MANCHESTER 1824

The University of Manchester

Migrating from the Informal to the Formal Ontology: Costs and Benefits

Robert Stevens

School of Computer Science

University of Manchester, UK

(Robert.stevens@manchester.ac.uk)

MANCHESTER 1824

The University of Manchester

Introduction

- Two communities of ontology builders
- Shared understandings: For humans; for machines
- Both useful
- What does each provide?
- Moving towards the formal, computationally amenable version
- What are the costs?

MANCHESTER 1824

The University of Manchester

Acknowledgements

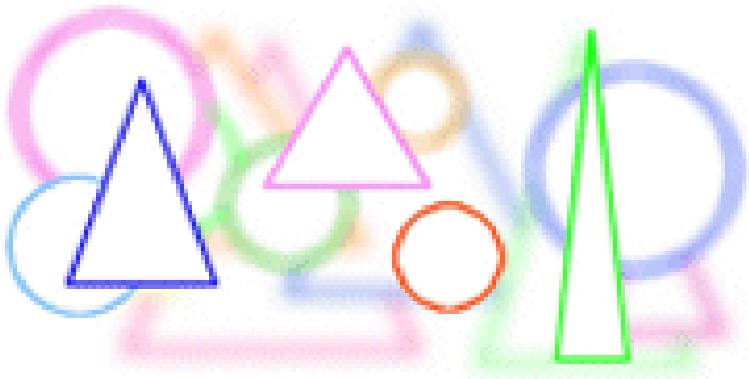
- Chris Wroe & Mikel Egana Aranguren
- Katy Wolstencroft
- Michael Ashburner, Jane Lomax, *et al*

MANCHESTER 1824

The University of Manchester

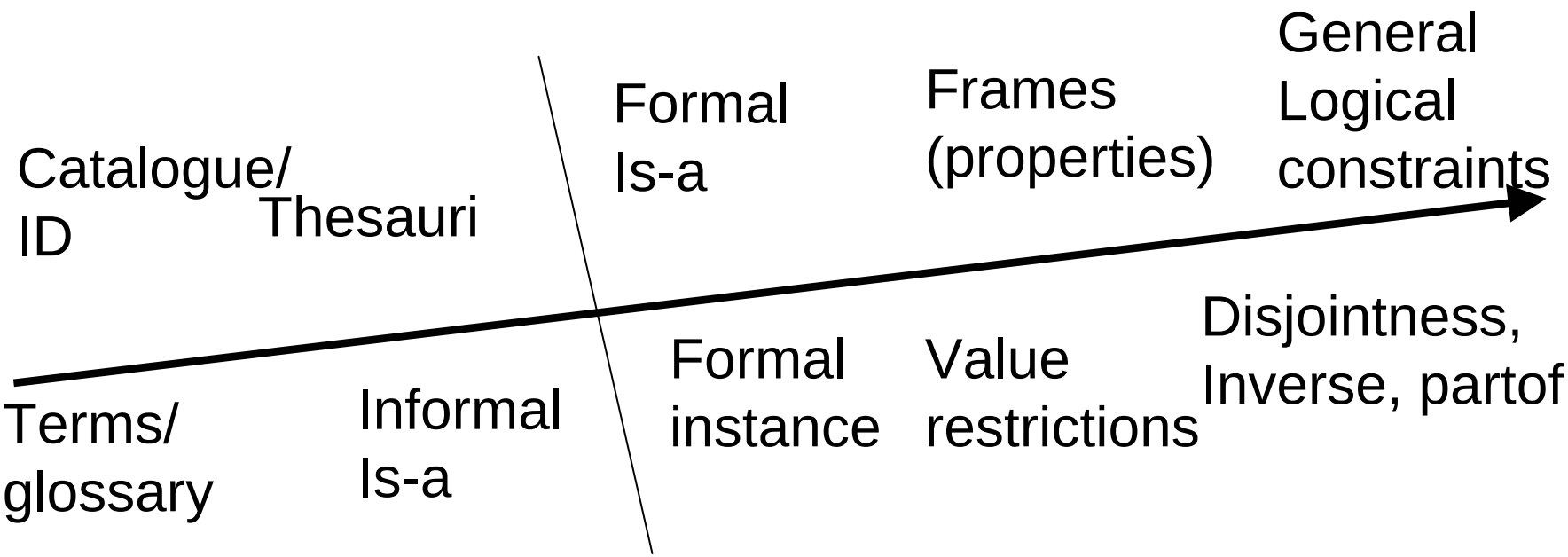
Practical questions

- what shapes of ontology exist?
- Which should I use for what purpose?
- What tools help build each shape of ontology?
- How can I change the shape of an ontology if requirements change?



