Example Ontologies and their Applications

- **Example ontologies**
  - Cyc
  - Gene Ontology
  - Foundational Model of Anatomy
  - Applications to KM
  - Enterprise Ontology

- **Definition of an ontology**
  - A *shared* specification of a conceptualisation
    - A conceptualisation is a world view, expressed in concepts
    - The specification explicitly constrains how the concepts are related and how they may be instantiated
    - The ontology usually excludes any particular instantiation or world state

We should add: Concepts denote classes of entities that exist in the real world.

Ontologies provide
- Structure (a framework for representation)
  - `subClassOf`, `partOf`
- Content (domain knowledge)

Applications may also need
- Heuristic or logical inference methods
  - Knowledge-based systems
- Methods for computing similarity between concepts
- Methods for indexing data/documents with ontology terms
  - Semantic search engines

Knowledge-based / Expert systems

**Ontology**

- Infection
- Meningitis
- Bacterial Meningitis
- Viral Meningitis
- Surgery
- Neurosurgery
- CardiacSurgery
- Human
- Real Number

**Expert System Rule:**
“If the patient has an infection and his Cerebrospinal Fluid cell count is less than 10, then it is unlikely that he has meningitis”

Reasoning needs heuristics, measures of certainty, and a control strategy in addition to definitional domain knowledge

Semantic search

Task: Users search for relevant documents or data
Assume: Documents can be tagged with ontology terms
User queries can also be mapped to ontology terms
Example: a query for “infection” -> Infection [ontology term] and returns documents tagged with this term or a subclass of it.

The edges in the ontology graph can be followed to expand the query to find other (hopefully) relevant documents.
Types of Ontologies

Before looking at some ontologies, review how they can be organised into types:

- Knowledge representation ontologies
  - OWL: Class, ObjectProperty, DatatypeProperty level
- Top-level ontologies
  - Cyc upper-level: Thing, Stuff and Object
  - In a future lecture: DOLCE
  - Sowa’s ontology, SUMO
  - Smith’s Basic Formal Ontology
- Domain-specific ontologies
  - Gene Ontology
  - Foundational Model of Anatomy
  - Enterprise Ontology

The Cyc Ontology

- Cyc is a very large KBS, based on a commonsense ontology
- The Cyc project started in 1984 and is still running
- The Cyc system includes both the ontology and facts (instances and assertions about them)
  - It contains 100K concepts and 1M rules and axioms
  - Formalised in first-order logic
  - Supported by specialised inference algorithms
  - ‘formalised commonsense’ or fundamental human knowledge
    - An object cannot be in two places at one time
    - Animals live for a finite unbroken period of time
    - Washington DC is a City
- Cyc is now developed commercially by Cycorp: www.cyc.com
  - An open source version is available: www.opencyc.org
    - This version contains 6000 concepts plus 60,000 assertions
  - Extensive tutorial material

“Common Sense Reasoning - From Cyc to Intelligent Assistant”
http://www.cyc.com/cyc/technology/pubs

Knowledge Representation Ontologies

- The classes and relations in an ontology can themselves be organised into a knowledge representation ontology
  - Sometimes called a meta-ontology
  - The names of classes become instances
  - The names of relations become instances also

The Cyc Ontology

- Cyc is applied to:
  - Database integration - making use of Cyc’s ontology as an interlingua
  - Intelligent search - retrieval of images and film clips
  - Natural language understanding
    - ontology concepts are linked to their linguistic form
    - grammars etc are defined
    - alternative parses of sentences are disambiguated using the semantics in the ontology
    - logical expressions are constructed and used as queries or assertions
      “Fred saw the plane flying over Zurich.”
      “Fred saw the mountains flying over Zurich.”

Clickable map of the ontology:
http://www.cyc.com/cyc/technology/whatiscyc_dir/whatscycknow
The Cyc Ontology

The Cyc knowledge base is structured into Microtheories, these are sets of sets of facts, and extensions to the ontology.

- Microtheories are organised in a graph
- Facts are inherited

![BaseKB](the upper-ontology)

Topics covered by the Cyc ontology:

