Zoomorphology
DOI 10.1007/s00435-008-0081-5

REVIEW

The future role of bio-ontologies for developing a general data standard in biology: chance and challenge for zoo-morphology

Lars Vogt

Received: 15 April 2008 / Revised: 30 October 2008 / Accepted: 31 October 2008

Abstract Due to lack of common data standards, the communicability and comparability of biological data across various levels of organization and taxonomic groups is continuously decreasing. However, the interdependence between molecular and higher levels of organization is of growing interest and calls for co-operations between biologists from different methodological and theoretical backgrounds. A general data standard in biology would greatly facilitate such co-operations. This article examines the role that defined and formalized vocabularies (i.e., ontologies) could have in developing such a data standard. I suggest basic criteria for developing data standards on grounds of distinguishing content, concept, nomenclatural, and format standards and discuss the role of data bases and their use of bio-ontologies in current activities for data standardization in biology. General principles of ontology development are introduced, including foundational ontology properties (e.g. class–subclass, parthood), and how concepts are defined. After addressing problems that are specific to morphological data, the notion of a general structure concept for morphology is introduced and why it is required for developing a morphological ontology. The necessity for a general morphological ontology to be taxon-independent and free of homology assumptions is discussed and how it can solve the problems of morphology. The article concludes with an outlook on how the use of ontologies will likely establish some sort of general data standard in biology and why the development of a set of commonly used foundational ontology properties and the use of globally unique identifiers for all classes defined in ontologies is crucial for its success.

Keywords Bio-ontology · Data standard · Linguistic problem of morphology · Morphology · RDF

L. Vogt (✉)
FU Berlin, Fachbereich Biologie Chemie Pharmazie,
Systematik und Evolution der Tiere,
Königin Luise Str. 1-3, 14195 Berlin, Germany
e-mail: lars.vogt@zoosyst-berlin.de

Introduction

In the past, the field of biological research and knowledge experienced a continuous process of differentiation and diversification into different disciplines and communities of researchers to a degree that is unrivaled within natural sciences, having led to significant differences in traditions and schools of thought, methods and techniques applied.

One obvious reason for the strong impact of this process is the fact that organisms exhibit a complex hierarchical organization (see 'scalar hierarchy' Salthe 1985, 1993; 'levels of organization' Wimsatt 1976, 1994; 'cumulative constitutive hierarchy' Valentine and May 1996; 'Theorie des Schichtenbaus der Welt' Riedl 2000). One can distinguish for instance a molecular and a cellular level of organization, a level of tissues and a level of organs, up to the level of organization of multicellular organisms or the organization of eco-systems. Thereby, an increase in organizational level is usually accompanied by an increase in structural diversity—the higher the organizational level, the more complex the structures and the higher the degree of structural diversity.

Biological objects of different organizational levels differ from one another quite substantially, having their own essential structural properties and functional relationships. Thus, it is not surprising that each level has its own typical scientific questions and research programs, and the methods for producing data for one level of organization are often

Published online: 20 December 2008

completely different from the methods used for producing data for another level. As a consequence, in correlation with the different levels of organization one has to distinguish fundamentally different types of data in biology.

While each level of organization possesses a set of unique and characteristic structural and functional properties that distinguishes it from another level, objects of different levels of organization are yet functionally and materially interlinked and constrain one another by up- and downward causation (Trewavas 2006). Objects of different levels therefore inter-depend on one another with respect to causal processes in which they are involved as well as with respect to their compositional integrity. Experience has shown that due to emergent properties (sensu Mahner and Bunge 1997) the knowledge of all relations, properties, and processes of all the objects of one level not necessarily enables one to deduce comprehensive knowledge for all objects of another level of organization—knowing the nucleotide sequence of a particular gene and all its chemical properties as such tells us nothing about its origin or about the function of its transcript (see e.g., Polyani 1968; Schutt and Lindberg 2000; Trewavas 2006). Thus, in order to gain comprehensive knowledge of a given biological system, it is insufficient to take in a reductionist approach and exclusively investigate only the molecular level—quite contrary research into all organizational levels of the biological system in question is required.

Due to increasing methodological specialization among biologists, however, a biologist often produces data referring only to a single level of organization and has no experience with methods and techniques that are required for producing data and no knowledge and experience with established quality standards for interpreting and analyzing data referring to other levels of organization. This is unfortunate since the functional interlinks between objects and processes of different levels of organization, as for instance between the DNA and the protein level or between the protein and the cellular and organ level, become more and more interesting for biologists, especially with regard to systems biology (e.g., Trewavas 2006). Thus, in order to successfully study these interlinks, it is essential that biologists from various backgrounds co-operate within interdisciplinary research programs. However, the progressive specialization of biologists significantly complicates the inquiry of functional interlinks, as it requires biologists with such different methodological and theoretical backgrounds as for instance molecular biologists and morphologists to co-operate and work together. In order to compensate the lack of practical experience with parts of the data and the methods and techniques applied for their production, communication about the different data types and their corresponding quality standards becomes very important for the biologists who are involved in such co-operations.

Thus, modern biology is situated within a field of tension between differences in (a) methods and techniques applied during research practice, (b) quality standards required for correctly interpreting and analyzing data, (c) types of explanations sought, (d) types of data produced and analyzed, and (e) philosophies and research strategies applied, all of which somehow depend on the choice of the level of organization studied. All this resulted in the development of specialized languages for communicating data and metadata within specialized communities—biologists working with different model organisms, different techniques, or with structures of different levels of organization often speak different (scientific) languages, to a degree that makes their co-operation difficult.

Unfortunately, biology is lacking a commonly accepted general data standard, which could facilitate such co-operations. Most types of biological data are lacking such a standard. For instance on the morphological level biologists have to deal with the linguistic problem of morphology, causing fundamental ambiguities regarding morphological terminology and resulting in a serious slowdown of scientific progress in morphology (Vogt et al. 2008). But also on the genetic level, biologists had to deal with a lack of standards regarding gene names and spellings (Brazma 2001; Stein 2003), not to speak of the conceptual problems of agreeing on a common gene concept (e.g., Beurton et al. 2000; Wilson 2005; Gerstein et al. 2007; Griffiths and Stotz 2007; Scherrer and Jost 2007; Prohaska and Stadler 2008). As a consequence, communication of data is not trivial and comparing data sometimes even impossible, which considerably hampers co-operations between biologists. Comparability and compatibility of data can best be accomplished by standardization of data production and representation, which would at the same time enable their reliable communication. The standardization of biological data would significantly increase effectiveness of the performance of integrated analyses over different types of data and particularly co-operations among biologists of different methodological backgrounds, not to speak of all biological research that is based on comparison. The standardization of biological data could provide all biologists with a common (meta-)language; and a common language is one of the most important prerequisites for the integration of a group.

Considering that even the different types of biological data are lacking a commonly accepted standard, the development of a general data standard for all types of biological data represents a real challenge. But it also represents an important task—at least as long as biology does not want to continue to waste a significant fraction of its resources for the correction of incorrectly used data and the translation of individual data formats, which is very time-consuming. Is it possible to develop such a general data standard for all types of biological data? What are the components of such

a standard? Would not such a standard also constrain future developments within biological research? These are the questions I am addressing in this article.

Recognizing the need for data standardization in biology

Standardization and data quality

A standardization of data is usually accompanied by a quality increase, which is the reason for developing a data standard in the first place. But what does quality of data mean in that respect?

Besides research dependent specific criteria for the assessments of data quality, such as various statistical measures of support of data for a given explanatory hypothesis (i.e., evidential weight), there exist some general criteria of data quality that are common to all empirical sciences. These usually involve reference to transparency and reproducibility of methods and techniques applied during data production and are relevant for comparability and compatibility of data. Biologists usually have this latter type of data quality in mind when they develop data standards in their respective field of research.

Assuming that the ultimate structure of reality is independent of humankind and human culture (ontological objectivity sensu Daston and Galison 1992), increasing the communicability of scientific data increases their objectivity. This is referred to as aperspectival objectivity (see Daston 1992, 1998; Daston and Galison 1992; procedural objectivity sensu Heintz 2000; for a critique of the claim for aperspectival objectivity in science see, e.g., Kukla 2006). Aperspectival objectivity claims that something is more objective than something else if it relies less on the specific individual who generated the results, their social position and character. While increasing aperspectival objectivity is either implicitly or explicitly part of most data standardization efforts in biology, what is referred to as mechanical objectivity (Daston and Galison 1992; methodical objectivity sensu Heintz 2000) is another part. Mechanical objectivity requires ruling out all individual and subjective influences of body and mind and forbids judgment and interpretation in documentation and reports of observation (Daston and Galison 1992). It involves the establishment of specific experimental and measurement standards, with the purpose to rule out the fallibility (in terms of deceivability) and deficiency of human cognition. Thus, most data standardization efforts are aimed at increasing aperspectival and mechanical objectivity of data, and when I refer to data quality within this paper it is always in reference to these two types of objectivity.

Minimum information checklists

The need for standardizing biological data has been recognized in various fields of biological research—especially within molecular biology and the model organism research communities. The introduction of the so called OMICS technologies (i.e., technologies from genomics and proteomics, including DNA microarray technologies for expression, genetic and epigenetic analysis, metabolic profiling techniques, etc.) resulted in an exponentially growing body of raw data produced by high through-put methods and required the development of new techniques and approaches for data management that enable easy sifting through of large amounts of data in order to be analyzed, compared, and disseminated through public data bases (e.g., Brazma 2001; Field and Sansone 2006). Especially the variation in the results of OMICS experiments, which are usually associated with details of the processing of samples and which are highly dependent on the condition, age, and history of the samples (Field and Sansone 2006), required transparency regarding such contextual information in order to guarantee unlimited usability of the respective data. Without corresponding annotations most OMICS analyses would be hard to understand and very difficult to interpret correctly. Within OMICS, these annotations, which represent data about data (i.e. metadata), usually refer to experimental design, the source, preparation, and treatment of the biological material being studied, the parameters and values of instruments used, and other information (Field and Sansone 2006). Thus, standardizing data to require the inclusion of (standardized) metadata became a crucial condition for the success of OMICS research programs. This became especially apparent in the context of multi-OMICS approaches, which attempt to understand complete biological systems and require data from various fields of biological research.