[After Deborah McGuinness, Stanford]

Spectrum of ontology Shapes

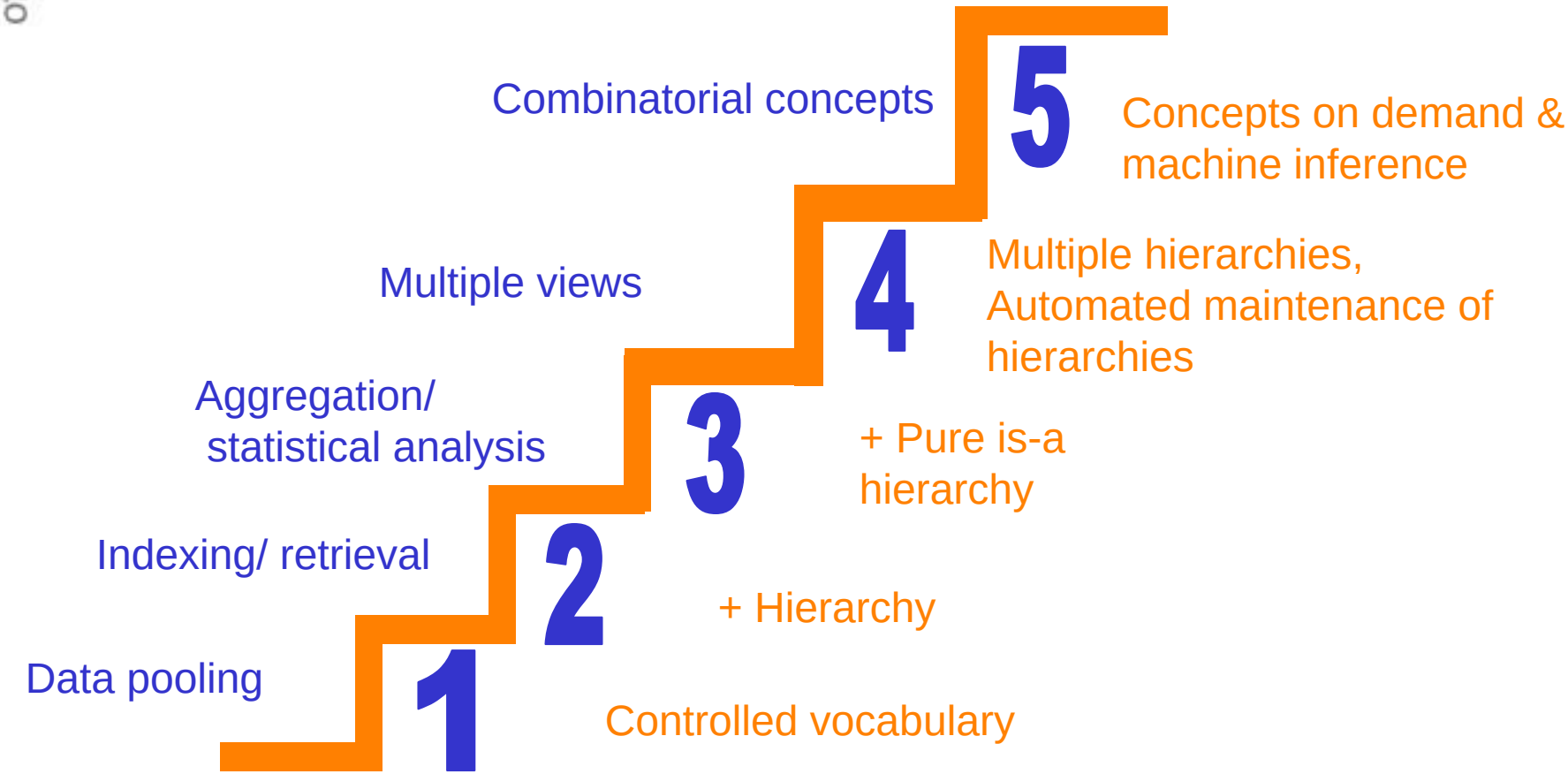


MeSH *Gene Ontology* *EcoCyc* *TAMBIS*

Mouse Anatomy *PharmGKB*

Common languages (OWL, OBO) and
Common tools Protégé, DAGEdit

From requirements to features: the feature escalator



Step 1: A *common vocabulary* for data pooling

www.godatabase.org

Gene Symbol		Datasource	Evidence	Full Name
ASA1		TAIR	TAS	None
ATR1		TAIR	IDA	None
BA1248	ATGCC / GOst	TIGR_CMR	ISS	None
F15D2.31		TIGR_Ath1	TAS	None
F24B9.11		TIGR_Ath1	TAS	None
