MANCHESTER 1824

The University of Manchester

Biomedical ontologies / Terminologies

Alan Rector
School of Computer Science / Northwest Institute of Bio-Health Informatics
rector@cs.man.ac.uk

www.co-ode.org

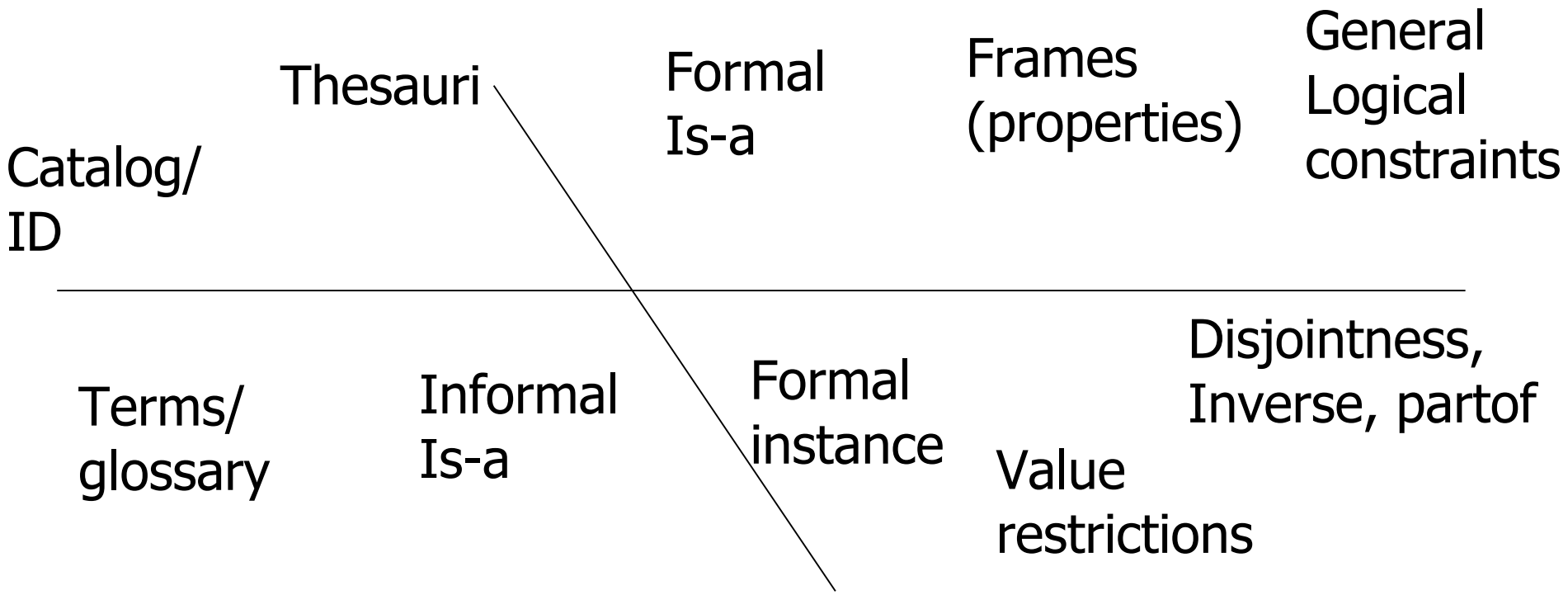
protege.stanford.edu

clinical.escience.org

NIBHI

[Deborah McGuinness, Stanford]

So what is an ontology?



THE UNIVERSITY of MANCHESTER

University of Southampton

My definition of an ontology

- Short version:
  "a representation of the shared background knowledge for a community for use in software"
- Long version:
  "an implementable model of the entities that need to be understood in common in order for some group of software systems and their users to function and communicate at the level required for a set of tasks"
- Alternative Version
  "a symbol system for modelling the definitions and descriptions of the entities represented in an information system"
- ... and "it doesn't make the coffee"
  Just one of at least three components of a complete system

What's it for?