It requires the development of a commonly accepted reporting structure that specifies which annotations (i.e., metadata) in which form should be included for publication and dissemination of data. A first step towards such a reporting structure is to identify and agree upon which information must be included in a data report. This has been referred to as minimum information convention and the minimum information about a microarrray experiment (MIAME) checklist (Brazma et al. 2001), developed by the Microarray Gene Expression Data Society, has been the first of its kind, specifying which minimum information should be reported about a microarray experiment. The MIAME checklist had a great impact on other OMICS standardization efforts and subsequently became a benchmark for the development of various minimum information checklists (see table 1 in Field and Sansone 2006).

Unfortunately, within zoo-morphology respective attempts are usually restricted to standardization of taxonomic

information (e.g. Kennedy et al. 2006) or to specific model organisms and their corresponding data bases—there is no taxon-independent standard for morphological metadata established so far and all attempts are in their beginnings. Much worse, it seems that the need for such a standard is not commonly recognized in the community of morphologists. Consequently, the motivation for participating and contributing to a standardization effort is very limited within the community. This is very unfortunate, since morphology, compared to other biological disciplines, suffers significantly from a low degree of aperspectival and mechanical objectivity (see Vogt et al. 2008, Vogt 2008).

Biological data bases

Data bases become more and more popular in biology. Besides general data bases for molecular data, a lot of data bases have been developed that are restricted to data referring to a specific model organism, as for instance FlyBase for *Drosophila* and Arabidopsis Information Resource for *Arabidopsis thaliana*. Other data bases have been developed restricted to a specific taxonomic group, as for instance Antbase, Fishbase, or AmphibiaWeb (see Table 1).

Every data base has to define what information can be uploaded by whom in which way, and it necessarily has its own standardized way of storing and presenting data and metadata. Thus, each data base inevitably sets its own data and metadata standard. In the beginning of the development of biological data bases the focus was on defining what type of information should be stored (this is comparable to minimum information checklists). However, terminology problems such as the lack of standards of gene names and spellings (Brazma 2001; Stein 2003) resulted in fundamental problems regarding the comparability and compatibility of the content of a data base. Thus, although a data base might have specified which type of information has to be included when uploading new data, relevant information concerning a specific gene can be present in a data base but still, due to terminological or conceptual ambiguities, not be found reliably, which turns the initial purpose of the development of the data base upside down. Thus, terminological and conceptual ambiguities made a change of focus necessary, and more and more data bases started to put a lot of effort into the development of defined and controlled vocabularies in order to deal with these problems.

The development of a defined and controlled vocabulary which also incorporates relationships (i.e., an ontology) requires the explication of the meaning of terms and concepts and the rules of their application. Using an ontology within a data base has the potential to significantly increase the formalization and standardization of data presentation, storage, and documentation.

In the past, various efforts have been made to impose ontologies on data bases. Some of the projects focus on taxonomic classification and nomenclature, attempting to provide a comprehensive list of all known species together with a validation of their respective taxonomic names and relevant metadata such as taxonomic ranks, vernacular names, and synonyms (see e.g., Bisby et al. 2002; Gewin 2002; Godfray 2002; Wilson 2003; see also taxonomic indexing, Patterson et al. 2006). Such efforts are important, since often the scientific past acts like a heavy weight on everybody who attempts to revise a taxonomic group. This weight is the result from hundreds of years of sedimentation of taxonomic and nomenclatural revisions, bringing about a dubious complex of synonymy and scattered type material (Godfray 2002). Many of the projects listed in Table 1 aim at bringing stability and unambiguousness of taxonomic names and affiliated metadata into this dubious complex. This is insofar a very important task, as taxonomy and nomenclature provide the fundamental reference system for all biology.

However, general terminological ambiguities become an increasingly severe problem with increasing amounts of data that have to be sighted and with different types of biological data that have to be analyzed simultaneously.

Morphological data bases

Within the last years, some interesting morphological data bases and projects started. MorphBank (http://morphbank.csit.fsu.edu) is an open web repository of images for the documentation of specimens and vouchers for sharing research results in taxonomy, morphometrics, morphology, and phylogenetics. Another project, MorphoBank (http://morphobank.com), is a repository for storing digital images (Pennisi 2003). It catalogues images and allows the labeling of structures on the images and the display of editable phylogenetic matrices, to which the images can be linked. A different project, Digital Morphology (DigiMorph; http://www.digimorph.org), is an archive of digital morphological images and 3D models. None of the abovementioned morphological data bases, however, have their focus on managing morphological descriptions and, thus, morphological data in the strict sense (see Vogt et al. 2008). Instead, they mainly manage all kinds of different morphological media files.

Bio-ontologies, RDF, and zoo-morphology

Bio-ontologies and knowledge bases

By now, many biological data bases use ontologies (not to be mistaken with Ontology in philosophy, which is the

Table 1 List of taxonomy and phylogeny projects and model organism data bases

Project	URL	Description
Species 2000 and the Integrated Taxonomic Information System (ITIS)	http://www.sp2000.org http://www.itis.gov	Two of the most prominent taxonomic projects with the aim to create an electronic global framework for taxonomy. They joined their forces in the catalogue of Life consortium with the aim to catalogue all known organisms by establishing a federation of interoperable data bases for documenting the world's taxonomic knowledge
All Species Foundation	http://www.all-species.org	They bring together taxonomists and high-tech people in order to name and describe all living species within the next 25 years by utilizing web-based approaches and recruiting 'parataxonomists'
Universal Biological Indexer and Organizer (uBio)	http://www.ubio.org	uBio is an initiative gathering all names for taxonomic indexing purposes (for taxonomic indexing see Patterson et al. 2006)
Global Biodiversity Information Facility (GBIF)	http://www.gbif.org	GBIF develops an interoperable network of different biodiversity data bases and information technology tools that require defined vocabularies in order to find and utilize the information stored in the various data bases, therewith developing standards for interoperability and digitizing biodiversity data
Species Analyst	http://Speciesanalyst.net	Species Analyst represents a research project developing standards and software tools for access to the world's natural history collections and observation data bases
Taxonomic Search Engine (TSE)	http://darwin.zoology.gla.ac.uk/~ rpage/portal	TSE provides a platform for searching several data bases including ITIS, Index Fungorum, and NCBI
Taxonomic Databases Working Group (TDWG)	http://www.tdwg.org	TDWG attempts to establish international collaboration among biological data base projects by developing, adopting, and promoting standards and guidelines for the recording and exchange of data about organisms
Tree of Life web project	http://www.tolweb.org	A data base of phylogenies
TreeBase	http://www.treebase.org	TreeBase is a data base for storing phylogenetic trees and phylogenetic character matrices
Projects developed by specialized taxonomic communities	Index Fungorum (http://www.speciesfungorum.org/Names/Names.asp), AlgaeBase (http://www.algaebase.org), ILDIS LegumeWeb (http://www.ildis.org), the International Plant Names Index (http://www.ipni.org), FishBase (http://www.fishbase.org), AntBase (http://antbase.org), AmphibiaWeb (http://amphibiaweb.org), some of which are affiliated to the Catalogue of Life consortium	
Database	**URL**	**Description**
FlyBase	http://flybase.bio.indiana.edu	A data base of genetic and molecular data for *Drosophila*
Saccharomyces Genome Database (SGD)	http://www.yeastgenome.org	A data base of the molecular biology of the yeast *Saccharomyces cerevisiae*
Mouse Genome Informatics (MGI)	http://www.informatics.jax.org	Provides integrated access to data on the genetics, genomics, and biology of the laboratory mouse
Arabidopsis Information Resource (TAIR)	http://www.arabidopsis.org	A data base of genetic and molecular biology data for *Arabidopsis thaliana*

study of being or existence) for providing defined and controlled vocabularies that incorporate also relationships. Ontologies consist of a vocabulary of terms and some specifications of their meaning that is used to describe a certain reality. A bio-ontology is a formal way of representing knowledge of a particular scientific field of biology through concepts and represents a data standard (Wang et al. 2005)—"a specification of a conceptualization" (Gruber 1993). An ontology is structured and formalized through a set of formal rules and assertions that describe the relationship between the concepts in a computer parsable form.

An ontology is build by a set of statements each of which is composed of two types and their relation to one another: '*Type_X* **relation** *Type_Y*'. The relations of an

ontology play an important role since they carry all the semantic content. The concepts of an ontology are described both by their meaning and their relationship to each other (see also Bard 2003; Bard and Rhee 2004). Therefore, by its rules and assertions, an ontology imposes a structure on a knowledge domain that constraints the possible interpretations of terms (Stevens et al. 2000). An ontology is usually intended to be explicit and complete. Its possible applications outclass those of an indexed set of terms and definitions as it is common for dictionaries or glossaries. The concepts of an ontology represent classes of defined terms and their inter-relationships and should not contain empirical data (i.e. instances) in principle. If empirical data and an ontology are combined in a data base one receives what is called a knowledge base (Stevens et al. 2000).

