

Minimal Anatomy Terminology (MAT): a species-independent terminology for anatomical mapping and retrieval

Jonathan B.L. Bard¹, James Malone², Tim F. Rayner² and Helen Parkinson²

1. Computational Biology Research Group, Weatherall Institute for Molecular Medicine, John Radcliffe Hospital, Oxford OX3 9DS, UK

2. EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

ABSTRACT

Motivation: The problem of integrating a multiplicity of non-orthogonal anatomy ontologies is well known in ontology development. There are now major public ontology repositories (e.g. the Ontology for Biomedical Ontologies) that require a multi-species anatomy ontology. We present MAT (Minimal Anatomy Terminology) an OBO format terminology (~400 terms) using SKOS *broader-than* relationships designed for annotating and searching tissue-associated data and timelines for any organism. Identifiers from >20 anatomy ontologies are mapped to each MAT term to facilitate access to and interoperability across tissue-associated data resources

Availability: www.ebi.ac.uk/microarray-srv/mat/

1 INTRODUCTION

Data in public biomedical databases typically has various classes of metadata associated with it that enable searching and analysis, and standards for different data types and domains are now becoming available (e.g. a series of *Minimum Information* protocols for this purpose, mibbi.sourceforge.net/resources.shtml). There is no such minimal standard for annotating anatomy because tissues are much harder than other (e.g. experimental) data types to formalize simply. This is partly because organisms have so many diverse tissues and partly because tissue organization is so complex. Nevertheless, because of the need to handle tissue-associated data in databases, user communities for all of the main model organisms have produced formalized and fairly complete anatomical hierarchies (ontologies) that are largely based on *part_of* and *is_a* relationships (Bard, 2005, 2007; Smith et al., 2007; Burger et al., 2007). These high-granularity ontologies are complex and their use presupposes considerable anatomical knowledge of the organism whose anatomy is represented, as well as some understanding of the representation format of the ontology. They are therefore mainly used by specialist curators annotating data for the main model organism databases using rich annotation tools (e.g. Phenote, www.phenote.org).

Elsewhere, anatomical annotation is essentially free text, or at best loosely controlled. Databases such as those from the NCBI (www.ncbi.nlm.nih.gov) typically do not control anatomical annotation, and text mining is needed to extract any anatomical information. It is unrealistic for these multi-species databases archiving high throughput data to develop annotation tools that provide intuitive access to all (anatomy) ontologies and expect biomedical users to use them consistently. ArrayExpress (www.ebi.ac.uk/microarray-as/aer/entry), for example, uses a text-mining strategy and string-matching methodology that adds no burden at the point of submission but does require representative ontologies for automated annotation (Parkinson et al, 2006).

For query purposes, a simple anatomical ontology is needed that allows searching and tree browsing, with its complexity limited to that which is comprehensible to a bench biologist. The simplest format for accessing annotation terms is a controlled vocabulary or terminology where informal relationships connect the terms (unlike an ontology whose formal relationships carry inheritance implications). Two such terminologies are currently available: the eVOC terminology set (Kelso et al., 2003) whose scope is limited to human and mouse, and the very short SAEL terminology (Parkinson et al., 2004) mainly intended for core mammalian anatomical annotation and which has no relations at all. Neither resource includes identifiers for other anatomy-based resources that can be used for cross-mapping and interoperability purposes.

This paper reports the development and validation of a terminology entitled MAT (*Minimal Anatomy Terminology*). It is similar in format to eVOC but expanded to include high-level tissues and timelines appropriate for the great majority of taxa rather than just mammals. Data associated with these tissue terms include synonyms and ontology identifiers for tissues from other anatomical ontologies currently downloadable from the Open Biomedical Ontologies (OBO) website (obofoundry.org/), and is thus compatible with them.

Figure 1 MAT terminology displayed in the CoBrA editor (www.xspan.org/cobra). The left panel shows the four top categories with 'anatomy basic component' expanded. The 'eye' has been expanded in the middle panel to show parent and child terms, synonyms and identifiers. The right panel shows the organizing classes 'taxon ontology' and 'time stages'

The MAT terminology is designed to facilitate the easy annotation, curation and searching of tissue-associated data while the ready availability of the various ontology identifiers will facilitate tissue-associated interoperability across databases

2 METHODS & RESULTS

Scope

The MAT terminology is intended to cover the basic anatomy for all common taxa from fungi to plants and animals to support anatomical information and mapping of data contained in existing public resources. MAT is not intended to represent formal knowledge about all these organisms with its inherent implications for inheritance.

Identifiers

Each term has an identifier of the form MAT:0000001, and is mapped to one or more identifiers from the anatomy ontologies currently available in the OBO foundry (Smith et al., 2007). It also contains identifiers from the Mammalian Phenotype Ontology (Smith, 2004) as these typically include anatomical information and may be useful in the context of mapping abnormal phenotypes within an anatomical context.