- Re-use and sharing
  - Common standards
  - Database integration / schema fusion
  - Common metadata for annotation
  - Maintaining large terminologies
    - "A terminology compiler"
  - Re-usable "value sets"
- Highly declarative representations for software
  - Fractal context and detail
  - More constrained and more general than standard Model Driven Architecture
    - but less mature

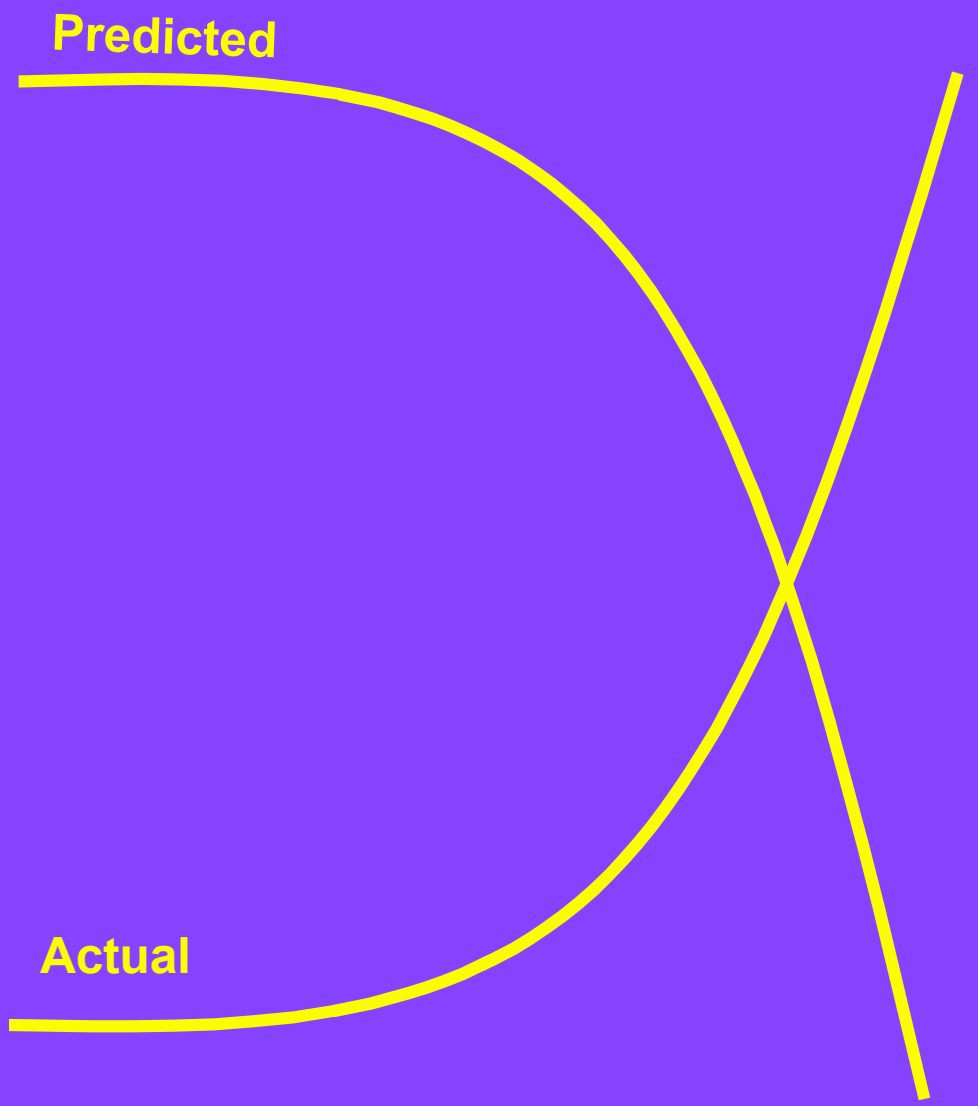
What's it for...
Tasks for Semantic Webs & "ontologies"

- Cataloguing, naming, describing
  - "Value sets" - "what can I say about this?"
- Database federation and Schema Unification
  - "What does this entity mean in terms of some other schema?"
- ***Semantics Rich Software and Querying***
  - What I mostly do
    - "Model driven architectures on steroids"
      - for fractal knowledge
    - Adaptable user interfaces
      - for fractal adaptation
- Annotation, Navigation and Information Retrieval
  - What most biologists mostly do with them
- Discovery of Resources
  - What web services mostly do with them
- Linguistic processing
  - Social markup and authoring - the current hot topic

The scaling problem: The combinatorial explosion

- It keeps happening!
  - "Simple" brute force solutions do not scale up!