Within the last decade research on ontologies has increased tremendously, and, as a result, more and more bio-ontologies become available. While most of them are restricted to one specific model organism, one of the most important bio-ontologies, the Gene Ontology (GO), represents an exception. GO is an ontology that integrates genetic data about gene products with knowledge of their properties (Bard 2003). GO provides a standardized, controlled vocabulary for genome annotation systems, cataloging information about the structural and cellular location of gene products, about the processes to which these products contribute, and the functions that they fulfill (Stevens et al. 2000; Bard 2003). Currently, GO contains over 1.6 million annotated gene products (Gene Ontology Consortium 2006), which are connected by class–subclass and parthood relationships. GO is also used for data analysis (Blake 2004). GO was initiated by the model organism data base informatics community and involved the yeast (SGD), fly (FlyBase), and mouse (MGI) data base and soon expanded to include other model organism data bases, such as for instance the *Arabidopsis* Information Resource (TAIR) (Blake 2004). In the mean time, other genome centers such as NCBI have also incorporated the GO annotation sets into their systems (Blake 2004).

Morphological ontologies

The first bio-ontologies referring to higher levels of organization have been used in data bases specialized for data of specific model organisms and are, thus, customized for handling model organism specific data. By now, many model organism ontologies exist, for example: for human data the GALEN CORE Model of the openGalen project (http://www.opengalen.org), the Foundational Model of Anatomy ontology (FMA, http://sig.biostr.washington.edu/projects/fm), the Human developmental anatomy ontology, and the Human disease ontology for various medical data bases; the mouse adult gross anatomy, gross anatomy and development, and pathology ontologies for the Mouse Genome Informatics data base (MGI; http://www.informatics.jax.org), the edinburgh mouse atlas project (emap; http://genex.hgu.mrc.ac.uk), and Pathbase (http://www.pathbase.net); the Zebrafish anatomy and development ontology for the Zebrafish Model Organism Database (ZFIN; http://zfin.org); the *C. elegans* development, gross anatomy, and phenotype ontologies for WormBase (http://www.wormbase.org); the Drosophila development, gross anatomy, and fly taxonomy ontologies for FlyBase (http://flybase.org); and the yeast phenotypes ontology for the *Saccharomyces* Genome Database (SGD; http://www.yeastgenome.org). All these ontologies are available from the Open Biomedical Ontologies website (OBO; http://www.obofoundry.org).

Unfortunately, these specialized ontologies cannot be applied to a broader taxonomic range, and none of the abovementioned morphological data bases (MorphBank, MorphoBank, and Digital Morphology) has a defined and formalized, taxon-independent morphological ontology implemented so far. This is owed to the fact that currently there is no anatomy ontology available that is comparable to GO in terms of GO's degree of taxonomic independence and terminological depth. However, morphologists solely start to recognize the requirement and utility of ontologies in morphology and respective projects on the way (e.g., Ramírez et al. 2007). Anyway, so far no morphological ontology has been developed that covers large zoological taxonomic units as for instance the Metazoa or the Eukaryota. The only anatomy ontology referring to a larger taxonomic group is the Bilateria anatomy ontology, which is currently under development and which will be implemented in 4DXpress (4dx.embl.de/4DXpress), a data base that does not focus on managing morphological data but that has been developed for comparing gene expression data from already existing data bases such as the abovementioned Zebrafish Model Organism Database.

Other morphology centered initiatives focus on mediating between existing species-specific anatomy ontologies, making them comparable and compatible: while the Uber anatomy ontology is a multi-species anatomy ontology for comparison across multiple species and is generated semi automatically from the union of existing species anatomy ontologies, therefore suffering from limited applicability (http://www.bioontology.org/wiki/index.php/UBERON:Main_Page), the Common Anatomy Reference Ontology (CARO; http://www.bioontology.org/wiki/index.php/CARO:Main_Page), on the other hand, is currently being developed to facilitate interoperability between existing species-specific anatomy ontologies, which requires the development of these species specific ontologies in the first place (to which it might actually provide some guidance).

The Morphological Descriptions Data Base (MorphDBase; http://www.morphdbase.de) is a relational data base, which attempts to provide a platform for uploading different types of phenotypic information including all kinds of media files and morphological descriptions based on a morphological ontology (i.e. MorphOntology) in the near future. MorphOntology is currently being developed and is accessible through the MorphDBase website. It will provide taxon-independent concepts for morphological structures that do not refer to homology assumptions and will standardize morphological descriptions.

The lack of a general zoological anatomy ontology is very unfortunate, since the generation of morphological data has a long lasting history, going even back to ancient times, and a rich deposit of valuable information lies buried in the grounds of the extensive literature of past centuries. Unfortunately, due to the linguistic problem of morphology (see below), the old literature has to be thoroughly studied in order to evaluate what part of the enclosed information meets modern scientific quality standards. The relevant morphological information then has to be translated into present day terminology. As a consequence, the extraction of morphological data from literature is often extremely time-consuming—and as long as the results of these efforts are not documented in detail and not made publicly available, morphologists will have to repeat extracting them again and again. In the near future, MorphDBase might provide an appropriate platform for documenting legacy data that has been extracted and translated from older morphological literature, which would make valuable data available to every body without having to repeatedly put the same time-consuming effort into its evaluation and translation.

Resource description framework (RDF)

The method of representing an ontology takes in a crucial role since it has to be highly standardized and formalized in order to be applicable with description logics and utilizable for many different software applications. The Resource Description Framework (RDF) (http://www.w3.org/RDF) has become the most accepted general method for modeling knowledge (with OWL and RDFS as its most common implementations—see below). In RDF, relationships between resources are described by connecting one resource to another through a property. A resource is anything that is identifiable by a uniform resource identifier (URI) reference (Manola and Miller 2004).

The basic information unit in RDF is an RDF statement consisting of the triple *Subject* **property** *Object* (in the remainder of the paper, *Subject* and *Object* will be written in italics while the **property** will be in bold font) and represents a special case of the '*Type_X* **relation** *Type_Y*' statement mentioned above. The *Subject* represents the object that is being described, the **property** specifies the relationship or property type between *Subject* and *Object*, and the *Object* specifies the value of the property and is either another resource or a literal string. For example: '*Coelom* **has_part** *Coelothel*', '*Arenicola_marina* **is_synonymous_with** *Lumbricus_marinus*', '*Protonephridium* **actively_participates_in** *Excretion*', '*Epithelial_Cell* **is_a** *Polarized_Cell*', and '*Arenicola_marina* **has_maxlength(m)** *0.25*' represent five RDF statements—the former four connect two resources, the latter a resource with a string.

Each RDF statement can be modeled as a graph comprising two nodes connected by a directed arc (Fig. 1). A collection of such RDF graphs can jointly form a directed labeled graph (DLG) (see Fig. 2). Such a DLG at its turn can, in theory, model most domain knowledge (Wang et al. 2005) and is a useful tool for analysis and equivalence calculation using graphs logics. A collection of RDF statements can be used to represent an ontology.

RDF is a (meta-)data model and not a specific description language for metadata. In order to make RDF computer parsable it requires syntax. Typically, RDF uses a defined XML syntax (Beckett 2004) or N3 (Berners-Lee 2005) and the semantics via reference to RDF Schema language (RDFS) (Brickley 2004) or Ontology Web Language (OWL) (McGuinness and van Harmelen 2004). RDFS and OWL represent languages that are based upon RDF and offer support for machine processing and inferences (Wang et al. 2005).

The advantage of RDF lies in its ability to provide a way of explicitly describing semantic relationships. Moreover, since RDF statements can be modeled as a DLG, adding further RDF statements (i.e. adding nodes and edges to a DLG) does not change the structure of the existing statements. In other words, extending the data structure induces no fundamental change of the entire data structure—new RDF statements neither change nor negate the validity of already existing RDF statements.

Since an existing RDF statement can serve as subject or object in another RDF statement, even reification is possible. This allows the documentation of both consensus and disagreement about the same resource. It further allows additional commenting and supplementing of any RDF statement. Since all these features meet common requirements in scientific communities, RDF seems to be perfectly suited for academic use (Wang et al. 2005).

Fig. 1 A RDF statement modeled as a directed labeled graph (DLG). *Subject* and *Object* represent the nodes and the **property** the edge that connects the nodes