F27K19.50		TIGR_Ath1	TAS	None
F4F7.39		TIGR_Ath1	TAS	None

Gene Symbol	Function
ASA1	tryptophan biosynthesis

Locus Name	Function
F15D2.31	tryptophan biosynthesis

Step 2. A *hierarchy* for navigation and retrieval

MeSH – Medical Subject Headings – annotating publications with terms from a thesaurus like structure for retrieval purposes

Should a search for documents dealing with A find all (or most) documents dealing with B?

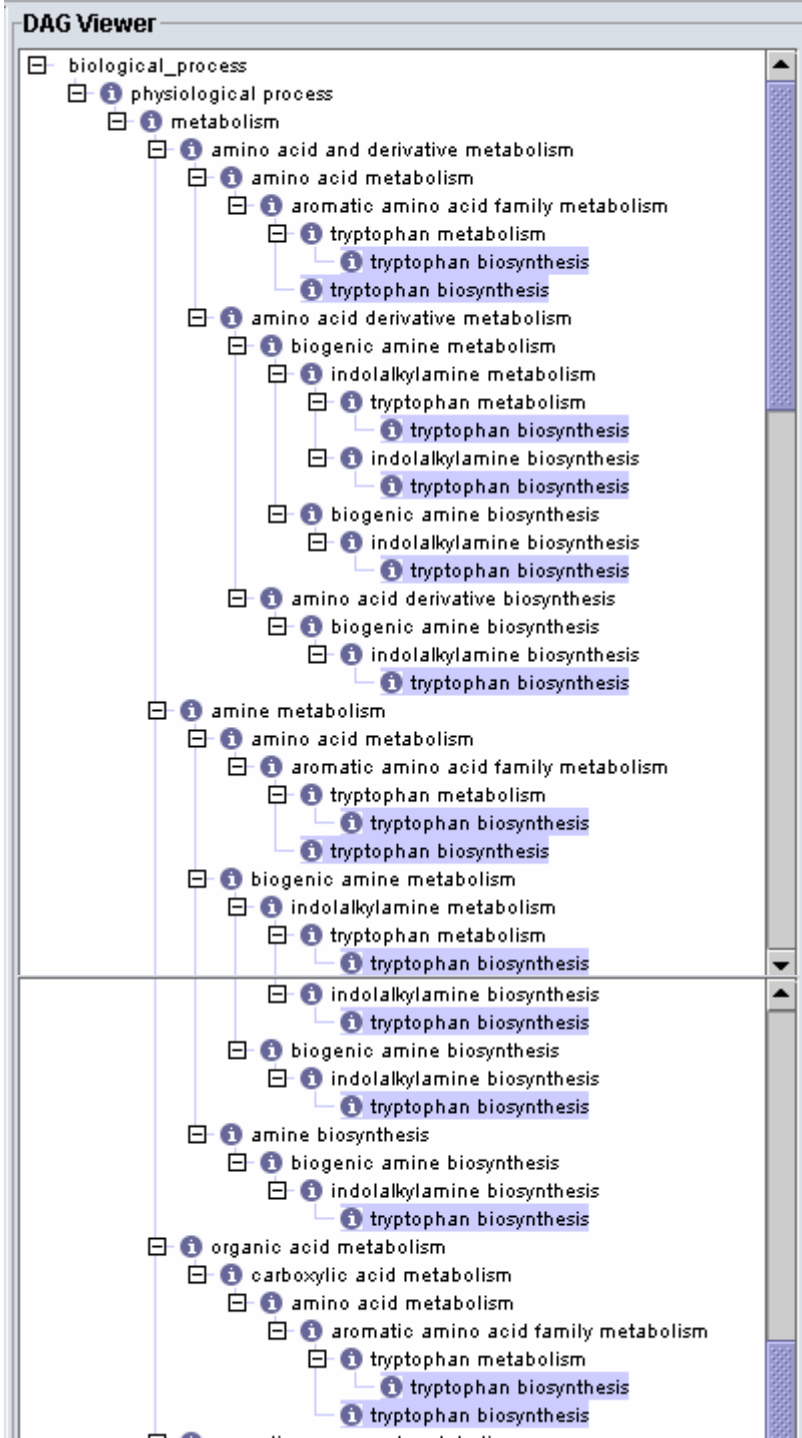
Step 3. A *pure subsumption hierarchy* for aggregation