- Conditions x sites x modifiers x activity x context
  - *Huge number of terms to author*
  - *Software CHAOS*

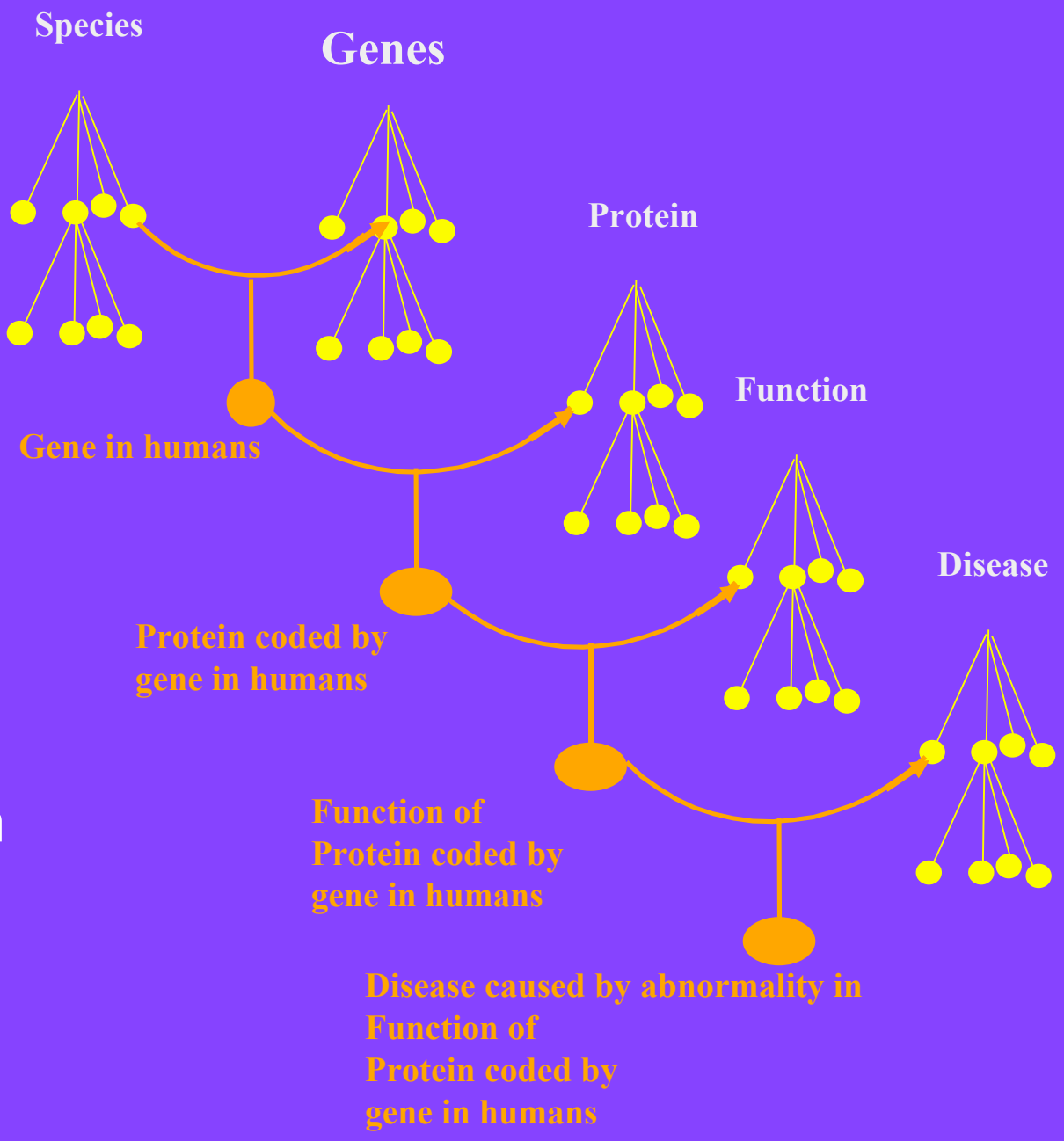


Logic as the clips for "Conceptual Lego"

"***SNPolymorphism*** of ***CFTRGene*** causing ***Defect in MembraneTransport*** of ***Chloride Ion*** causing ***Increase*** in ***Viscosity*** of ***Mucus*** in ***CysticFibrosis***…"

"***Hand*** which is ***anatomically normal***"

