Ontology Workshop Living Document

This ‘Living Document’ is intended to be used by participants at the BioMedBridges Knowledge Exchange Workshop: Practical Solutions with Ontologies. Hopefully, it can be utilised to capture our collective thoughts on the workshop. Ultimately this information will be collated and used to produce part of the WP12 deliverables. It will be used to provide guidance and advice on ontologies to members of the BMB project.

- **Ontologies**
  - Where to find existing ontologies
  - Ontology expertise within BMB
  - Which ontologies might be good ontologies for BMB to use and why?
  - Ontology expertise outside BMB

- **Technologies**
  - Ontology technologies
  - Technology expertise

- **Getting Started Checklist**

- **Delegate Slides**
  - Plasmodium berghei
  - UMCG 1
  - UMCG 2
  - Software registry
  - InterMine
  - TUM
  - Clinical trials
  - PhenoBridge
  - WP6

- **Use cases**
  - ArrayExpress & EFO
  - KupKB
  - Ontology use by IMPC
  - Ontologies in Instruct/PiMS
  - Ontology use in BBMRI(-NL) and EU-BioSHaRE
  - EuBI cellular phenotype ontology

- **Ontologies FAQ**
  - Why bother with ontologies?
  - What do ontologies need to describe?
  - What’s the difference between an ontology and a thesaurus?
  - What kinds of information can ontologies capture?
  - Which ontologies are ‘open’? What can I do with ‘closed’ ontologies?
  - When is it better to knock together my own ontology vs reuse something that exists already?
  - Lots of ontologies - are they going to be condensed in the future?
  - When should I use the protein ontology vs the gene ontology?
  - Is it possible to bridge ontologies and link them together?
  - Simplifying ontologies, is it possible?
  - What happens when you delete terms?
  - How do you prioritise what to annotate?

- **Glossary of Terms**
Finding Ontologies Exercise
Course Survey
Post-it sorting exercise
  Things there are ontologies for...
  Things there are not ontologies for...
Ontologies

This section of the document is for recording ontologies of use to the project and people with expertise in using them.

Where to find existing ontologies

Please list any resources and information about existing ontologies:

- Open Biomedical Ontologies - [http://www.obofoundry.org](http://www.obofoundry.org) - tightly defined set of ontologies; open, community efforts; orthogonal to each other.
- NCBO BioPortal - [http://bioportal.bioontology.org](http://bioportal.bioontology.org): OBO ontologies +; includes lots of small, one-person efforts
- Terry Meehan - mouse phenotyping using ontologies
- The Software Ontology (SWO) is a resource for describing software tools, their types, tasks, versions, provenance and data associated: [http://theswo.sourceforge.net/](http://theswo.sourceforge.net/). SWO imports EDAM (below)
- EDAM (EMBRACE Data and Methods) is an ontology of common bioinformatics-specific types of data, data identifiers, data formats, data operations and topics:
  - [http://edamontology.org/](http://edamontology.org/)
- Ontology Lookup service - like BioPortal but OBO ontologies only [http://www.ebi.ac.uk/ontology-lookup/](http://www.ebi.ac.uk/ontology-lookup/)
- Swoogle - [http://swoogle.umbc.edu](http://swoogle.umbc.edu) - contains ontologies that aren’t in BioPortal and OLS; covers more than the biomedical domain.

Ontology expertise within BMB

Please list any members of BMB that are currently working with ontologies or have prior experience in using them. If possible, include information on data types and objectives of the work:

- Helen Parkinson, James Malone, EMBL-EBI - construction of application ontologies from reference ontologies
- Check out the ontogenesis blog - [http://ontogenesis.knowledgeblog.org](http://ontogenesis.knowledgeblog.org)
- Jon Ison, EMBL-EBI, developer of EDAM ontology and BioMedBridges Software Registry, email: jison@ebi.ac.uk
- Terry Meehan - all things mousy!