<table>
<thead>
<tr>
<th>Fundamentals</th>
<th>Clothing</th>
<th>Transformations</th>
<th>Biology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top Level</td>
<td>Weather</td>
<td>Changes Of State</td>
<td>Chemistry</td>
</tr>
<tr>
<td>Rule Macro Predicates</td>
<td>Geography</td>
<td>Transfer Of Possession</td>
<td>Physiology</td>
</tr>
<tr>
<td>Time and Dates</td>
<td>Paths and Traversals</td>
<td>Movement</td>
<td>General Medicine</td>
</tr>
<tr>
<td>Spatial Relations</td>
<td>Transportation</td>
<td>Parts of Objects</td>
<td>Materials</td>
</tr>
<tr>
<td>Quantities</td>
<td>Information</td>
<td>Composition of Substances</td>
<td>Waves</td>
</tr>
<tr>
<td>Mathematics</td>
<td>Perception</td>
<td>Agents</td>
<td>Devices</td>
</tr>
<tr>
<td>Microtheories and Contexts</td>
<td>Agreements</td>
<td>Organizations</td>
<td>Construction</td>
</tr>
<tr>
<td>Groups</td>
<td>Linguistic Terms</td>
<td>Actors</td>
<td>Financial</td>
</tr>
<tr>
<td>&quot;Doing&quot;</td>
<td>Emotions</td>
<td>Roles</td>
<td>Food</td>
</tr>
</tbody>
</table>

America in the 19th Century e-business

The top-level of the Cyc Ontology is structured (in part) by opposing concepts:

- The top concept is Thing
- Collection -- Individual
  - Collection is the class of all Cyc collections (classes). Cyc collections can be thought of as sets, having subsets etc. but are not to be equated as mathematical sets can be equated (by having the same elements).
  - Individual is something that is not a set.

- Intangible - things that are not physical or made of matter (nor encoded in matter).
- Intangible and Individual - things that are both Intangible and Individual, things that lack mass, volume, colour etc. Examples: ideas, algorithms, integers, distances.
- PartiallyIntangible - things with an intangible component but which exist in time, e.g. the laws of Texas, your bank account
- CompositeTangibleAndIntangibleObject - things with tangible and intangible components, e.g. people have bodies and minds, music on a CD, text in a book. These objects exist in time. Subclasses are Agent, PublishedMaterial, Sculpture.

Events are things that ‘happen’. Events change the state of the world.

- Scripts are composite events
  - part-of relation
  - Buying
  - Choosing
  - -Paying

Stuff -- Object

- Stuff is the category of things that can be subdivided spatially or temporally and the resulting portions are a still instances of the same class
  - Mass nouns
    - e.g. water, wood, walking
  - Object is the category of things that if they are divided, the pieces are not of the same category
    - Count nouns
      - e.g. book, bus, leap year

- Stuff and Object do not form a partition
  - ExistingObject: temporally stuff-like (existing through time) but spatially object-like, e.g. a book
  - ExistingStuff: temporally and spatially stuff-like, e.g. a piece of wood

The Gene Ontology

- The Gene Ontology (GO) was proposed as a tool for the unification of biology by the Gene Ontology Consortium
  - www.geneontology.org

Molecular biology generates large volumes of data, and many diverse research groups are involved

- Ashburner et al. note that “progress in the way biologists describe and conceptualise the shared biological elements has not kept pace with sequencing” (2000)
- Gene sequences for yeast, worm, mouse, human are known, and held across many databases
  - a gene can be described by the amino acid sequence, e.g.
  - Sequences can be compared using algorithms, but to share information about what a gene does (its functions and the processes it is involved in) it is necessary to define a shared vocabulary
The need to share descriptions (annotations) of experimental findings came from research - the realisation that organisms were similar in terms of the functions performed by comparable (orthologous) genes.

- The roles of many worm genes can be inferred from the roles of the equivalent gene in yeast.

This gives the opportunity to automatically transfer knowledge about one organism (e.g. one of the model species) to another. The uses include agriculture and healthcare.

At present, the Gene Ontology is:
- Manually curated
- Updated and released daily
- Used in many data analysis and visualisation tools (1000’s of papers)

GO is a directed acyclic graph (DAG) of terms:
- Terms should be class-level
- Two relationships:
  - `is-a` (subClassOf)
  - `part-of`

GO comprises three ontologies (Nature 2000):
- Biological process ontology
  - “Biological process refers to a biological objective to which the gene or gene product contributes. A process is accomplished via one or more ordered assemblies of molecular functions. Processes often involve a chemical or physical transformation.”

- Molecular function ontology
  - “Molecular function is defined as the biochemical activity ... of a gene product. This definition also applies to the capability that a gene product (or gene product complex) carries as a potential. It describes only what is done without specifying where or when the event actually occurs.”

- Cellular component ontology
  - “Cellular component refers to the place in the cell where a gene product is active. These terms reflect our understanding of eukaryotic cell structure. As is true for the other ontologies, not all terms are applicable to all organisms; the set of terms is meant to be inclusive.”

Annotation of genetic data about cells is the main role of the ontology.
- A gene (or gene product) may have a process, function and location assigned independently.
- These aspects are complementary and non-overlapping.