Build complex representations from modularised primitives



OWL and Description Logics

- Web Ontology Language (OWL)
  - W3C standard
    - one layer up from RDF(S) in the “web stack”
- Based on description logics
  - An abstract syntax, two XML serialisations, a user oriented syntax,...
- Three variants
  - OWL Lite, OWL-DL, and OWL Full
    - and now OWL 1.1
- See

  http://www.co-od.org

  http://www.w3.org/2004/OWL/
  http://www.w3.org/2001/sw/BestPractices/
  http://www.w3.org/2001/sw/BestPractices/OEP/
-

Example

- Pneumonia = Infection THAT *has_locus* SOME Lung
- Pneumonia = Inflammation THAT
  has_cause SOME Infection AND
  has_locus SOME Lung
- Pneumonia = Inflammation THAT
  has_cause SOME Infection AND
  has_locus SOME
  Parenchyma THAT *is_constituent_of* SOME Lung
- Pneumonia = Inflammation THAT
  has_cause SOME Infection THAT
  has_locus SOME
  Parenchyma THAT *is_constituent_of* SOME
  (Lung OR *is_subdivision_of* SOME Lung)

Generated text for presentation and quality assurance

- Pneumonia = Inflammation THAT
  has_cause SOME Infection THAT
  has_locus SOME
  Parenchyma THAT *is_constituent_of* SOME
  (Lung OR *is_subdivision_of* SOME Lung)
- "Pneumonia is any inflammation of the parenchyma of the lung that is caused by an infection"
  - GALEN version
    - Sadly no OWL tools yet do this well
    - Does require recording linguistic information in each language
      - but because it is a building block set this is not an onerous task
        - » GALEN: 2 weeks to first version, 2 months to full version
    - Current experts are Open University

...and now a word from our sponsor

Protege4 is now available!

- Now the recommended editor for OWL ontologies if using Protege (except for simultaneous multi-user editing)
  - Despite being called alpha it is now more stable than 3.3beta
  - Complete rewrite with
    - Manchester Syntax & greatly improved UI
    - Direct loading of OBO and KRSS (SNOMED) files
    - Easy modules and importing
    - Very fast classification Small files <1sec. NCI Thesaurus <2min
    - Ontology queries
    - Web publishing for browsers and (RSN) Web Services
    - OWL 1.1 -
      - numeric ranges and more general annotation and cardinalities
    - OWL oriented API and plugin framework
      - Much easier to implement applications
      - Now joint with Swoop/Pellet OpenOWLTools.org - developing rapidly
- http://protege.stanford.owl _ downloads _ Protege4Alpha http://www.co-ode.org downloads - for more plugins
  - If you have problems contact us and/or mailing list

Medical and Ontologies and Standards

- Established clinical data standards
  - **"...or you don't get paid"**
    - **ICD9-CM - (International Classification of Diseases)** http://icd9cm.chrisendres.com/
      - all US hospital data on diseases
        - » Also ICD10-ACM for Australia
    - **CPT - Clinical Procedure Terminology** http://www.ama-assn.org/ama/pub/category/11249.html
      - All procedures in the UK
    - **SNOMED-CT in UK** http://www.ihtsdo.org/_ http://www.snomed.org
      - Recently reorganised as an international standard by international subscrition
    - Various Continental European national standards
  - **"...or your system won't win the procurement"**
    - **LOINC - Lab data - Logical Observation Identifiers Names & Codes** http://www.loinc.org
    - **HL7 *V2* - clinical messaqing standard**
  - **...or Regulatory authorities won't accept your data**
    - **MEDDRA** - adverse drug reactions
    - ? SNOMED - certain data for FDA - new

UMLS Metathesaurus and PubMed

- Cross reference of all established clinical and many research terminologies / ontologies plus MeSH
  - Part of UMLS knowledge sources (UMLSKS)
    - http://umlsinfo.nlm.nih.gov/
- *Concept Unique Identifiers* (**CUIs**) & *Lexical Unique Identifiers* (LUIs)
  - The best common route to MeSH and PubMed
  - I wouldn't build a medical ontology that didn't expect map to CUIs and LUIs
    - But if you establish yourself, NLM will probably do at least part of the work
- Mapping to lexical / linguistic resources
  - UMLS Semantic Network
    - ≥200 High level categories that make sense to linguists but not to ontologists
- Free but must sign a complex license
  - Contains proprietary material that you must promise not to rip off
    - Maintained by US National Library of Medicine (part of NIH)
- See web site of ≥40 specialised terminologies indexed