ICD10: Modern mortality statistics

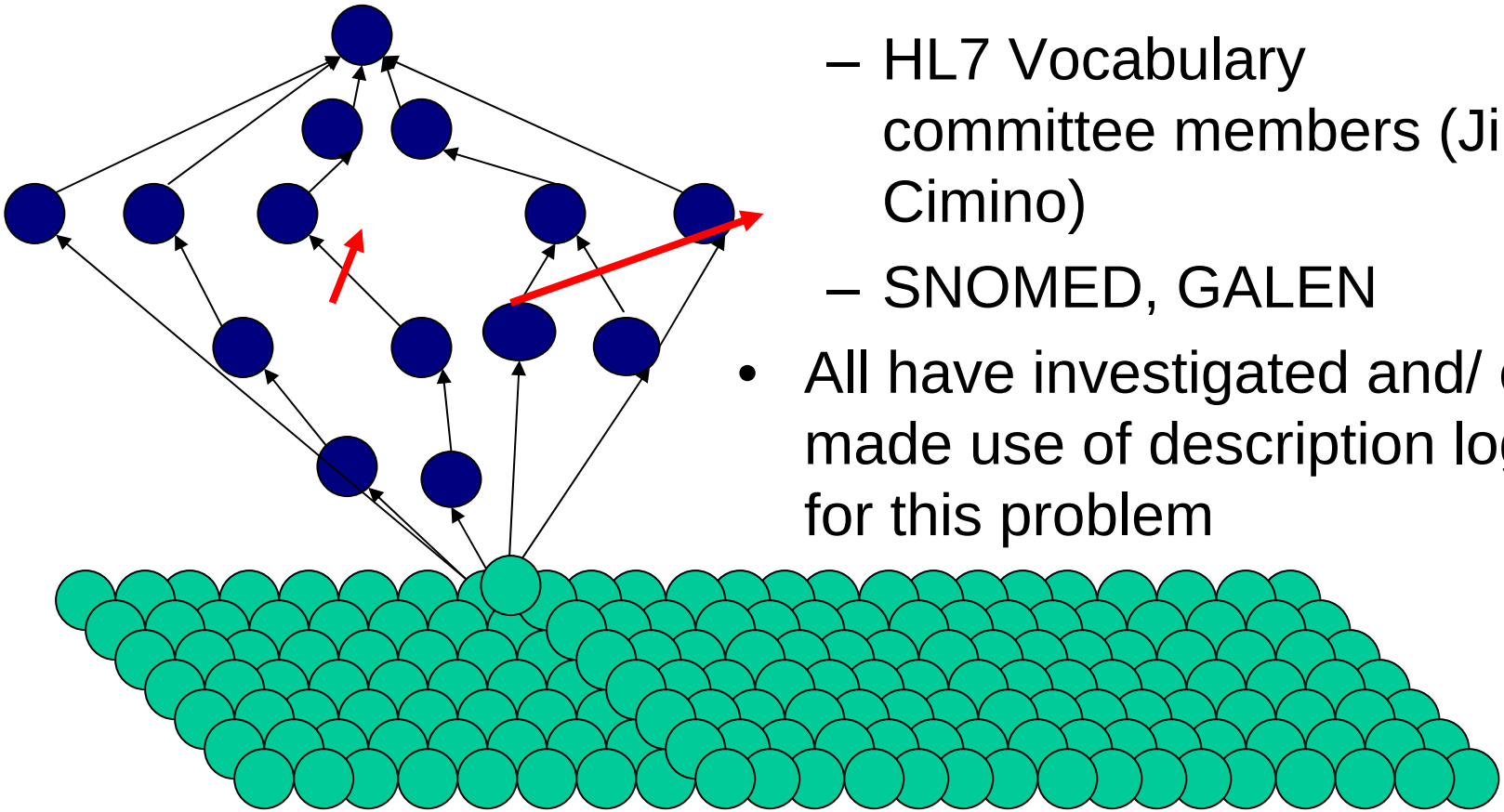
V01-X59	Accidents	
	V01-V99	Transport accidents
		V01-V09 Pedestrian injured in transport accident
		V10-V19 Pedal cyclist injured in transport accident
		V20-V29 Motorcycle rider injured in transport accident
		V30-V39 Occupant of three-wheeled motor vehicle injured in transport
		V40-V49 Car occupant injured in transport accident
		V50-V59 Occupant of pick-up truck or van injured in transport accident
		V60-V69 Occupant of heavy transport vehicle injured in transport accic
		V70-V79 Bus occupant injured in transport accident
		V80-V89 Other land transport accidents
		V90-V94 Water transport accidents
		V95-V97 Air and space transport accidents

