Hancock J.M. and Davidson D. Using ontologies to describe mouse phenotypes, Genome Biol 6, R8, (2005)
6. Grenon,P., Smith,B.: SNAP and SPAN: towards dynamic spatial ontology. Spat. Cogn. Comput. 4, 69--103. (2004)
7. Masolo C., Borgo C.; Foundational Aspects of Ontologies (FOnt 2005) Workshop at KI (2005)
8. Masolo C.: Founding properties on measurement. Proceedings of the Sixth International Conference (FOIS 2010)
9. Probst F.: Observations, measurements and semantic reference spaces. Applied Ontology 3, 63-89, (2008)
10. Mizoguchi R.: Yet Another Top-level Ontology: YATO, Proceedings of the 2nd Interdisciplinary Ontology Meeting, 2, 91-101. (2009)
11. Stevens, S.S. On the Theory of Scales of Measurement. Science 103, 677--680. (1946)
12. Mizoguchi, R.: Tutorial on ontological engineering – Part 3: Advanced course of ontological engineering, New Generation Computing, OhmSha & Springer, 22, No.2, pp.198-220. (2004)
13. IAO, http://code.google.com/p/information-artifact-ontology/
14. Kozaki, K., Sunagawa, E., Kitamura, Y., Mizoguchi, R.: Role Representation Model Using OWL and SWRL, Proc. of 2nd Workshop on Roles and Relationships in Object Oriented Programming, Multiagent Systems, and Ontologies, 39-46 (2007)
15. PATO2YAMATO, http://www.brc.riken.go.jp/lab/bpmp/ontology/ontology_pato2yato.html
16. Smith C.L., Goldsmith C.A., Eppig J.T.: The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. Genome Biol. 6, R7. (2005)
17. Mungall C, Gkoutos G.V., Smith C., Haendel C., Lewis C., Ashburner M.: Integrating phenotype ontologies across multiple species. Genome Biology, 11. R2. (2010)
18. IMPC, http://www.mousephenotype.org/
19. Masuya H., Makita Y., Kobayashi N., Nishikata K., Yoshida Y., Mochizuki Y., Doi K., Takatsuki T., Waki K., Tanaka N., Ishii M., Matsushima A., Takahashi S., Hijikata A., Kozaki K., Furuichi T., Kawaji H., Wakana S., Nakamura Y., Yoshiki A., Murata T., Fukami-Kobayashi K., Mohan S., Ohara O., Hayashizaki Y., Mizoguchi R., Obata Y., Toyoda T.: The RIKEN integrated database of mammals. Nucleic Acids Res. 39, D861-870. (2011)
20. Smith B., Ashburner M., Rosse C., Bard J., Bug W., Ceusters W., Goldberg L.J., Eilbeck K., Ireland A., Mungall C.J.; OBI Consortium, Leontis N., Rocca-Serra P., Ruttenberg A., Sansone S.A., Scheuermann R.H., Shah N., Whetzel P.L., Lewis S.: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol. 25, 1251-1255 (2007)