Which ontologies might be good ontologies for BMB to use and why?
• GENES AND GENE PRODUCTS: The Gene Ontology - it’s big and it’s become widely accepted for describing genes and gene products across many species.
• PHENOTYPES: PaTO (phenotype ontology) - the OBO ontology for phenotype;
• DISEASE:
  o Human Disease Ontology (OBO; open) maps to SNOMED; more research focused, biomolecular view on disease
  o SNOMED - disease ontology mandated for use in >60 countries BUT not open in some countries; for example, EBI can use it because we’re based in UK but we can’t redistribute it for use in pan-European and international projects; it’s built for clinical practice.
  o ICD-10 (ICD-9 is open; ICD-10 isn’t); used for describing disease and billing in healthcare practice
  o NCI Thesaurus
  o UMLS; used by libraries
• SMALL MOLECULES: ChEBI - not just a dictionary of the chemistry, but also roles such as drug or pesticide. Therefore useful in the context of biomedical research.
• CELLS: cell ontology
• PROTEINS - protein ontology for uniquely defined biomarkers
• PATHWAYS: BioPax - not an OBO ontology but useful (why?)
• SOFTWARE / SERVICES
  o Software Ontology (SWO)
  o EDAM: Bioinformatics operations, types of data, topics, and data formats. To request terms email Jon Ison (jison@ebi.ac.uk) or see http://www.ebi.ac.uk/panda/jira/secure/IssueNavigator.jspa?reset=true&mode=hide&pid=10426&sorter/order=DESC&sorter/field=priority&resolution=-1&component=10905
• TAXONOMY - NCBI taxonomy
• PROVENANCE:
  o ECO - the evidence code ontology (OBO)
  o PROV-O - W3C, very generic evidence ontology
  o Open annotation mode;
• REPRODUCIBILITY
  o OBI
  o BMB software registry, which uses SWO-EDAM
  o the myGRID ontology: note this is subsumed within EDAM
• RELATIONSHIPS - RO, the OBO relations ontology.
• MICE
  o Mouse phenotype ontology - used for phenotyping mice in the IMPC
  o Adult mouse anatomy ontology - used for tagging photos and to associate with annotations from the MPA
  o Allow users to do faceted searches of the IMPC data - can look by phenotype and/or anatomy
  o Uberon - pan-vertebrate anatomy ontology: helps to make connections between human and mouse phenotype - see Terry Meehan’s mandible example and Damian Smedley paper

Ontology expertise outside BMB

Please list any researchers outside of BMB that are currently working with ontologies or have prior experience in using them. If possible, include information on data types and
objectives of the work:

- Robert Stevens - University of Manchester - definition and applications of ontologies; ontology editing tools; myGRID ontologies?
- Jane Lomax and Janna Hastings, EBI - how to bring together two ontologies with different structures, e.g. ChEBI and GO
Technologies

This section of the document is for recording ontologies of use to the project and people with expertise in using them.

Ontology technologies

Please list any ontology technologies. Where possible provide links to resources, websites and documentation files:

- OWL - the web ontology language; see tutorial at [http://www.co-ode.org/resources/tutorials/intro/](http://www.co-ode.org/resources/tutorials/intro/)
- SKOS - simple knowledge organisation system, for representing lightweight vocabularies that are not ‘true’ ontologies. [http://www.w3.org/2004/02/skos/](http://www.w3.org/2004/02/skos/)
- OBO - [http://www.geneontology.org/GO.format.obo-1_2.shtml](http://www.geneontology.org/GO.format.obo-1_2.shtml) - came from the GO developer community; simpler and more human readable than OWL; has now been mapped to a subset of OWL and OWL can now be used for reasoning and checking OBO ontologies.
- OBO-edit - open source editing tool for ontologies - [http://oboedit.org](http://oboedit.org)
- Protege - [http://protege.stanford.edu](http://protege.stanford.edu) - allows you to create terms and organise them in a hierarchy - essentially a graphical user interface for interacting with the OWL API
- OWL API - the API you need to interact with an OWL ontology
- Spreadsheets - yes, really! You can use spreadsheets to get data from domain experts that can then be sucked into OWL.
- Pellet - an OWL 2 reasoner for Java
- ELK
- Hermit OWL reasoner
- OWL-EL -
- Bubastis - [http://www.ebi.ac.uk/efo/bubastis/](http://www.ebi.ac.uk/efo/bubastis/) - see the difference between two versions of an ontology

Tools for annotation

- NCBO annotator: [http://bioportal.bioontology.org/annotator](http://bioportal.bioontology.org/annotator) - quick way of grabbing annotations for a piece of text; no scoring system though
- Whatizit - [http://www.ebi.ac.uk/webservices/whatizit/info.jsf](http://www.ebi.ac.uk/webservices/whatizit/info.jsf)
- Terminizer - terminizer.org - limited to OBO ontologies

APIs

- Jena - [http://jena.apache.org](http://jena.apache.org)

Editing tools
URigen - tool to generate URIs for ontology terms
Ontology design patterns used by EFO
Bamboo - a continuous integration framework which can be used to perform checks during code commits and to perform build plans, such as releasing the EFO ontology
Phenotator - phenotypes annotation tool; allows you to create ‘entity-quality’ pairs for each phenotype.