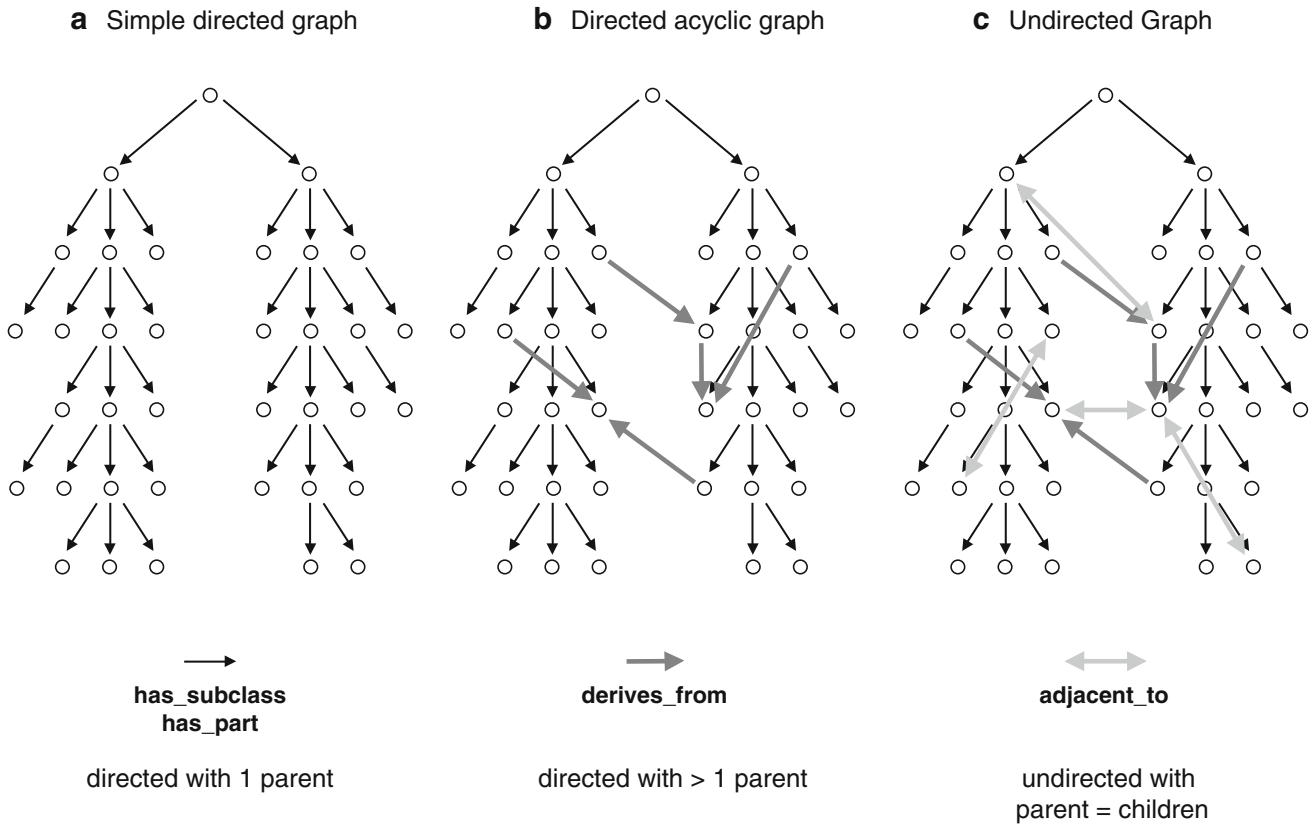

Fig. 2 Different types of graphs. **a** A unidirectional rule that allows only a single parent (e.g., **has_subclass**, **has_part**). It can be modeled as a simple directed graph representing a tree. **b** A unidirectional rule that allows for more than one parent (e.g., **derives_from**) can be modeled as a directed acyclic graph in which the graph itself can be traversed in several ways, with more than one path linking two nodes. **c** A bidirectional rule that imposes no directional constraints (e.g., **adjacent_to**), resulting in an undirected graph

General principles of ontology development

Foundational ontology properties

Within an ontology, a term has to be defined by means of its relationship to other terms. This is achieved through the **property** part of the RDF statement. Thus, properties take in a central role in every ontology, since all classes defined in an ontology refer to them. There is a variety of possible properties to choose from, and one should be very careful in deciding which properties one incorporates in an ontology and how to define/characterize them to avoid redundancies and inconsistencies.

Ontology properties describe the interactions or relations between concepts or a concept's properties (Stevens et al. 2000). The **instance_of** relationship, the class–subclass relationship **is_a**, and the parthood relationship **part_of** belong to the most foundational properties in an ontology.

A particular individual object or event (i.e. a datum) is referred to via the **instance_of** property with which it is associated to a concept of the ontology. It thus relates an instance to a class. With this property empirical data can be associated to an ontology implemented in a data base, thereby forming a knowledge base.

The **is_a** property, on the other hand, relates a class (i.e. concept) to another class. With the **is_a** property one can specify hierarchical relationships of classes and their subclasses, which, from class to subclass, results in a taxonomy of more and more specialized concepts, implying a hierarchical organization of terms (i.e. taxonomic inclusion, Bittner et al. 2004). This hierarchy can be modeled as a simple directed graph with nodes and edges representing parent-child relationships, relating concepts across a tree structure (Fig. 2a). In the class–subclass relationship every child has only a single parent, with the defining properties of a parent being inherited downstream to its children (i.e. downward propagation).

The parthood property **part_of** describes relationships between two or more individual instances (i.e. objects or processes) in which one instance represents part of another instance. Parthood relations (and all topological relations) never cross the dichotomy between objects and processes—an object is never part of a process and vice versa (Grenon et al. 2004). Thus, following that basal dichotomy one

distinguishes parthood relations between objects (e.g., **spatial_part_of**) from parthood relations between processes (e.g., **phase_of**). Processes are partitioned in time, resulting in a sequential order of sub-processes of a lower level of complexity (i.e. phases of a process). The life history of an organism, for instance, can be partitioned into the sequential order of larval/embryonic phase, juvenile and adult phase. On the other hand, spatial parthood is defined on the basis of topological inclusion (i.e. mereological inclusion). All the spatial parts of one level of structural organization have spatio-topological relations to one another that are multidimensional and, unlike temporal parts, cannot be ordered into a sequence—my brain, my nose, and my ears are all **spatial_part_of** of my head, but I cannot specify any natural sequential or hierarchical order for them. Though both types of parthood properties represent relations between individual instances, considering certain rules (e.g., Smith et al. 2005; Bittner 2004) one can also talk about parthood relations between universals (i.e. concepts, classes).

Parthood relations between individual instances and between concepts can also be modeled as a simple directed graph. Contrary to the class–subclass relation, however, properties are inherited upstream (i.e. upward propagation), from children to parent, and not downstream as in the class–subclass relationship. The resulting hierarchy of classes based on the parthood relation is usually called a partonomy (Smith and Rosse 2004), and one can distinguish temporal and spatial partonomies.

There are, of course, many more properties which are commonly used in ontologies but cannot be discussed here, as for instance **adjacent_to**, **derives_from**, **actively_participates_in**, **is_synonymous_with**, and others (e.g., Stevens et al. 2000; Smith 2004). Unfortunately, currently many ontologies use their own properties and property definitions/characterizations, which considerably hampers comparisons of data across different ontologies and the transfer of data between data bases that use different ontologies (but see Smith et al. 2005).

Logical properties

Each property can be classified according to its logical properties. For instance is the **is_a** property *transitive* (if A_1 **is_a** A_2 and A_2 **is_a** A_3, then A_1 **is_a** A_3), *reflexive* (A_1 **is_a** A_1), and *antisymmetric* (if A_1 **is_a** A_2 and A_2 **is_a** A_1, then A_1 and A_2 are identical); it is not symmetric. Both, the parthood relation between individual instances and the one between concepts are transitive, reflexive, and antisymmetric as well. Antisymmetry implies a direction and can be modeled as a directed graph (Fig. 2a, b) in which the nodes and edges represent parent–child relationships (for symmetry see Fig. 2c). The **instance_of** property, on the other hand, is neither transitive, nor reflexive, nor symmetric, nor antisymmetric.

Besides transitivity, reflexivity, and symmetry or antisymmetry, further logical properties can be specified for every RDF property. Regarding a specific RDF property one can define the concepts that are allowed to be used as *Subject* (i.e. the domain of the property) and as *Object* (i.e. the range of the property) in a RDF triple in connection with this property. The specification of the domain and the range of a RDF property constraints its applicability with the intention to minimize logical inconsistencies within sets of RDF triples. The property **actively_participates_in**, for instance, specifies a material object that participates in a process. Thus, the domain of **actively_participates_in** has to be restricted to material objects and its to range processes.

Defining concepts in an ontology

The definition of a term/concept takes in an important function in every ontology. One can distinguish two types of terms or concepts. On the one hand primitive concepts that have only necessary conditions, and on the other hand defined concepts that have necessary and sufficient conditions (Stevens et al. 2000). Regarding explicitness and clearness one should always prefer defined over primitive concepts when developing an ontology.

Necessary conditions are specified via the class–subclass relationship **is_a**. The specification of class–subclass relationships represents an essential part of what is commonly referred to as Aristotelian definitions and includes, due to downward propagation which is inherent to the **is_a** property, all defining properties of all of its parent classes. An Aristotelian definition consists of two sets of defining properties (Fig. 3): genus, necessary for assigning an entity to a class (properties inherited from the parent class), and differentiae, necessary for distinguishing the subclass (child) from other subclasses assigned to the same class (e.g., Smith 2004; Rosse and Mejino 2003). By connecting a class to another class through the class–subclass relationship **is_a** and restricting it through additional properties that are specifically required for the subclass, different subclasses can be distinguished and defined. An individual object necessarily has to meet the requirements of the class to qualify as an instance of its subclass. This allows for a convenient way of defining a new term by relating it through the class–subclass relationship to an already defined term (necessary condition) and than adding its distinguishing property (in combination resulting in the sufficient condition). The human kidney, for instance, can be defined as: *Polarized_junctioned_cell* **is_a** *Junctioned_cell* AND *Polarized_junctioned_cell* **orientation** *Apico-basal*. Here '**is_a** *Junctioned_cell*' represents the *genus* part of the definition and '**orientation** *Apico-basal*' the *differentia* part (Fig. 4).