GO’s top-level categories are domain-specific:
- no "entity" or "object" classes

- GO:0008150 biological process (10593)
- GO:0007610 behavior (2414)
- GO:0000044 biological process unknown (28719)
- GO:0009897 cellular process (38756)
- GO:0017275 development (16478)
- GO:0007582 physiological process (70981)
- GO:0007879 regulation of biological process (14629)
- GO:0016032 viral life cycle (225)
- GO:0005575 cellular_component (98453)
- GO:0005623 cell (71949)
- GO:0043190 ATP-binding cassette (ABC) transporter complex (78)
- GO:0005933 bud (273)
- GO:0000267 cell fraction (1768)
- GO:0042995 cell projection (448)

Genes
Gene Ontology

- Yeast gene
- PR5
- 5.9e-05
- FlyBase
- frutti gene

- ATP-binding cassette
- cell
- cellular_component

- GO annotations
- DNA replication, DNA repair, transcription

- GO:0003673 Gene Ontology (14821)
- GO:0008744 molecular function (9507)
- GO:0017275 biological process (16369)
- GO:0016032 cellular component (14323)
Goals of the GO Consortium
- To compile a comprehensive structured vocabulary of terms describing different elements of molecular biology that are shared among life forms
  » Terms may have synonyms (broader/narrower)
- Separate vocabularies
- To describe biological objects using these terms
- To provide tools for querying and for curation
- Not to unify databases; dictate a standard; define evolutionary relationships

GO's design principles (ontologies are considered to be DAGs)
- A child term may have multiple parents through:
  » is-a "a child is an instance of parent", but really is-a means subclass; and
  » part-of "refers to when a child is a component of the parent".
- Terms have unique identifiers GO:0000235
- Definitions - all terms should be defined, preferably from a standard reference
  » Oxford Dictionary of Molecular Biology
  » Only through precise definitions can annotation across communities succeed, as often the same terms are used with different meanings

Working principles
- All paths must be true (the true path rule)
  » the is-a and part-of relations to the immediate parents, and from there up to the root, must be true
  » if not, the hierarchy may be restructured
- Terms should not be species-specific
  » Include terms that apply to more than one taxonomic class of organism
  » Words may have different meanings in different species “(sensu)”
- Terms should have class-level coverage
- All attributes must be accompanied by appropriate citations
- All annotations must incorporate controlled statements of evidence as well as citations

Annotation
- Resulting from an experiment, a gene is given 0 or more GO Process / Molecular Function / Cellular Component terms as an annotation
- Each aspect is assigned independently
- The most specific term that correctly describes the biology should be chosen
  » High-level terms may be used
- ‘unknown’ may be used - indicates someone has tried and failed to make an annotation
- The annotation should reflect normal activity and location
  » Functions observed only in mutants are not usually included
  » Viral processes are annotated (although these are abnormal for the host)

Evidence codes (provenance)
- These are selected from a controlled vocabulary, including:
  » TAS: traceable author statement
  » IDA: Inferred from direct assay
  » IEA: Inferred from electronic annotation

GO has several syntaxes, including OWL
- GO flat file syntax
  » Indentation indicates subclass (%) or part-of (<)
- OBO flat file syntax (http://www.geneontology.org/GO.format.html)
  » The ontology file begins with a <header> followed by <stanza>s
  » Each term is defined by a stanza:
    [Object type]
    <tag>: <value>
    <tag>: <value>
  » Tags include: id; name;(required tags) plus def; is-a; relationship;

Evidence code: [Term]
- id: GO:0003674
- name: molecular_function
- def: "The action characteristic of a gene product." [GO:curators]
- namespace: biological_process
- [Term]
- id: GO:0000018
- name: regulation of DNA recombination
- namespace: biological_process
- def: "Any process that modulates the frequency, rate or extent of DNA recombination", the processes by which a new genotype is formed by reassortment of genes resulting in gene combinations different from those that were present in the parents." [GO:curators, ISBN:0198506732]
- is_a: GO:00015052
- relationship: part_of GO:0006310
Gene Ontology

- Conversion to OWL-DL requires part-of to be expressed as value restriction or as an existential commitment: \( \exists \text{part-of}.A \) or \( \forall \text{part-of}.A \)
- Problem: GO’s definitions are largely informal
  \[ A \sqsubseteq B \sqcap E \quad [B \text{ part-of } A \text{ and a subclass of } E] \]
  \[ A \sqsubseteq C \quad [C \text{ is part-of } A \text{ but has no further explicit information}] \]
  - B must be defined by \( E \sqcap \exists \text{part-of}.A \)
  - C must be defined by \( \exists \text{part-of}.A \)
  - The definitions of B and C are very similar - more information is needed
  - Classification may give errors
- One solution, a term such as ‘Carbohydrate Metabolism’ can be analysed:
  Carbohydrate Metabolism \( \sqsubseteq \) Metabolism \( \sqsubseteq \) actsOn.Carbohydrate \( \sqsubseteq \) actsOn.Carbohydrate