**Technology expertise**

Please list any members of BMB currently working with these technologies:

- Jon Ison, EMBL-EBI, email: [jison@ebi.ac.uk](mailto:jison@ebi.ac.uk)
  - OWL, OWL-API, OBO, OBO-Edit,
- Krish, STFC & Chris Morris, PiMS-STFC email: [narayanan.krishnan@stfc.ac.uk](mailto:narayanan.krishnan@stfc.ac.uk) or [chris.morris@stfc.ac.uk](mailto:chris.morris@stfc.ac.uk)
- RDF, Apache Jena - API
Getting Started Checklist

Ontogenesis blog - http://ontogenesis.knowledgeblog.org/

Please add items to the list below to cover basics of ‘getting started’ with ontologies

- There is no simple solution for dealing with the enormous amount of overlap between existing ontologies - it’s an ongoing issue in the field.
- Identify your needs first: we use...
- Competency questions
- Use cases
- Don’t worry about annotating to detail - see Terry Meehan’s slides; OK to annotate to kidney and then have a kidney expert look at the annotations to differentiate different parts of kidney
- Things to consider when deciding to use an ontology or incorporate it into an application ontology - see James’s list of four from his tuesday morning slides. Is it actively developed; do terms have definitions; two other things.

Delegate Slides

Plasmodium berghei

ID GO terms in pseudogenes of different Plasmodium falciparum isolates. Problem: low annotation.

UMCG 1

- Software
  - At UMCG we host biobank catalogues and patient registries (phenotypes/genotypes)
  - all implemented in open source software MOLGENIS, see http://www.molgenis.org and http://github.com/molgenis
  - Current demo at http://molgenis19.target.rug.nl:8080/lifelines (may be offline again when you read this).
- Data
  - In BBMRI-NL and CTMM (=Dutch EATRIS) we have a shallow biobank catalogue of 188 sample collections with basic characteristics (materialtype, general categories, size, gwas, ...), 400.000 cases
  - In EU-BioSHARE we have a deep catalogue of 5 biobanks, having full descriptions on each data item (average 1000 items per biobank)
At UMCG we have >150,000 samples in LifeLines and Prevend biobanks, for which we have full information (although we cannot share individual data)

- **Goals**
  - BioBankers want data pooling for statistical power in analysis
  - Translational researchers want to find patients matching their own, e.g. to steer diagnosis and treatment decisions

- **Practical pilots proposed:**
  1. implement (REST) service to find sample collections across biobank catalogues (in NL but also in EU). Would like to agree on a standard syntax. Optionally, also have standard ontologies for typical
  2. pilot a search across biobank catalogues that contain sample collections given our research needs (data items, inclusion/exclusion criteria, properties such as ‘materialType’, ‘disease’). Can query expansion help?
  3. match variables across biobanks automatically by using knowledge from ontologies. Should we annotate the biobanks first or is query expansion all we need (we have working implementation that Chao Pang can demo to you)

**UMCG 2**

- Next to biobanks we also host model organism and multi-omics consortium databases for Plant, Worm, Rat, typically QTLs.
  - See for example [http://www.wormqtl.org](http://www.wormqtl.org)
  - We would like to bridge data across species and data modalities
  - e.g.
    - GO - tools to bridge from model organism to human and back
    - Anatomy terms
    - Gene

**Software registry**

[http://wwwdev.ebi.ac.uk/fgpt/toolsui/](http://wwwdev.ebi.ac.uk/fgpt/toolsui/)

Short term aim: include tools provided by BMB partners or required by BMB Use Cases.
Longer term: include more tools (all bioinformatics software)

Uses ontologies to describe tools:

- **EDAM:**
  - Software/service categories
- Software function(s)
- Software inputs and outputs (types of data)
- Input and output data formats
- SWO:
  - Software type
  - Provider names
  - License
- Other:
  - NCBI taxonomy (organism names)

Open question - please help!:

*Are there other vocabularies we should use to describe the tools?*

UMCG - has tools to provide

**InterMine**

FlyMine + model organism databases
How to interoperate between databases?
Based on ontologies and how to bridge them

**TUM**

No ontology experience so far
What tools to use to get started?
Shared ontology development?
BBMRI - ontologies useful for biosamples etc

**Clinical trials**

Visualisation of ontologies
Can you model process of clinical trials using ontologies?