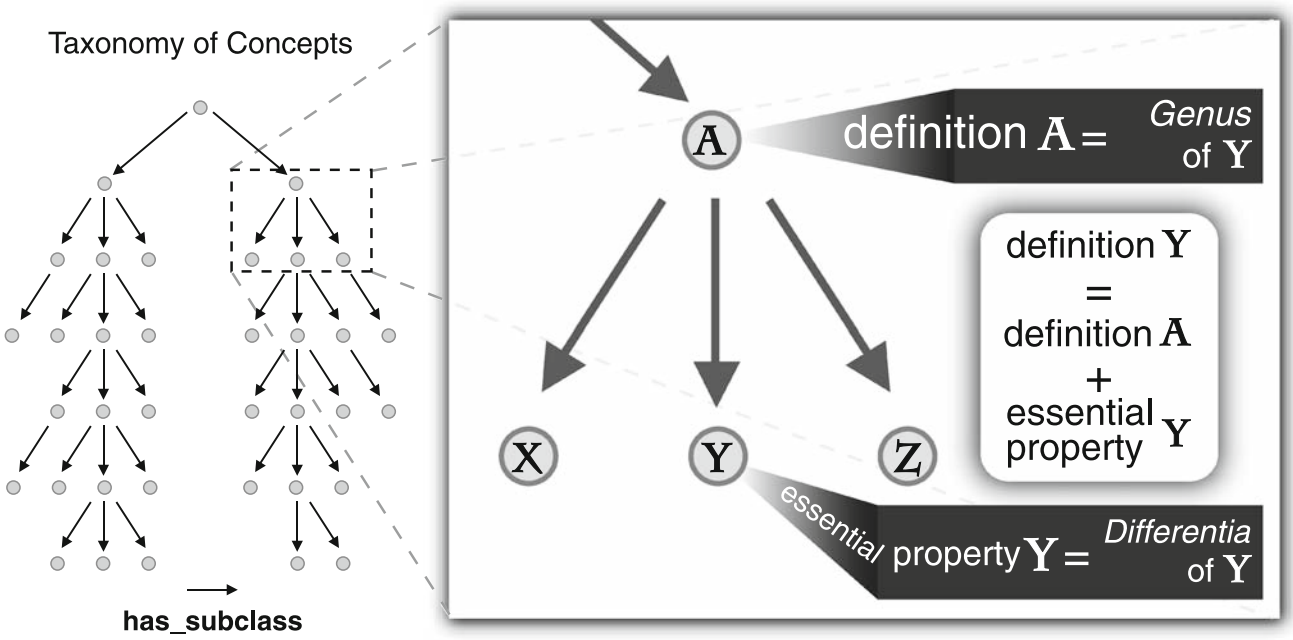

Fig. 3 Aristotelian definition: the set of defining properties (i.e. essence) of a 'parent' class are passed on to all its 'child' classes (downward propagation) and provides the *Genus* part of their definitions. All instance of every 'child' class *necessarily* has to posses all of these properties. The 'child' classes are distinguished from one another by their respective essential properties, which provide the *Differentia* part of their definitions. Only the combination of *Differentia* and *Genus* is *sufficient* for membership to a class and represents the essence of a kind, and thus the definition of the respective class

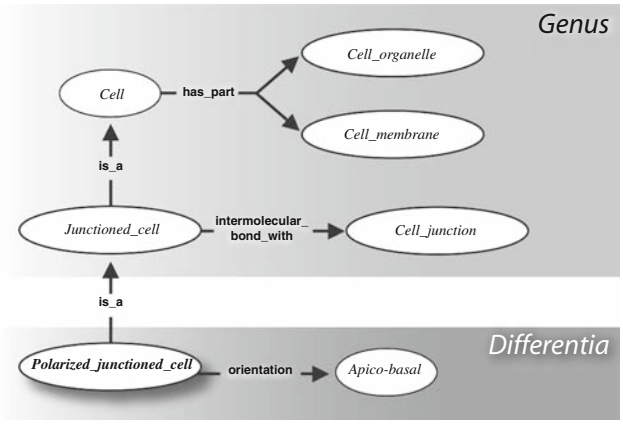

Fig. 4 RDF graph of a definition of 'polarized junctioned cell': a 'polarized junctioned cell' can be defined as a 'junctioned cell' that possesses an apico-basal orientation. In this case, all defining properties of 'junctioned cell' as well as those of all of its parent classes would provide the genus part of the definition of 'polarized junctioned cell', while the property of possessing an apico-basal orientation would represent the differentia part. Understood that way, 'polarized junctioned cell' would represent a special kind of 'junctioned cell'

Zoo-morphology

The linguistic problem of morphology

Lately, the representation of phenotypic information represents an area of intense discussion (see Blake 2004). Especially in the context of inquiring causal/functional interlinks between objects belonging to different levels of organization, such as gene products and morphological structures, the demand of high quality phenotypic data increases constantly. But also for all kinds of comparative studies over a large taxonomic range, phenotypic data with high degree of comparability is highly requested.

Morphological data constitutes a very large proportion of phenotypic data. In morphology, only descriptions of the structures or the functions of morphological features represent data in the strict sense (see Vogt et al. 2008; Vogt 2008). Compared to molecular data, morphological descriptions seem to be inferior with respect to transparency and reproducibility of their production. The apparent weakness of morphological characters has been addressed in a series of papers by Wiens and Jenner within the framework of phylogenetic methodology (e.g., Wiens 1995; Hillis and Wiens 2000; Poe and Wiens 2000; Wiens 2001, 2004; Jenner 2002, 2004a, 2004b, 2004c; see also Pimentel and Riggins 1987; Stevens 1991; Thiele 1993; Kesner 1994;). According to these authors, morphological character matrices build by one phylogeneticist are often treated like a black box by other phylogeneticists, and it becomes a question of authority rather than scientific argumentation whether one should rely on a phylogenetic character statement of another phylogeneticist.

The problems of phylogenetic morphological characters addressed by Wiens and by Jenner within a phylogenetic

framework result from a more fundamental and general problem concerning morphological descriptions, the linguistic problem of morphology (Vogt et al. 2008). The linguistic problem of morphology results from the lack of a standardized and commonly accepted morphological terminology, the lack of a rational for the delimitation of morphological traits, and a standardized and formalized methodology of morphological description. As a consequence, morphological terminology and morphological descriptions vary from author to author, the meaning of a term often changes through time, and morphological terminology is often restricted to a specific taxonomic group and cannot be easily adapted to other groups. This non-standardization and therefore diverse usage of morphological terminology and the lack of a commonly accepted method of description can, in scientific practice, lead to identically described structures that are in fact not identical or divergent descriptions of one and the same morphological structure (Vogt et al. 2008; see also Ramírez et al. 2007).

The linguistic ambiguities in morphological descriptions turned out to pose fundamental problems for comparative morphological studies, since it is the source for repeated misunderstandings among morphologists, undermining the possibility to reliably communicate morphological data. Reliable communication of data, however, represents a necessary prerequisite for division of labor not only among morphologists conducting comparative studies over a broad taxonomic range, but also for all kinds of co-operations and studies in which morphologists are involved or morphological data is analyzed.

Another problem, which is somehow related to the linguistic problem of morphology, is often found in comparative studies like for instance phylogenetics. In phylogenetics morphological descriptions are often not clearly separated from hypothetical conclusions and explanatory hypotheses based on them, with the consequence that observational data and hypotheses blend and cannot be differentiated anymore. This happens for instance whenever morphological descriptions use terms that imply homology of the described structures (Vogt 2008; Vogt et al. 2008). Homology statements represent hypotheses that transcend the perceptually given by providing a possible explanation for the sameness of structures through the implicit assumption of a common evolutionary origin and should not be mistaken with morphological descriptions, which are by nature descriptive and not explanatory.

RDF ontology provides a promising solution to the linguistic problem

Hitherto, no traditional dictionary or thesaurus for morphological terminology solved the problems resulting from the linguistic problem. Apart from the fact that no taxon-independent dictionary/thesaurus is available, this may be due to the fact that different traditional dictionaries and thesauri propose different terms and concepts and that they often use terms that imply homology. In any case, the most important reason for their failure is that they are not commonly accepted and therefore not used by a majority of morphologists.

Instead of the traditional dictionaries and thesauri, the combination of a data base for morphological/phenotypic data and the use of a taxon-independent RDF ontology for morphology could have the potential to provide a regional solution to the linguistic problem by explicitly imposing a defined and controlled vocabulary (Vogt et al. 2008). This combination would provide an integrative platform—although restricted to the particular data base—within which comparative morphological/phenotypic studies would be possible. Within life sciences such a morphological data base could take in a central methodological function comparable to GenBank (http://www.ncbi.nlm.nih.gov/Genbank) for molecular data.

The necessity for a taxon-independent data standard

So far, a lot of effort and resources went into the development of bio-ontologies with a molecular focus or with a focus on modeling knowledge for a specific model organism (see above). Some of these ontologies reached a level of sophistication that, by now, allows their successful application within respective data bases (e.g., FlyBase, Mouse Genome Database, *Arabidopsis* Information Resource, WormBase, etc.). On the other hand, data bases that are designed for the management of comparative data over a large taxonomic range are still rare. This is particularly true for morphological data. For the use of data in comparative studies across a broad taxonomic range a taxon-independent data standard would be required. Thus, comparative biology requires ontologies to be as much as possible taxon-independent. Taxon-independence in this context means that the applicability of terms and concepts referring to specific types of structures should not be restricted to a particular taxonomic group—the definition of terms should be free of conditional predicates such as 'insect'-head or 'arthropod'-segmentation, but should provide taxon-independent criteria for their application—even though some terms might in practice only be applied to a certain taxonomic group.

A general structure concept for morphology

The development of a taxon-independent anatomy ontology that does not rest on homology assumptions presupposes a general structure concept for morphology (see

Vogt 2008). In this context, structure should be understood to represent a way to conceive properties and relations of a complex whole: the set of relations between different parts and aspects of a given complex whole determines its structure. A structure concept is required for the conceptualization of a complex whole, and should facilitate in generating data of a specific type and quality that are relevant for a specific scientific discipline or research program. Every structure concept is thus necessarily context-dependent (Vogt 2008). As such, the role of a structure concept regarding data in the strict sense (data s.str.) is comparable to the role of minimum information checklists regarding metadata.