Granularity

Determining the appropriate level of granularity for MAT is critical: too light and its archiving and searching uses would be inadequate; too heavy a complexity would prohibit use by non-anatomists. The main indicator for tissue selection is that formal species-specific ontologies (*Drosophila*, mouse etc) include these terms at a high level in their respective representations. A second indicator is that the selected tissues should be accessible for molecular analysis. A third was that their meaning was obvious and unambiguous to a biological user.

The current version of MAT has ~400 anatomical child terms of the class *anatomy basic component* (Fig. 1). The majority of these are used in their stage-independent form. This is possible as most of the external ontologies to which MAT is mapped have either restricted their scope to adults or are structured so time and tissue are handled independently (Burger et al., 2003).

Organizing principles

The terminology is intended to be intuitively navigable by a biologist, and obvious choices for high level terms in the hierarchy were *organ* and *major tissue systems* as these both underpin anatomical organization in an intui-

tive way and are used by most anatomical ontologies. MAT includes ~300 animal, ~75 plant and ~20 fungal systems and tissues. Where tissues naturally fall into more than one system (e.g. the mouth is both a craniofacial tissue and part of the alimentary system), multiple inheritance has been used.

MAT includes two high level nodes in addition to *anatomy basic component: taxon ontology*, and *time stage* (Fig. 1). Detailed staging for each organism is outside the scope of MAT, but a generic set of 11 stages each for animals and plants that extend from the zygote to adult are used. This allows distinctions to be made between, for example, the embryonic, the juvenile and the adult testes.

It should be noted that a few ontologies (e.g. adult human and adult mouse) only handle adult tissues, and their identifiers should not be used for developmental tissues.

As the MAT terminology is designed for mapping and annotation rather than for logical inference, the use of formal relationships such as *is-a* and *part-of* were replaced by the single *broader-than* relationship used as defined by SKOS, the Simple Knowledge Organization System, (www.w3.org/2004/02/skos/, a part of the Semantic Web (www.w3.org/2001/sw/). This allows us to represent the terminology as a tree with a single, informal relationship carrying no inheritance implications.

The MAT terms are intended to be species-independent, *trachea* in the *respiratory system* has the associated identifiers from the *Drosophila*, human and mouse anatomy ontologies even though the insect and vertebrate tracheae are very different – they are analogues and not homologues. A *sensu* tag is used in only twice: the vertebrate and invertebrate limbs are so different in structure and development that it seemed unreasonable to include them under the same term, while the insect and amphibian fat bodies are neither homologues nor analogues. MAT also contains some transitional development-specific tissues with no timing details (e.g. somite). These terms were included as they would thus not be present in adult organism lists.

Different anatomy ontologies use different terms and spellings for what are essentially equivalent terms (e.g. oesophagus and esophagus, digestive system and alimentary system). We have made a subjective decision to use the most common term as the standard (e.g. *eye* rather than *visual system*, see Fig. 1), but synonyms are included in the file and can be searched. In assigning identifiers from other anatomy ontology to MAT terms (~1600 in all), there was sometimes a choice as to which term to map to. In the *Drosophila* ontology, for example, there is a term for the digestive system and sub-terms for the embryonic/larval digestive systems and the pupal/adult digestive systems. Where alternatives exist the broadest

term is used preferentially. A very few tissues have been included that are not present in other anatomical ontologies as they may be interesting in a wider evolutionary context (e.g. *phyllid*, the gametophyte leaf).

The MAT terminology has very few text definitions as almost all the terms are in common use by biologists. Indeed, it was often impossible to provide anything but a very loose definition for tissues from different taxa with the same name (e.g. mammalian and invertebrate *trachea* are both involved in the respiratory system, and this is explicit in the terminology). The definitions that are provided cover tissues that may be unfamiliar (e.g. *phyllid*) or whose meaning is slightly technical (e.g. *mesonephros* – adult).

We explicitly decided not to use the CARO upper level anatomy ontology (Haendel et al., 2007) as it is not intuitive to the biologist and is therefore not useful for use in annotation tools or browsing data, and is actually intended for use as a template in developing anatomy ontologies rather than for representing multi-species mappings. We also decided not to adopt the view that multiple parentage of terms is undesirable as we are not trying to represent full anatomical knowledge, rather to produce a resource to aid data integration pragmatically, and biologists intuitively comprehend multiple parentage as is, for example, present in the Gene Ontology (Ashburner, et al., 2000).

Validation

Prior to the construction of the MAT terminology, the ArrayExpress user supplied annotation of ‘OrganismPart’ comprising 817 unique terms used in the annotation of >60,000 samples obtained from >200 species was mapped to multiple anatomy ontologies using a Perl implementation of the MetaPhone ‘sound-alike’ algorithm (Phillips, 1990). The FMA was found to provide the highest coverage of all existing anatomy ontologies, but still covered only 38% of ArrayExpress anatomical annotations. MAT was mapped to the ArrayExpress annotations three times during the development of the MAT terminology and the results curated to identify categories of coverage. The final coverage is ~39%. This figure is comparable with the FMA, and uses only 400 terms to achieve the same coverage. The FMA in contrast contains 25,000 terms and is far less tractable in the context of annotation tools and usability for the general biomedical scientist.