**PhenoBridge**

Transferring diabetes classification of molecular causes and consequences into an ontology
Available languages and standards
Visualisation of ontologies
Reference Ontologies - Mouse and Human Phenotype Ontologies
  Many terms already align

**WP6**
Demonstrate interoperability imaging data
KO in cell culture
Compare phenotypes of tissues - cell division
Currently annotated using free text
How to move to an ontology
Use cases

This section of the document is for recording information on use case scenarios for ontologies within BMB. Provide details of who the users are, the data being worked with, what they need/want to do with their data, how it might intersect with different areas of the project and how ontologies can be of benefit to them.

ArrayExpress & EFO

Building ontologies - focussed on gene expression
Array Express
   Direct submissions and GEO
   Curated
   Many species and experiments
   Diverse annotations, samples etc
Subset of AE into Gene Expression Atlas
   Further curation, annotation with ontologies, analysis and summary
Motivation - 2007 use case
   Query support - high level searches return many results
   Working towards intelligent template generation for different experiments
   Nonsense detection
Primary needs
   Ontology that serves use cases
   Reuse where existing resource appropriate
   Efficient build vs accuracy
EFO is an application ontology
   Built towards own application needs
   Needs to work - treated like software engineering project
   Check and self-test code
   Predictable release cycle
Tools
   URiGEN - used to generate new identifiers in ontology
   Standardised design - Minimal metadata set for each class
   Bamboo - checking software (like Jenkins)
   Bubastis - Checking external ontologies as they update
Ontology used to annotate date in AE/Atlas
OBO Foundry used extensively
Objective analysis is hard - but check
   Coverage
   Actively Developed
   Textual definitions
   Is it used elsewhere?
Activity to check how stable an ontology is over time - additions/deletions
   Q - can take an archived ontology for use?
Adding synonyms enriches searching
Upper ontologies - BFO (?)

KupKB
Small data integration task - 3 people
Kidney + Urinary Pathways - ID of biomarkers in urine and kidney
Combines data on: human + mouse; urine, tissue, cell; gene & protein
Trust PubMed over generalist databases - most of the data lost (supplementary data)
  Attempt to re-capture this data
Nano-publication - what has been observed, where and when?
  Highlight key bits of info
  Collect this info to create knowledgebase
Connected reference ontologies in building an application ontology
Knowledge>Ontologically>Differentia>Real Examples
Populous tool - ontologies by stealth - built by biologists
  Converted into ontology ready for use
195 experiments - pulled all data into a template - created into KUPKB
  Browser for data - iKUP
Resulted in new biological discovery - evidence found in un-indexed supplementary data

**Ontology use by IMPC**

International Mouse Phenotyping Consortium
  Create mouse models to understand mammalian gene function
  Create KO mouse strain for every protein coding gene
  Systematic broad-based phenotyping
  Both OBO-based ontologies
Mammalian Phenotype Ontology (rodent)
  40k genotypes, 8k genes, 8700 terms
MA- Adult Mouse Antatomy
  Describe anatomical sites of gene/protein expression
  3k terms
MPO (Mouse Phenotype Ontology) - Large number of tests once mice reach maturity
  Each test has multiple parameters, each parameter has an ontology term
  Can take multiple stat test and collapse down to biological terms
  These terms can then be searched for
AMAO (Adult Mouse Anatomy Ontology)
  LacZ replacement
  Determine where gene is expressed - annotation at production centres
Breakdown of ontologies by viewable images
  Explore gene expression by images
  Good vs Perfect annotation - ongoing work to update initial annotation
  Workshops or crowdsourcing in future
2ndary phenotyping
  Will require more ontologies to describe phenotypes
    EMAP, EMAPA, extending MP, Cell Type Ontology, HPO, UBERON
HPO - use OMIM as basis, extending across rare diseases towards common diseases
  Find connections between HPO and MP terms; could use HPO to map phenotypes in different biobanks to a single ontology. Helen suggests that Mauris starts with Diabetes and AML samples to fit with the other BMB projects.
Lots of applications in translational research - Terry Meehan working with diagnostics community and others e.g. cancer, eye disease. Workshop coming at NHGRI.
Challenges
Ontologies change
MA = partonomy - built up of parts, not child/parent (can someone explain?)
EMAP = developmental stage specific
Annotation can vary
Complex queries need additional work
Semantic representations

WP7 Phenobridge
Working w/ clinician and their diabetes terms - investigating the terms in MP and HPO
Sweet Preference challenge, Anorexia and Diabetes in HPO, but not MP
Diabetes can be classed as a disease, not just as a phenotype

Detecting changes in ontologies as they are updated
Jenkins (link?) used to do this
Terms retired or made obsolete - don’t delete terms as this will destroy annotation in other resources

Requesting changes - request term through tracker - see SourceForge - often ontologists will annotate with reasons on why something is accepted/rejected