On a very basic level, morphological data s.str., like many other types of data s.str. in natural sciences, represent existence statements. In morphology they are called morphological descriptions and they represent hypotheses about the existence of entities and their properties, which are based on observational judgments—just like answers to questions about the entities properties and relations. A general morphological structure concept (like any other structure concept) should provide a scheme for standardizing and formalizing such descriptions by providing a list of relevant perceptual categories (Fig. 5)—non-redundant questions to be answered on the basis of morphological investigations. Morphological data would than consist of a list of answers, which can be understood to represent the values to each category that has been evaluated in a morphological investigation. Ideally, the structure concept also specifies which answers (i.e., values) are possible, thereby restricting the possible value-space (Vogt 2008).

A general structure concept for morphology would thus have to consist of a standardized set of relevant empirical categories (questions like: what is directly adjacent to the entity to be described; is it continuous with some other entity; out of what parts is it composed; does it actively participate in a specific type of biological process) together with their respective value-spaces (to each question a standardized set of possible answers). A value-space can be either a Boolean YES or NO, numerical values like for instance the natural numbers or a specific numeral interval, or a limited set of defined terms. It is obvious that a morphological ontology presupposes the development of a general structure concept for morphology, and first steps in this direction have already been taken (see Vogt 2008).

The future of zoo-morphology

Until the late nineteenth century, morphology was considered to be the most important and prolific scientific field of what would later become 'biology' (e.g., Nyhart 1995).

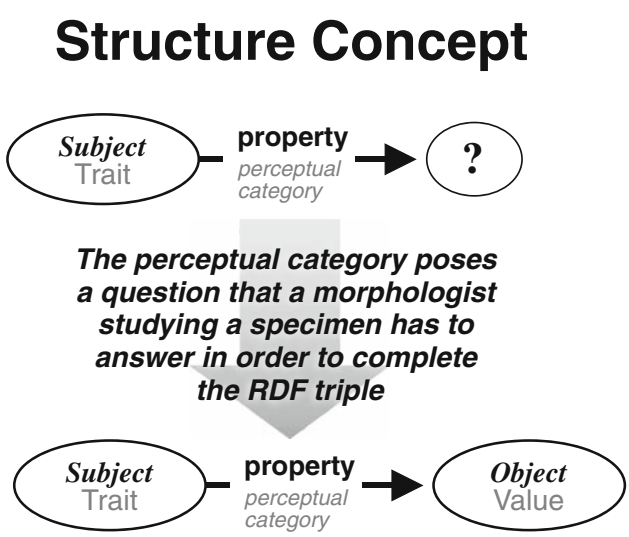

Fig. 5 Using a structure concept in RDF statements: The entity to be described (e.g., a morphological trait) is represented by the '*Subject*'. The '**property**' corresponds with one perceptual category taken from the structure concept and functions as an empirical question that can only be answered by investigating the entity to be described. The answer to such a question can be represented by the '*Object*' of the RDF triple. It corresponds with one of the 'values' that the structure concept allows as a possible value ('value-space') for this particular category

However, with the advent of physiology as an independent scientific discipline and its rapid development in the twentieth century, and not least with the rise of genetics, morphology more and more lost its prominent role.

The difference in data quality in terms of aperspectival and mechanical objectivity might be the reason why some biologists question the relevance of organismic biology in general and morphology in particular with respect to their potential to significantly contribute to future developments in biology. I am convinced that the linguistic problem of morphology accounts for a significant proportion of the popularity of this opinion. Apparently, if morphology manages to come up with a solution to the linguistic problem, it is likely that it will continue to make valuable contributions to important future developments in biology. In doing so, it may actually become indispensable and could take in an essential role, since it has the potential to provide high-quality phenotypic data that is highly demanded for various interesting biological research programs.

Standardized morphological data would not only improve comparability of morphological data. It would also facilitate many co-operations of morphologists with for instance developmental biologists or geneticists. Very interesting research programs concerning for instance mechanismic interrelationships of morphogenesis and underlying gene organization could benefit from a morphological data standard. The development of a RDF ontology for morphology represents a first and important step towards that direction.

A general data standard in biology

Standardizing bio-ontologies

The general use of RDF ontologies within biological data bases facilitates interoperability across different data bases and provides the formal basis for the final goal of establishing a general and unified data standard for all types of biological data. A unified data standard, at its turn, would provide the formal basis for an integrative approach to biology (Bard and Rhee 2004), enabling communication and integrated comparison of all kinds of biological data. The establishment of commonly accepted general criteria for the development of well-structured and standardized ontologies is essential for ontologies to be successful in this respect.

Unfortunately, although all well constructed ontologies are highly formalized, different ontologies often differ quite substantially in their use of properties and their hierarchy (i.e. taxonomy) of concepts. Practical demands of researchers for exchanging and integrating data require a coordinated development of shared ontologies. The Open Biological Ontologies initiative (OBO; http://obo.sourceforge.net) provides a virtual umbrella organization for increasing transparency of ontology development and coordinating standardization efforts for ontologies for shared use across different biological and medical domains. At the OBO website one can find various bio-ontologies ranging from anatomy to development, genomics and proteomics, experimental conditions, metabolomics, phenotype, and taxonomic classification. OBO specifies criteria that have to be met by an ontology in order to be accepted as one of the OBO (obo.sourceforge.net/crit.html). These criteria guarantee for instance (a) that all their ontologies are for sharing and are resources for the entire community, (b) that tools, as for instance query machines, can be usefully applied to all their ontologies, facilitating shared software implementations, and (c) the possibility of two different ontologies (e.g., an anatomy and a process ontology) to be combined through additional relationships.

OBO is showing the way of how to impose something like a seal of approval for ontologies that meet certain criteria. Thus, the basis for the establishment of a general data standard in biology does already exist, and under the OBO umbrella it is actually already under development. Yet, at the current stage of development, the data standard is still very unspecific and cannot utilize all its potential. However, at present nobody can predict what exactly the final schema of a general standard for biological data will be like, other than that it will most likely involve RDF ontologies.

What is a data standard?

All standardization activities have to deal with many methodological as well as practical problems and any criteria or sets of questions that could guide this process are welcome. In analogy to manufacturing standards, Brazma (2001) suggests that the huge amounts of information produced by high throughput experiments have to be transformed into executive summaries, and he identifies three criteria for the generation of such summaries.

First, a summary should include all elements essential for understanding the studied phenomena, as for instance all relevant metadata about the studied biological samples (e.g., species determination, sampling coordinates, etc.) and some quality indicators. This is the objective of minimum information checklists. However, all minimum information checklists are restricted to metadata and thus only represent the metadata part of a content standard (for the data s.str. part of a content standard see structure concept).

A data base that has an ontology implemented and that uses it for data representation and documentation provides its data in a computer parsable form. This corresponds with Brazma's (2001) second criterion that the information should be available in a way that it can be parsed by a computer program. Since RDF statements can be modeled as directed graphs and since RDF allows the application of description logics, RDF ontologies provide a broad basis for all kinds of computer applications, including very powerful query engines that enable a researcher to find most specific peaces of information in very large data sets, without having the scientist to go through all available data herself. One could, for instance, pick one morphological description from the data base and let the data base search for most similar descriptions, or search for all structures described as adjacent to heart (i.e. all RDF triples with the property **adjacent_to** and the object *Heart*).

The use of RDF ontologies within data bases results in a significant increase not only of the transparency and reproducibility of the generation of data, but also the degree of comparability and communicability of biological data in general. Thus, RDF ontologies have the potential to significantly increase aperspectival and mechanical objectivity of biological data. Understanding aperspectival and mechanical objectivity as an index for data quality, RDF ontologies can establish a quality standard for biological data, which represents Brazma's (2001) third criterion.

Wang et al. (2005) suggest that before developing a data standard one has to answer two questions. The first question concerns the content: what should be standardizes? This question addresses the requirement of a content standard and thus corresponds with Brazma's first criterion, except that it also includes the data s.str. part of a content standard.

The second question of Wang et al. (2005) concerns the methodology: how should the standardization be formatted? This corresponds with Brazma's second criterion of computer parsability. Agreeing upon a common format standard is essential for increasing aperspectival and mechanical objectivity, and RDF provides such a format standard. While the content standard specifies, which type of information is required, the format standard provides in form of a standardized file format the specific syntax for transmitting and communicating data (Field and Sansone 2006).

Anyhow, a very important distinction that one should always keep in mind while developing data standards is much more basic: the establishment of a data standard not only requires a detailed documentation of the methods and techniques applied for data production (i.e., metadata standard), but also the development of clear and transparent definitions of scientific terminology for data representation (i.e., data s.str. standard). Thus, the abovementioned criteria and types of standardizations apply to both data and metadata. Most current standardization activities, especially within the OMICS fields of research, however, only focus on standardizing metadata. This might be due to the fact that what belongs to data s.str. and how to unambiguously represent it is less problematic in the OMICS fields of research, which already established standards for their data s.str. (e.g., DNA sequence data), than it is for instance in morphology, where it represents the main challenge (see Vogt et al. 2008). Only in case one manages to standardize both, the expected increase in data quality will be reached.

Moreover, considering the terminology problems of lack of standards of gene names and spellings mentioned above, it is obvious that agreeing upon a content standard and a format standard is not sufficient—biological data requires further standardization in order to meet the demands for communicability and comparability of data. In order to avoid terminological problems, concepts standards and nomenclatural standards are required in addition to content and format standards. The implementation of ontologies within data bases has proven to be very successful in that respect.

Do general standardization criteria exist that apply to all biological disciplines?