Anatomy Ontologies

- GO describes molecular functions at the cellular level
  - Purpose and objectives defined?
  - Fit for purpose?
  - Is it a ‘shared specification of a conceptualisation’?
- Medical applications need a representation at a larger granularity - anatomy - example applications:
  - GALEN, a project to capture descriptions of surgical procedures across Europe. Knowledge of the procedures and human anatomy was captured (in a DL known as GRAIL).
  - Digital Anatomist, developed to support the teaching of anatomy by the University of Washington
    - A large knowledge base of symbolic models and images
    - 2D and 3D images, annotated with anatomical terms
    - Quiz mode - the user tries to select the correct object
    - Further applications include clinical systems, and biomedical research
    - Based on a conceptual framework: Foundational Model of Anatomy
      http://www.digitalanatomist.com/

Anatomy Ontologies

- Foundational Model of Anatomy (Rosse, C. University of Washington)
  - Aims to support the development of intelligent medical systems
  - This requires a richer representation than that provided by medical terminologies such as the UMLS (Unified Medical Language System)
    - These have a limited capacity for inference or for consistency checking.
    - The FMA intended to be a sharable resource,
    - providing deep anatomical knowledge.
- The FMA is developed following disciplined modelling procedures
  - Implemented using Protégé
  - High-level scheme for concepts and relationships
    - Subsumption
    - Definitional attributes (relations)
  - Anatomy Taxonomy: the class hierarchy
  - Anatomical Structural Abstraction: the structural (part-of) relations

Anatomy Ontologies

- **Anatomy Taxonomy**
  - The upper levels of the ontology have a clear organisation, and are used to structure the more specific classes.

- **Modelling principles**
  - Unified context principle
    - Structural view, not functional, surgical, or radiological
  - Abstraction level principle
    - Canonical anatomy (the idealised structure)
  - Species specificity
    - Human
    - But may serve as a framework for other vertebrates

---

**Examples**

- **Organ**
  - is an **Anatomical structure** which **consists of** the maximal set of **Organ parts** so **connected** to one another that together they **constitute** a self-contained unit of macroscopic anatomy morphologically distinct from other such units.

- **Organ Part**
  - is an **Anatomical structure** which **consists of** two or more types of **tissues** **spatially related to** one another in patterns determined by coordinated gene expression; together with other contiguous organ parts it **constitutes** an Organ.

---

**Modelling principles**

- Definition principle
  - The defining relations (attributes) of a class should be specified in terms of the physical and other structural relations
- Dominant concept principle
  - **Anatomical Structure** is the dominant class, the reference for other definitions
- Content constraint (i.e. granularity)
  - Largest structure is the Body (the whole organism)
  - Smallest structure is macromolecule
- Relationship constraint, 3 relationships should be modelled:
  - Class subsumption
  - Static physical relationships (parts)
  - Relationships describing transformation (development)

---

**Aristotelian Definitions**

- Dictionary definitions define all meanings of a term
- but in this ontology, definitions capture the essence of the Concept
- 10 requirements must be satisfied, primarily: **genus** and **differentiae**

---

**Foundational Model of Anatomy**

- A “specification of a conceptualisation”, both formally and informally
- Product of a group, rather than a community
- Deep model that supports reasoning (cf. annotation)
Skills Management in an insurance company (Swiss Life)
- To improve the sharing access and reuse of knowledge in formal, semiformal and informal notations
- Produce a system to aid the staffing of projects and career development

Method
- Feasibility study following CommonKADS
- Knowledge acquisition was performed at two workshops involving three departments
  - Private insurance, human resources, IT
  - Three hierarchies were constructed covering these three domains
  - Issues: granularity, representation as a class or an instance?
- A web-based application was built and deployed, gaining 200 users
- Employees grade their skill levels themselves
  - Basic knowledge; Practical experience; Competency; Top specialist;

Lessons
- Communication issues: terms such as 'customer' and 'policyholder' had different meaning to different groups
  - Different legislation applies to different types of insurance
- Large ontologies are difficult to browse
  - Define 'user views'

Ontology-enhanced searching
- Cyc has been used for semantic search
  - An experimental version was launched by Yahoo
- Annotation with ontological concepts
  - Search for “JAVA”
    - Java(island)
    - Coffee(drinks)
    - Java(programming language)
  - Uses the Cyc ontology