Ontologies in Instruct/PiMS

- PiMS has laboratory data from recombinant Protein production experiments.
- The data already has references to resources such as Uniprot
- Use case to provide semantic links to these resources as a way of integration

Protein Information Management System
capture experimental information for protein-protein recombination interactions
Living organisms>instruments>publications
https://twitter.com/search?q=%23overlyhonestmethods
How to trap data for future use?
Recording electronically now, for future return on investment
Knowledge transfer becoming more important
Data warehousing - dump for research data - can it be made useful and of use outside of original context

Crystallogenesis - well defined experiment - Cubic Phase techniques changed this
Protein expression is highly varied, unlike simple DNA/RNA expression models
Embrace ‘not yet’ philosophy

Targets - every ORF is a potential target - confuses areas of science

Access Rights - generally a closed process within a lab, science is an open process
Cognitive dissonance
Tech
- RDBMS with schema in DDL
- Java w/ business objects - DAOs, DTOs
- UML data model
- Strong schemata - but inflexible

e-IRG recommendations
- Persistence of metadata - doi, business plan for continuation
- Interoperability - annotation in multiple languages
- decouple object from metadata

Benefits of RDF
- Law of triples representing entire database
- Machine/user readable - request basis
- Subject - RDF Type - Object
  - Linked via URL

Writing up a lab notebook - how to incorporate this methodology
  - Struggle to go backwards

Ontology use in BBMRI(-NL) and EU-BioSHaRE

- Use cases
  - BBMRI-NL catalogue (188 cohorts) with free text descriptions: use query expansion to find matching cohorts?
  - LifeLines biobank catalogue (1000 phenotypes) with free text descriptions: use query expansion to find data items?
  - EU-BioSHARE harmonization project (5 biobanks) with free text descriptions: annotate the desired variable; then scan the 5 biobanks for candidate matches?
  - BBMRI-EU meta-search: can we use ontologies to enable cross-country searches across the national and EU catalogues?
- Questions:
  - should we annotate all variables in all biobanks (which would be tremendous effort and social engineering)?
  - or can we have much more return on investment using a query expansion + string matching approach?

EuBI cellular phenotype ontology

Marking up existing data...

- Edinburgh Mouse Atlas (EMAP) - who?
- Virtual fly Brain (www.virtuallyflybrain.org) - David Sutherland
• UniProt GO Annotation team at EBI
• GWAS catalogue [www.genome.gov/gwasstudies/](http://www.genome.gov/gwasstudies/) - mapping of SNPs to different traits - Helen Parkinson’s team are mapping the traits to ontologies so that you can see where traits cluster on the genome

New knowledge using ontologies
Gene set enrichment analysis - using GO to determine whether a list of genes is associated with a specific phenotype/condition/prognosis

**Towards a cellular phenotype ontology - Systems Microscopy**

Using existing ontologies to represent data from WP6 and Systems Microscopy
Gain a systems biology level understanding of living cell
Automated live cell imaging
4D assays
Develop repository for data from a infant -omics domain
No standards and varied approach to analysis
Attempt to develop unified approach
Receive raw images - extract data from images

Data
Metadata
Study description
siRNA library information - RNAi screens to silence screens
Gene symbol, siRNA ID, phenotype as searchable terms
Currently stored as free text - difficult to search
Attempt to create ontology for this info

Challenges
Integration at quantitative level
Integration at level of phenotypic descriptions

Cellular Phenotype Ontology
Pre-composed based on GO BP, GO CC, GO extensions and PATO
Split into morphological and processes

Community needs a tool to annotate ontology
Distribute to data producers

Phenotator tool - searching through BioPortal > map for annotation

Entity and Quality annotations
Include accuracy scores to rate how useful/accurate terms are

Problems
Orphan phenotypes not linked to biological process due to lack of info to make association
Observing cell populations - move to single cell observations
Quantitative phenotypes?
Granularity of ontology not sufficient for some areas of cell biology - migration
Temporal information is important, but not covered well by existing ontologies
Goal

Generic upper level ontology
Open access tool for annotation

Prototype URL: http://www.ebi.ac.uk/fg/sym

Q - are you limited to components when describing entities?
How to generate entities then ascribe a quantity

RDF and the Semantic Web

Integration of data

1st Gen - linked documents
2nd Gen - web as a platform
3rd Gen - semantic web
   - platform for publishing/querying data in structured format
   - Semantic markup - adds meaning to data for machine processing

Data can be fit into a graph of linked items
   - Needs to be machine processable