Format

The MAT terminology uses the OBO (oboedit.org) flat file format which allows SKOS relationships, and was constructed using the COBRA editor which has good annotation capabilities that facilitate the mapping of properties such as identifiers and synonyms to MAT terms (www.xspan.org/cobra).

3: DISCUSSION

The aim of the MAT controlled vocabulary is not only to produce a standard terminology which can be assigned to any anatomical parts from any organism, but to provide primary search terms for those interested in accessing tissue-associated data. It is intended as a way of integrating data and allowing interoperation between many ontologies.

MAT is also intended to help with the strategy of determining the molecular basis of some process in one organism by using information relevant to its development and function gleaned from other organisms. Here, MAT provides candidate tissues and identifiers, although MAT tissue groupings may or may not be viewed as equivalent in any particular context, and the onus is on the user to choose which tissues may be relevant to their own and, in turn, which associated data is helpful.

MAT may also be useful in the wider context: as more data is being generated and funders and journals require data to be archived, it becomes impossible for database curators to keep up with the annotation needed for archiving the files, and indeed, harder for the funding agencies to be able to provide the necessary financial support. A practical solution to this problem is that people who deposit material in databases annotate their own data in at least in part. This has always been difficult to achieve formally for tissue-associated data, and we hope the use of terminologies such as MAT will be helpful here.

We expect that the majority of users will be interested in a limited number of taxa, and an editing tool (e.g. COBRa or OBO-edit) can be used to select only the tissues for particular organisms. Note that MAT does not seek to replace the existing ontologies. A more common problem may be that MAT's granularity may be too coarse, and terms may need to be added. This could be solved by using a species-specific ontology, free text, or evolving the MAT for a specific groups needs. Suggestions, criticisms and requests should be emailed to j.bard@ed.ac.uk.

The MAT terminology does not address the issue of developing an all-encompassing multi-species ontology that precisely describes orthologous anatomical parts across evolutionary time. This is a much larger task and has been attempted in the development of the Bilateria ontology used in the 4DExpress database of developmental gene expression data (Haudry et al., 2008). We and others are participating in discussions to make this a more general effort. We applaud these efforts and hope that MAT will be useful in the interim.

ACKNOWLEDGEMENTS

Thanks to Dawn Field for comments on the manuscript, and to Dawn, Stuart Aitken, Nick Kruger, Robert Stevens and Steve Taylor for discussions.

FUNDING

JB thanks the Leverhulme Trust, HP, JM, TFR are funded in part by EC grants FELICS (contract number 021902), EMERALD (project number LSHG-CT-2006-037686), Gen2Phen (contract number 200754) and by EMBL.

REFERENCES

- Ashburner, M, Ball, et al. (2000) *Gene ontology: tool for the unification of biology*. The Gene Ontology Consortium, *Nat Genet*, 25, 25-29.
- Bard JBL (2005) *Anatomics: the intersection of anatomy and bioinformatics*. *J Anat* 206: 1-16.
- Bard J (2007) *Anatomy ontologies for model organisms: the animals and fungi*. In: Burger A, Davidson D, Baldock RA editors. *Anatomy Ontologies for Bioinformatics*. Springer. pp 3-26.
- Burger A, Davidson D, et al. (2003) *Formalization of mouse embryo anatomy*, *Bioinformatics* 19: 1-9.
- Burger A, Davidson D, et al. (2007) (editors) *Anatomy Ontologies for Bioinformatics*. Springer. pp 356.
- Haendel, MA., Neuhaus, F, et al (2007). *CARO – the common anatomy reference ontology*. In: Burger A, Davidson D, Baldock RA (editors). *Anatomy Ontologies for Bioinformatics*. Springer. pp 327-350.
- Haudry Y, Berube H, et al. (2008). *4DXpress: a database for cross-species expression pattern comparisons*. *Nuc Acids Res* 36: D847-53.
- Kelso J, Visagie J, et al (2003). *eVOC: a controlled vocabulary for unifying gene expression data*. *Genome Res* 6A: 1222-30.
- Parkinson H, Kapushesky M, et al. (2006). *ArrayExpress - a public database of microarray experiments and gene expression profiles*. *Nucl Acids Res* 35: D747-750.
- Parkinson H, Aitken S, et al. (2004) *The SOFG Anatomy Entry List (SAEL): An Annotation Tool for Functional Genomics Data*. *Comp Funct Gen* 5: 521-527.
- Phillips L (1990) *Hanging on the Metaphone*. *Comput Lang* 7: 39-49
- Smith B, Ashburner M, et al. (2007). *The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration*. *Nat Biotechnol* 25: 1251
- Smith, C, Goldsmith, C-A and Eppig, J (2004) *The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information*, *Genome Biology*, 6.R9