Developing a data standard actually requires the development of a whole set of different standards: (a) what represents relevant information (i.e. content standard), (b) what does a concept mean (i.e., concept standard), (c) which label (i.e., word, symbol, acronym) should we use to refer to this concept (i.e., nomenclatural standard), and (d) how is the syntax with which data should be transmitted and communicated.

Content standard

The content of what should be standardized has to be defined by each ontology and data base separately, depending on its scope and purpose. In case of metadata, content standards are determined by minimum information checklists, which specify what types of annotations are required for establishing a high degree of transparency and reproducibility of data production. For data s.str., on other hand, content standards are determined by structure concepts, which specify what types of perceptual categories together with their corresponding value-spaces should be used for describing data s.str., in order to establish a high degree of comparability of data.

Both, minimum information checklists as well as structure concepts are research and data type specific. Depending on a given research program or theoretical framework, they specify which empirical phenomena and procedural information is relevant and which should be ignored. Obviously, a general content standard that is universally applicable to all biological disciplines cannot exist. However, whenever information is relevant in several biological disciplines, as for instance specimen information, the same content standard should be applied, thereby generalizing standardization across various biological disciplines and avoiding duplication of efforts (Field and Sansone 2006). Some standards, as for instance inferential standards, entail other standards, as for instance standards for proper experimentation. Standards thus can have a hierarchical nature with some 'higher level' standards entailing 'lower level' standards. Projects already exist that provide support for integrating various standardization efforts, as for instance the Reporting Structures for Biological Investigations (RSBI) project (Sansone et al. 2006).

Concept standard

Concept standards provide the meaning to scientific terms used for metadata and data s.str. They are essential, as they solve terminological problems. Ontologies provide concept standards.

Although ontologies for data s.str. are based on specific structure concepts that are necessarily research specific and vary with data type and with the theoretical background of an investigation, they can still be based on a general taxonomy of foundational concepts, which can be commonly used in order to develop more specific ontologies. This applies particularly to the **properties** used for defining concepts. If a given **property**, as for instance **actively_participates_in**, has the same meaning throughout all bio-ontologies, we would have a standardized terminology for defining biological concepts within bio-ontologies. The necessity for stan-

dardizing **properties** has been recognized and efforts exist for developing a standard set of foundational properties (see, e.g., obo relationship types ontology; Smith et al. 2005).

The use of RDF ontologies in data bases for the purpose of data management also improves possibilities for the transfer or mapping of data across different data bases, which increases aperspectival objectivity between the contents of these data bases.

Nomenclatural standard

The nomenclatural standard provides a stable link between a term and its corresponding concept. Ontologies provide nomenclatural standards. By providing local identifiers for the concepts defined in an ontology, each ontology establishes a nomenclatural standard that is restricted to the data bases in which it is implemented. However, what is required is a nomenclatural standardization across different ontologies and different data bases. To provide a stable and resolvable identifier for concepts across ontologies and data bases it would be highly desirable to adopt globally unique identifiers (GUIDs) or life science identifiers (LSIDs) for all biological concepts defined within bio-ontologies.

Format standard

The format standard provides a standardized file format, which specifies the syntax for transmitting and communicating metadata and data s.str. The current trend is to use XML based formats, like the resource description framework (RDF).

The advantage of XML based syntax standards is that they provide a formalized standard that is computer readable but still allow a lot of freedom regarding the content that the syntax organizes. This is especially important considering that all standardization activities in biological research necessarily represent an ongoing process (Brooksbank and Quackenbush 2006). What seems to be important today may be less important tomorrow and new technologies and theoretical insights may bring about new types of data, which requires the development of new concepts and properties for existing ontologies or even new ontologies. Thus, XML based syntax standards represent a good answer to Wang et al's (2005) second question concerning data format.

Conclusion

Looking into the history of biology one can observe the continuous differentiation and subsequent methodological divergence of different biological disciplines, with organismic and molecular biology forming the most divergent poles. Besides all the obvious advantages of a diversification within biological research, this process also holds its drawbacks, because it results in a diversification of scientific languages, to the effect that communication between the various biological research communities becomes more and more difficult. But only co-operation involving organismic and molecular biologists will maximize output in biological sciences and will do justice to the biological diversity and organizational integration that is caused by evolution. In that context focusing only on model organisms is problematic, since model organism studies usually generate knowledge with high explanatory force, but this knowledge has a limited explanatory range since it only refers to a very limited portion of biological diversity (for 'explanatory force' and 'explanatory range' see Platts 1997; Rieppel 2005). Thus, in order to give good estimates regarding larger portions of biological diversity, comparative approaches are required, which at their turn not only depend on the knowledge gained by model organism studies but also require data that is highly comparable across various species and taxa.

The need for standardization has been recognized in many fields of biological research by now and all respective activities are gaining credibility, in particular in the light of emerging policies on open access and data sharing (NAS 2003; NIH 2006; OECD 2003; Wellcome 2003; Field and Sansone 2006). Unfortunately, many proposals for developing data standards still do not well in traditional peer-review settings, which mostly favor hypothesis-driven research (Brooksbank and Quackenbush 2006). However, funding agencies seem to start to realize that the expenses that are generated by incorrect data interpretations and analyses due to lack of data standards by far outweigh the costs for developing standards—thus, also from a purely financial point of view, investing into standardization activities makes good sense. Moreover, a common biological data standard would also provide instrument manufacturers and lab suppliers with common criteria for their products, reducing time and cost for implementing standard-compliance (Field and Sansone 2006).

To be successful as a data standard, RDF ontologies must gain widespread acceptance in their respective biological communities. Only if ontologies can overcome the resistance of the major generators of the corresponding data so that they are willing to use the ontologies, they will be a success. While this has been recognized in the OMICS fields of biological research, morphologists seem to be especially reluctant (which might be due to the fact that the OMICS communities have already been more integrated to begin with than for instance morphologists).

Further development of RDF ontologies is inevitable, also because the amount of data in life sciences will keep

increasing exponentially in future. Ontology related tools will enable easy sifting through the data bulks and the integration of data from different sources. Data bases in combination with description logics and underlying bio-ontologies will provide the basis for the development of such tools. This development provides a great opportunity for morphologists to disseminate their data through various biological communities in a way that they are actually usable for non morphologists. And when more morphological data are used by other biologists, not only new knowledge and insights will be gained, but also more opportunities for co-operations and maybe even more funding for morphological research. Thus, the entire field of morphological research could benefit from a morphological data standard. Whether morphology will be part of this ongoing process of standardization and integration will depend on morphologists managing to develop and agree on a common data standard.

Acknowledgments I thank Thomas Bartolomaeus and Markus Koch for discussion and valuable suggestions for an earlier draft of this paper. I also want to thank Ronald Jenner, as well as two other anonymous referees, for reading, criticizing, and commenting on an early manuscript. It goes without saying, however, that I am solely responsible for all the arguments and statements in this paper. This study was supported by the SPP 1174 of the German Research Foundation (DFG) (VO 1244/3-2). I am also grateful to the taxpayers of Germany.

References

Bard J (2003) Ontologies: formalising biological knowledge for bioinformatics. Bioessays 25:501–506. doi:10.1002/bies.10260

Bard J, Rhee SY (2004) Ontologies in biology: design, applications and future challenges. nature reviews genetics 5:213–222

Beckett D (2004) RDF/XML syntax specification (revised). W3C recommendation published online 10 February 2004. http://www.w3.org/TR/2004/REC-rdf-syntax-grammar-20040210/

Berners-Lee T (2005) Primer: Getting into RDF & Semantic Web using N3. Published online 29 June 2005. http://www.w3.org/2000/10/swap/Primer.html

Beurton PJ, Falk R, Rheinberger H-J (2000) The concept of the gene in development and evolution. Cambridge University Press, Cambridge

Bisby FA, Shimura J, Ruggiero M, Edwards J, Haeuser C (2002) Taxonomy, at the click of a mouse—informatics and taxonomy are working together to achieve more than either could alone. Nature 418:367. doi:10.1038/418367a

Bittner T (2004) Axioms for parthood and containment relations in bio-ontologies. In: KR-MED 2004 workshop on formal biomedical knowledge representation. University of Aachen, Aachen, pp 4–11

Bittner T, Donnelly M, Smith B (2004) Individuals, universals, collections: on the foundational relations of ontology. In: Varzi A, Vieu L (eds) Proceedings of the international conference on formal ontology in information systems. IOS Press, Amsterdam, pp 37–48

Blake J (2004) Bio-ontologies—fast and furious. Nat Biotechnol 22:773–774. doi:10.1038/nbt0604-773

Brazma A (2001) On the importance of standardisation in life sciences. Bioinformatics 17:113–114. doi:10.1093/bioinformatics/17.2.113

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FCP, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M (2001) Minimum information about a microarray experiment (MIAME)–towards standards for microarray data. Nat Genet 29:365–371. doi:10.1038/ng1201-365

Brickley D (2004) RDF Vocabulary description language 1.0: RDF Schema. W3C recommendation published online 10 February 2004. http://www.w3.org/TRF/rdf-schema/

Brooksbank C, Quackenbush J (2006) Data standards: a call to action. OMICS 10(2):94–99

Daston L (1992) Objectivity and the escape from perspective. Soc Stud Sci 22:597–618. doi:10.1177/030631292022004002

Daston L (1998) Fear and loathing of the imagination in science. Daedalus 127:16–30

Daston L, Galison P (1992) The image of objectivity. Representations (Berkeley) 40:81–128. doi:10.1525/rep.1992.40.1.99p0137h

Field D, Sansone S-A (2006) Foreword: a special issue on data standards. OMICS 10(2):84–93