Resource Description Framework - RDF
   - RDF statements - describe properties and property values of a resource
   - All statements in RDF are triple
   - URI id’s concept - open platform for referring to objects
     - Others can then use it to relate information to that object
     - Mechanism to identify ‘things’
   - Ontologies provide the meaning
     - Allow inferencing over data

Dbpedia
   - Create RDF from wikipedia data boxes
   - Links to other structures
   - Can be then queried - SPARQL is query language
     - Essentially queries the graph nodes
     - Machine can ‘crawl’ through graph data

Data resources are now starting to publish their data according to this format

Gene Expression Atlas and other EBI databases are being opened in this manner
   - Linking data and information together
   - RDF schema created - assigning ontology terms
   - Reactome already publishing RDF
   - Can detect inconsistencies in data annotation

Q - where does microdata fit into RDF? Is it a step to far?
   - RDFa is microdata, slightly different way of representing data
Q - practical information on RDF stores?
   Millions, if not billions of statements. Can now go for real scalability.
Q - Is there a future in the Semantic Web, there is a need to have lots of data to make it worthwhile?
   Nobody is quite sure how to do it, but appreciate moves must be made to work towards it. Biology has always been very good at sharing data - not true for other areas
Ontologies FAQ

Why bother with ontologies?

Data are useless unless we describe what they are; we have to agree on what the concepts are and what are the labels that we use for them.

What do ontologies need to describe?

The entities themselves - genes, proteins, processes, cells, properties
The investigations that produced the entities
The informational origins and history of those entities and their descriptions (data and provenance)

What’s the difference between an ontology and a thesaurus?

An ontology describes both the entities and the relationships between them; a thesaurus doesn't make the distinction between a concrete thing (e.g. an enzyme) and a process (e.g. the catalysis of a reaction).

What kinds of information can ontologies capture?

Process - e.g. Gene ontology cellular process
Information - e.g. information artifact ontology
Material - e.g. Chemical Entities of Biological Interest; NCBI Taxonomy; GO cellular component
Property - e.g. GO Molecular function; Phenotypic Quality (PaTO), Human Disease Ontology
Site - e.g. Gazetteer Ontology

Which ontologies are ‘open’? What can I do with ‘closed’ ontologies?

The use of many ontologies is open and they are free to use. However, they may be limited for other uses - i.e. not necessarily possible re-distribute an ontology.

When is it better to knock together my own ontology vs reuse something that exists already?

Better to reuse something that exists if you can; this may mean accepting that some things are not described in quite the way you want them to be, but it may be worth this compromise to avoid reinventing the wheel. Alternatively you can create your own and crosslink to an existing one (useful if the community developing the similar ontology is a bit slow moving); another thing that may influence your decision to make your own is whether you can afford to buy a proprietary one?

BioMedBridges
Lots of ontologies - are they going to be condensed in the future?

Many different deployments, but there are different reasons for them existing. They are used in many different forms and in different ways. Unlikely to stop using some ontologies. Example - we have ChEBI, but industry has their own descriptors and unlikely to stop using them. Different perspectives can be applied to a single ontology

When should I use the protein ontology vs the gene ontology?

UniProt uses GO, so what’s PO for?

Is it possible to bridge ontologies and link them together?

This is an ongoing research problem. It is possible to find common ground between two different ontologies, but this is not always possible. It is usually easier to build an ontology as part of a use case, rather than trying to leverage two existing ontologies together.

Simplifying ontologies, is it possible?

You can remove terms and remove relationships between entities.

What happens when you delete terms?

Need to explain about obsoletion process. One of the OBO rules is that terms should be obsoleted, not just deleted.

How do you prioritise what to annotate?

Focus on low-hanging fruit:
Query support - what are your users asking for?
Data visualisation
Other things from James’s Tuesday morning slides

What should we hide from our users?

Need an answer from James

How can I gather information from domain experts for my ontology without exposing them to scary tools like protege?

Glossary of Terms

Please use this section of the document to list any terms that would be useful to combine into a BMB Ontology Glossary. If you have terms on which you are unsure of and would like clarification, then add them in and hopefully someone can provide the information you are after.

Ontology
A set of logical axioms designed to account for the intended meaning of a formal vocabulary used to describe (conceptualisation of) a certain reality (Guarino 1998). Ontology describes the nature of what is, not how it appears; a thesaurus doesn't make this distinction. For example, a thesaurus (MeSH, for example) doesn't distinguish between a concrete thing (e.g. an enzyme) and a process (e.g. the catalysis of a reaction).

Reference ontology
Built with the aim of becoming an authority on a given domain (e.g. the Gene Ontology). OBO (http://www.obofoundry.org) lists the major reference ontologies in the biological and biomedical sciences.