Documents need to be indexed, this can be done manually, but automation is needed in many cases
- The words in the document may not correspond to the name of the ontology concept
  - Thing, AnatomicalEntity
  - But for many concepts, the name and the corresponding noun phrase will be similar
- If a set of documents are tagged manually, machine learning can be used to learn the term → feature set (e.g. bag of words) association
- More complex procedures for deriving feature sets for documents, matching these to concepts, and analysing which features are positive or negative indicators of a concept

Doug Lenat on Cyc and information retrieval
http://video.google.com/videoplay?docid=-7704388615049492068
One case study on ontology-enhanced search in an energy company (EnerSearch, a virtual organisation providing research material to the energy industry in Europe), had the evaluation goals:

- Users make fewer mistakes, i.e. answered the questions posed to them incorrectly - found to be true
- User find information more quickly - partial evidence for this

In a case study for the Norwegian oil industry, ‘ontology profiles’ were shown to improve search performance in comparison with simple keyword search

- Users scored the top ten search results, the average score for a small number of queries by a small number of users was better for the ontology-supported method (7/10 vs 5/10)

Ontologies establish shared conceptual models

- This is seen as key to enabling semantics-driven knowledge processing

Two problems arise

- Coping with multiple ontologies - as one single ontology may not be imposed
- Coping with changing requirements

Individual departments may accumulate information in their own systems, rather than contribute to a central KM system

- Reuse of this knowledge proves difficult as the original source model/ontology may differ from that required in a new context
- Some solutions:
  - Include the source model/ontology completely, all definitions are included in the new ontology
  - Map the source ontology to the target ontology - this is a more complex solution that maps a selected subset of classes

Ontologies evolve to meet changing conditions

- Concepts may be added, deleted, redefined, split...
- Changes need to be propagated (instances may be reclassified)
- Validation is important as users might not understand the implications of a change

Maedche, A. et al. IEEE Intelligent Systems March/April 2003

The Enterprise Ontology - developed to support methods and tools for enterprise modelling (1997)

- Product of a collaboration: IBM, Lloyds Register, Logica, Unilever, AII
- Enterprise modelling aims to take an enterprise-wide view of an organisation that can then be used as the basis for making decisions, some important goals:
  - Integration of views
  - Communication between people
  - Flexibility
  - Support users
- A shared understanding aids integration and planning, hence the EO

Role of the Enterprise Ontology

- To act as a communication medium
  - Between people and computers; computers and computers
  - To assist the acquisition of enterprise-related knowledge
  - To assist the organisation of libraries of knowledge
  - To provide explanations of rationale

The ontology was intended to support more specific enterprise models.

- To achieve stability, the ontology was driven by factors important to a business, independently of any implementation.
  - Stability promotes inter-operation of systems
- Ontology terms were to be an interlingua (shared representation) that would allow tools to inter-operate
- Both informal and formal versions of the Enterprise Ontology were created
- Scope - core concepts in the following work areas were defined:
  - Activity; Marketing; Organisation; Strategy; Time
  - Terms were chosen to reflect natural English usage
  - Definitions were produced - not dictionary definitions but normative definitions
  - Definitions were made precise through the use of a small number of building blocks (language primatives or meta-ontology)
  - Definitions were formalised in FOL
- Resulting in around 90 top-level concepts
Enterprise Ontology

Some example concepts:

- **Organisation work area**
  - Legal Entity: has rights and responsibilities, including legal rights
  - Organisational Unit: a recognised unit within an organisation (no legal status)
  - Person and Corporation: subclasses of Legal Entity
  - Machine: a non-human non-legal entity capable of playing certain Roles
  - Ownership: of rights and responsibilities may only lie with Legal Entities
  - Resources: rights and responsibilities for these may be assigned to Organisational Units
  - Non-legal Ownership
  - Management Links: represent the management structure
  - Manage: represents assigning Purpose to Organisational Units

Additional examples

- UNSPSC: global multi-sector standard for efficient, accurate classification of products and services
  - Ontologies for Corporate Web Applications. Leo Obrst, Howard Liu, Robert Wray

- ISO15926 Lifecycle data for process plants
  - An upper ontology based on ISO 15926. Rafael Batres et al.
    doi:10.1016/j.compchemeng.2006.07.004
    See also: https://www.posccaesar.org/wiki/ISO15926

Example ontologies

- Choose one of those covered in the lecture (or another that is well described in the literature, e.g. law)

- Ensure you can describe:
  - Area covered (scope), aims, level of granularity
  - Applications
  - Method of development
  - Degree of formalisation