Consortium Gene Ontology (2006) The gene ontology (GO) project in 2006. Nucleic Acids Res 34:D322–D326. doi:10.1093/nar/gkj021

Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, Korbel JO, Emanuelsson O, Zhang ZD, Weissman S, Snyder M (2007) That is a gene, post-ENCODE? History and updated definition. Genome Res 17:669–681. doi:10.1101/gr.6339607

Gewin V (2002) All living things, online. Nature 418:362–363. doi:10.1038/418362a

Godfray HCJ (2002) Challenges for taxonomy. Natur 417:17–19. doi:10.1038/417017a

Grenon P, Smith B, Goldberg L (2004) Biodynamic ontologies: applying BFO in the biomedical domain. In: Pisanelli DM (ed) Ontologies in medicine. IOS Press, Amsterdam, pp 20–38

Griffiths PE, Stotz K (2007) Gene. In: Hull DL, Ruse M (eds) The Cambridge companion of the philosophy of biology. Cambridge University Press, Cambridge

Gruber T (1993) A translation approach to portable ontologies. Knowl Acquis 5:199–220. doi:10.1006/knac.1993.1008

Heintz B (2000) Die Innenwelt der Mathematik. Zur Kultur und Praxis einer beweisenden Disziplin. Springer, Wien

Hillis DM, Wiens JJ (2000) Molecules versus morphology in systematics—conflicts, artifacts, and misconceptions. In: Wiens JJ (ed) Phylogenetic analysis of morpological data. Smithsonian Institute Press, Washington DC, pp 1–19

Jenner RA (2002) Boolean logic and character state identity: pitfalls of character coding in metazoan cladistics. Contrib Zool 71(3):67–91

Jenner RA (2004a) The scientific status of metazoan cladistics: why current research practice must change. Zool Scr 33:293–310. doi:10.1111/j.0300-3256.2004.00153.x

Jenner RA (2004b) When molecules and morphology clash: reconciling conflicting phylogenies of the Metazoa by considering secondary character loss. Evol Dev 6(5):372–378. doi:10.1111/j.1525-142X.2004.04045.x

Jenner RA (2004c) Accepting partnership by submission? Morphological phylogenetics in a molecular millennium. Syst Biol 53(2):333–342. doi:10.1080/10635150490423962

Kennedy J, Hyam R, Kukla R, Paterson T (2006) Standard data representation for taxonomic information. OMICS 10(2):220–230

Kesner MH (1994) The impact of morphological variations on a cladistic hypothesis with an example from a mycological data set. Syst Biol 43:41–57. doi:10.2307/2413580

Kukla R (2006) Objectivity and perspective in empirical knowledge. Episteme J Soc Epistem 3:80–95

Mahner M, Bunge M (1997) Foundations of biophilosophy. Springer, Berlin

Manola F, Miller E (2004) RDF Primer. W3C Recommendations published online 10 February 2004. http://w3.org/TR/rdf-primer/

McGuinness DL, van Harmelen F (2004) OWL Web ontology language overview. W3C recommendation published online 10 February 2004. http://www.w3.org/TR/owl-features/

NAS (2003) NAS Committee on responsibilities of authorship in the biological sciences. Sharing publication-related data and materials. Available at: www.nap.edu/books/0309088593/html/

NIH (2006) NIH roadmap for bioinformatics and computational biology. Available at: http://nihroadmap.nih.gov/bioinformatics/index.asp

Nyhart LK (1995) Biology takes form. The University of Chicago Press, Chicago

OECD (2003) OECD Group on issues of access to publicly funded research data. Promoting access to public research data for scientific, economic, and social development. Available at: http://dataaccess.ucsd.edu/Final_Report_2003.pdf

Patterson DJ, Remsen D, Marino WA, Norton C (2006) Taxonomic indexing—extending the role of taxonomy. Syst Biol 55:367–373. doi:10.1080/10635150500541680

Pennisi E (2003) Modernizing the tree of life. Science 300:1692–1697. doi:10.1126/science.300.5626.1692

Pimentel R, Riggins R (1987) The nature of cladistic data. Cladistics 3:201–209. doi:10.1111/j.1096-0031.1987.tb00508.x

Platts M (1997) Ways of Meaning. An introduction to the philosophy of language, second ed. MIT Press, Cambridge

Poe S, Wiens JJ (2000) Character selection and the methodology of morphological phylogenetics. In: Wiens JJ (ed) Phylogenetic analysis of morpological data. Smithsonian Institute Press, Washington DC, pp 20–36

Polyani M (1968) Life's irreducible structure. In: Grene M (ed) Knowing and being (1969): essays by Michael Polyani. Routledge & Kegan Paul Ltd, London chapter 14

Prohaska SJ, Stadler PF (2008) Genes. Theory Biosci 127:215–221. doi:10.1007/s12064-008-0025-0

Ramírez MJ, Coddington JA, Maddison WP, Midford PE, Prendini L, Miller J, Griswold CE, Hormiga G, Sierwald P, Scharff N, Benjamin SP, Wheeler WC (2007) Linking of digital images to phylogenetic data matrices using a morphological ontology. Syst Biol 56(2):283–294. doi:10.1080/10635150701313848

Riedl R (2000) Strukturen der Komplexität–Eine Morphologie des Erkennens und Erklärens. Springer, Berlin

Rieppel O (2005) Proper names in twin worlds: monophyly, paraphyly, and the world around us. Org Divers Evol 5(2):89–100. doi:10.1016/j.ode.2004.03.003

Rosse C, Mejino JLV Jr (2003) A reference ontology for biomedical informatics: the Foundational Model of Anatomy. J Biomed Inform 36:478–500. doi:10.1016/j.jbi.2003.11.007

Salthe SN (1985) Evolving hierarchical systems: their structure and representation. Columbia University, New York

Salthe SN (1993) Development and Evolution: complexity and change in biology. MIT Press, Cambridge

Sansone S-A, Rocca-Serra P, Tong W, Fostel J, Morrison N, Jones AR, RSBI Members (2006) A strategy capitalizing on synergies: the reporting structure for biological investigation (RSBI) working group. OMICS 10(2):164–171

Scherrer K, Jost J (2007) The gene and the genon concept: a functional and information-theoretic analysis. Mol Syst Biol 3:87. doi:10.1038/msb4100123

Schutt CE, Lindberg U (2000) The new architectonics: an invitation to structural biology. Anat Rec New Anat 261:198–216

Smith B (2004) The logic of biological classification and the foundations of biomedical ontology. In: Westerståhl D (ed) Invited papers from the 10th International Conference in Logic Methodology and Philosophy of Science, Oviedo, Spain, 2003. Elsevier, North Holland

Smith B, Ceusters W, Klagges B, Köhler J, Kumar A, Lomax J, Mungall C, Neuhaus F, Rector A, Rosse C (2005) Relations in biomedical ontologies. Genome Biol 6. doi:10.1186/gb-2005-6-5-r46

Smith B, Rosse C (2004) The role of foundational relations in the alignment of biomedical ontologies. Proc Med info 2004:444–448. IOS Press, Amsterdam

Stein LD (2003) Integrating biological databases. nature reviews genetics 4:337–345

Stevens PF (1991) Character states, morphological variation, and phylogenetic analysis: a review. Syst Bot 16:553–583. doi:10.2307/2419343

Stevens R, Goble CA, Bechhofer S (2000) Ontology-based knowledge representation for bioinformatics. Brief Bioinform 1:398–414. doi:10.1093/bib/1.4.398

Thiele K (1993) The holy grail of the perfect character: the cladistic treatment of morphometric data. Cladistics 9:275–304. doi:10.1111/j.1096-0031.1993.tb00226.x

Trewavas A (2006) A brief history on systems biology. Plant Cell 18:2420–2430. doi:10.1105/tpc.106.042267

Valentine JW, May CL (1996) Hierarchies in biology and paleontology. Paleobiology 22(1):23–33

Vogt L (2008) Learning from Linnaeus. Towards developing the foundation for a general structure concept for morphology. Zootaxa 1950:123–152

Vogt L, Bartolomaeus T, Giribet G (2008) The linguistic problem of morphology—a challenge to the future role of morphology in life sciences (submitted)

Wang X, Gorlitsky R, Almeida JS (2005) From XML to RDF: how semantic web technologies will change the design of 'omic' standards. Nat Biotechnol 23:1099–1103. doi:10.1038/nbt1139

Wellcome (2003) Wellcome trust. Sharing data from large-scale biological research projects: a system of tripartite responsibility. Available at: http://www.genome.gov/Pages/Research/WellcomeReport0303.pdf

Wiens JJ (1995) Polymorphic characters in pylogenetic systematics. Syst Biol 44:482–500. doi:10.2307/2413656

Wiens JJ (2001) Character analysis in morphological phylogenetics: problems and solutions. Syst Biol 50:689–699. doi:10.1080/106351501753328811

Wiens JJ (2004) The role of morphological data in phylogenetic reconstruction. Syst Biol 53:653–661. doi:10.1080/10635150490472959

Wilson EO (2003) The encyclopedia of life. Trends Ecol Evol 18:77–80. doi:10.1016/S0169-5347(02)00040-X

Wilson RA (2005) Genes and the agents of life: the individual in the fragile sciences: Biology. Cambridge University Press, New York, NY

Wimsatt WC (1976) Reductionism, levels of organization, and the mind-body problem. In: Globus G, Maxwell G, Savodnik I (eds) Consciousness and the brain: a scientific and philosophical inquiry. Plenum Press, New York, pp 202–267

Wimsatt WC (1994) The ontology of complex systems: Levels, perspectives, and causal thickets. Can J Philos 20(supplemental):207–274