Application ontology
Built towards a specific application use case, such as for tooling or database needs. They often consume reference ontologies (e.g. the Experimental Factor Ontology).

Orthogonal
Two ontologies are described as orthogonal when they’re independent of each other; they don’t contain overlapping terms. For example, when describing anatomy we don’t include cell types because there’s a cell ontology.

Slim
A slimline version of an ontology comprising the higher-level terms.

Class/concept-term/type
xxx
other things from Steve Jupp’s slides - Steve, can you add these with definitions?

Description logic
http://en.wikipedia.org/wiki/Description_logic

OWL
The web ontology language; an XML-based language for describing ontologies. It includes constructs for asserting classes, individuals and the relationships between them.

RDF
The resource description framework (come back to this - need better definition).

Triple
a data entity composed of subject-predicate-object, like "Bob is 35" or "Bob knows Fred". They are stored in databases called triplestores - http://en.wikipedia.org/wiki/Triple_Store
DAG (Directed acyclic graph)
http://en.wikipedia.org/wiki/Directed_acyclic_graph

Semantic reasoner
A tool that allows you to infer relationships between different classes in an ontology.

Use case
Typically (in software engineering) interaction between a user and a system to achieve an objective. In ontology engineering, use cases can be divided:
Interaction with data and ontology
interaction with system and ontology
Interaction with user and ontology
Often require different considerations

Competency questions
Does it cover all my data?
Do definitions of classes correspond to my data?
Is x a subclass of y
...

Finding Ontologies Exercise

Stuff what was useful:

Stuff that was NOT useful or hard:
Post-it sorting exercise

Things there are ontologies for...

Phenotype - multiple ontologies - Yeast, human, mouse...

Data and assay types - OBI, BioAssay
Disease Ontologies - too many!
Material types - characteristics of samples, depending on species
Life course events and treatments - meta-application ontology - may come from patient records coding
Clinical trial processes - Clinical trials ontology, HL7 includes clinical interventions, OBI has recruitment terms
Clinical data from human patients - some ontologies may cover this; join in instead of making your own; link and claim interoperabilities
Drugs (clinical trials) - ChEBI (will add to)
Symptom Ontologies - data dictionaries, ontology of tests (ask Tomasz Adamusiak), covered in SNOMED - but this is difficult in a usable form
Mouse Models of Disease - MGI of OMIM diseases, currently ongoing work, no formal ontology
Organs - anatomy ontologies, vertebrate ontology, UBERON - cross-species anatomy ontology, VHOG; these both are being integrated to function better, looking at developmental ontologies
Ontology for omics data - OBI and EFO covers some of this
Ontology for NGS (process and end point annotation) - OBI or EFO, Sequence Ontology
Mapping of Bioassays - Terry to elaborate tomorrow, Too many!
Ontology for neuro-development - access is limited
Tissues & Bio-samples - Foundation Model of Anatomy (non-disease), developmental anatomy POV, not a clinical/surgical; UBERON; Cell Type Ontology - combine ontologies to map out cell derivations - GOOD CASE POINT - Ontologies often developed from a POV, might not be that useful for other applications
Bio-samples - FMA has bodily fluids (Blood, Sweat and Tears Ontology - Barry? Bodily Fluids Ontology)

Things there are not ontologies for...

Cellular phenotypes - One is being built! - Morphology, Structure, single cell vs population
Drug/Disease relationships - NCI Thesaurus covers some cancers; Database markup for clinically approved drugs?
Ontology for databases - useful for discovering services
Disease and altered pathways -
Ontology for Biobank - might need breaking down into separate ontologies (emerging ontology for biobank processes)
Informed Consent - defining data that can be accessed/used (is an ontology the right way to approach this? Would it be part of a database with this information)
Questions on Ontologies

Evaluating and Benchmarking an ontology

- Problem is no real agreement on how to do so
- Do people agree with you?
- Has it been built according to personal methods, viewpoints or is it fit for purpose
- OBO - adhere to naming conv, semantic variable, not deleting terms, data labels, being open and public
- Fitness and purpose are perhaps the best approach?
  - Usual way is via competency questions
  - Can it answer X, provide Y information
  - Have you got all the answers?
  - How often does it change, get updated? Is it actively supported?
- No existing technology to ‘benchmark’ an ontology
  - Unlikely to be in the future

What is best practice?

- OBO Consortium are holders of best practice in the community
  - Openness, activity, independent, naming conventions, semantic free IDs, versioning are all good
  - Follow the above for a ‘good’ ontology
  - Use OWL or standard OBO formats

Is there a graphical ontology creation too?