Aspiring but slow to emerge (1)

- SNOMED-CT - Clinical Terminology
  - "Merged" from SNOMED (US) and Read Codes (UK)
  - Major momentum - major political force
  - Licensed to countries but not really open source
    - International Health Terminology Standards Development Organisation (IHTSDO)
      - Originally a for-benefit development of the College of American Pathologists
    - UK, US, Australia, Denmark, Sweden, Netherlands (?), Canada (?)... Not France
  - Huge volume, broad coverage, low quality
    - Best for primary care & pathology
      - Typical applications outside core areas find 25%-50% of what they need
    - Many systematic and sporadic problems
      - Real standardisation is coming in the development of "subsets"
        - » A bottom up use
    - Main significance is probably the identifiers, versioning, and update mechanism

Aspiring but slow to emerge (2)

- HL7 **V3** - Clinical messaging - interoperation of clinical systems - mostly information models
  - A successor to HL7 v3
    - The “Reference Infomation Model” (RIM)
  - “Structural codes”
    - The vocabulary that holds the model together
  - External codes
    - Payload information from other terminologies - ICD9-CM, SNOMED-CT,...
  - Over-engineered and complicated, but gaining traction
    - Basis of BRIDG
      - Clinical Trials information model

Translational

- **National Cancer Institute (NCI) Thesaurus**
  http://www.nci.nih.gov/cancerinfo/terminologyresources
  ftp://ftp1.nci.nih.gov/pub/cacore/
  - Part of CaBIG/CaCORE
    - Three tiers
      - Ontology/vocabulary, clinical data elements, data models
        - » CADSR - Registry of data elements to ISO Metadata Standard (ISO 11179)
        - » Also mirrored at National Cancer Research Institute / CancerGrid UK
- **BRIDG**
  https://cabig.nci.nih.gov/inventory/infrastructure/bridg/
  http://gforge.nci.nih.gov/projects/bridg-model/
  - Common information model for clinical trials
    - Based on HL7v3; Ontology/Terminology role unclear
- **Clinical Trials**
  - OCRe - Ontology of Clinical Research
    - Recently established between TrialBank in US and CancerGrid in UK
      - Still seeking its place in the ecology
  - OBO-Foundry - Clinical Trials Ontology (CTO)
    - Role unclear - political issues
- **OMIM**
  - Important de facto link between genomics and human disease
    - Cross linked in UMLS Metathesaurus

Others

- Public health / international Statistics
  - ICD 9/10 now developing 11
    - “International classification of diseases”
- Adverse event recording and drugs
  - MEDDRA
    - International standard and FDA and MCA required
- Primary Care
  - ICPC
- Nursing
  - ICNP
- ...

Available for mining and re-use

- OpenGALEN
  - General representation of medicine at a clinical level
    - http://www.opengalen.org
  - Open Source with tools
    - Available in OWL
    - Soon to be classifiable with newest classifier
  - But not actively maintained
    - Outcome of an EU project that ended in 1997

Aspiring OBO-Foundry - little impact yet in medicine, but time will tell

- OBI - ontology of biomedical investigations
  - Relation to Clinical trials, LOINC, SNOMED, and HL7 unclear
    - Barry has an anti-hl7 blog
- PATO - ontology of descriptors for phenotypes
  - Fulfills what seems to be a real gap
- CTO - Clinical Trials Ontology -
  - has alienated BRIDG, Trialbank, and other key potential users

Other

- EU Actions of which I know little
  - ACGT - advancing clinical trials in cancer http://eu-acgt.org/
    - Cancer trials
  - EHRH+G
    - Linking EHR and Genomic data
  - Semantic Health
- Other medical
  - Immunology, Infectious disease, Adverse drug reactions...
- Other general
  - Quantities and Units
    - Several but a common mechanism critical
- US National Center for Biomedical Ontologies
  - Umbrella project
  - http://www.bioontology.org/

Terminology / Ontology Servers

- Easy access, versioning, comparison, browsing, etc.
  - Usually with Browser and Web Services interfaces
- Alternative versions
  - NCI / Apelon DTS
  - Mayo CTS
  - Ocean Informatics Terminology Server
- Web resources
  - National Center for Biomedical Ontology
  - Protege 4 Web publishing