- Yes! And No...
- Experimental tools exist
  - Might not scale very well
  - Good for outline structures
  - Rich axiomatization doesn’t work well
  - Protege doesn’t have this as standard

I would like to rank matching ontologies

- NCBO Recommender Service - http://bioportal.bioontology.org/recommender
- Need to add underlying information
- Terms can match elsewhere, especially in large ontologies
How do you find mappings between ontologies

- Using BioPortal
  - Uses lexical mapping
    - Sort of... matches, not always clear
- Look at Ontology Alignment
  - Structural techniques can be used
  - Alan Recter - Developed Galen
    - Overlapping regions used
    - Different labels for regions
    - Give different levels of alignment

What is the most common representation of an ontology?

- Looking at Biomedical Ontologies
- OBO
  - OBO Format
    - Could be viewed as an OWL syntax - is the question moot?
  - OBO make ontologies available in both OBO Format and OWL
  - Can become unmanageable in OBO format
- BioPortal - framebased systems
- SNOMED + NCI Thesaurus are OWL
- Manchester Syntax

How can I describe the result of a mice phenotyping procedure?

- Read the paper and make an assessment
- IMPC - assign pheno terms to outcome of assay
  - Auto-annotate based on outcome
  - 1+ associations can be made
- Understand the paper and ontology
  - Allows granular assignments and good annotation
    - But hard and expensive

How can ontologies deal with phenotypes based on quantitative traits?

- Use data properties to assign values to an annotation
- Straightforward, but requires agreement
- Ontologies better at class level, problematic at quantitative level
- Can be done, but there is a computation cost
  - P-values can be included
  - No defined classes based on numerical values
  - Difficult to compare between data from different curated sources
- Could be done with an RDF
Can I find out which data items in a study map to things of interest?

- Yes, you can!
- If data is there, ask the question
- PHENX - standardisation of stats
  - Can map to PHENX and then get values out
- Joining in community efforts
  - Set of vocabularies created
  - Connections between them - eg ChEBI and GO

How do you find groups of patient suitable for a stratified clinical trial?

Need patients marked up with an ontology...

How do you match phenotype ontologies between different species? Are there any tools that can help?

- UBERON + OWL-SIM
  - Map shared attributes from ontologies
  - MouseFinder
  - Pheno-"something"
  - Aim is to look at disease>human pheno>mouse pheno > model disease
  - OBO-Discuss - email list
  - YouTube video - http://www.youtube.com/watch?v=1wkzBoi33xA

What tools are you likely to use when wanting to annotate tools/services?

Drugs and disease they treat

- No!
- There may be proprietary software

Thoughts from the course

Previous course - one-day ontology course ECCB, very philosophical!
  Difficulties of using them because of the background
  Practical basis!
Overview of many more ontologies not thought about before
WP7 Mouse/human

Julie  Overview v.good
      Good to learn about other progress
      Ontologies for applications

Phillip  OWL and other technical
         Useful for WP4 in this regard
         Different view on ontologies

         Ontology - based on cell physical properties
         Relating back to previous works
         Good to meet people with similar problems from many backgrounds

         Tools and info on ontologies
         would have liked more exercises in the tools section

Willi  V Good overview
       Impossible to do in two days
       Resources for people to go to

James  Use cases - other people’s problems
       Spread of topics and different areas covered by participants
       Big gaps in clinical end of ontologies
       Work moving into this area - but it will be hard

CLOSED ONTOLOGIES
   Does proprietary databases/ontologies create a problem
   Nothing to stop you doing it yourself
   Resourcing issues
   SNOMED is very large, big input from governments
   ChEBI - industry standards are often closed so a big area
   Prioritise work that you need to do
   “Build something that is useful and just use it”
   “Perfect is the enemy of the good”

Sarah  Understand them more

Simon  Brushed over lots of technologies
       Inner workings could be covered again in future

Chao  Great workshop
      I’ve been using them without realising some parts
      More resources
      Tutorial session for applications of ontologies
      Follow up session?

      More hackathon style?
Demonstration of ontology use very helpful

Dennis
Inspiring to hear from different people
   Relational database people

Coffee in the morning!

Q&A
   Practical session
      Would have been good to demonstrate via a resource rather than just name dropping resources

Gather information before the workshop from across the project
   Be concise as possible

Keep up active discussion with participants across BMB
   Wiki, blog, discussion forums in Alfresco

Course Survey

Please spend a few minutes filling out the survey at the link below and let us know how you felt the course has gone:

Survey Monkey