Multiple views

- No one way to classify, especially if complex
- Concepts organised in different ways to suit different users
- Example
  - Metabolism type
  - Functional chemical classification
  - Compositional chemical classification
- More annotators more concepts

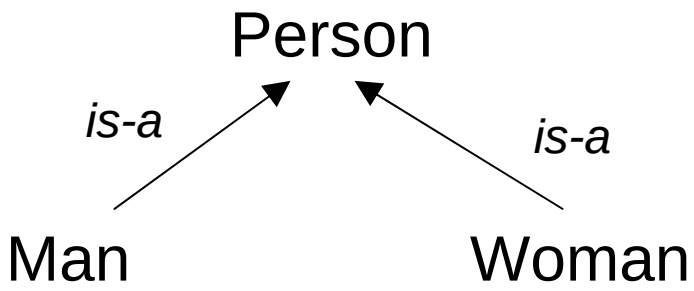


Step 4:*Formal definitions and reasoning* to support multiple hierarchies



- Its hard work! So says:
  - GO Curators
  - HL7 Vocabulary committee members (Jim Cimino)
  - SNOMED, GALEN
- All have investigated and/ or made use of description logic for this problem

What are we saying?



- Are all instances of Man instances of Person?
- Can an instance of Person be both a Man and an instance of Woman?
- Can there be any more kinds of Person?

What are we saying?

Man —*has-chromosome*→ Y chromosome

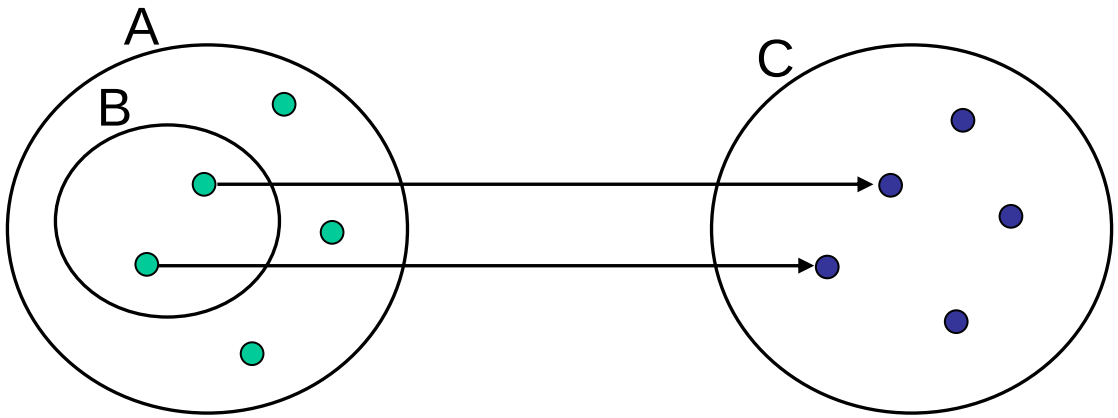
- What kinds of class can *fill* “has chromosome”?
- How many “Y chromosome” are present?
- Does their have to be a “Y chromosome”?
- What properties are *sufficient* to be a Man and which are simply *necessary*?

Man —*has-chromosome* 1→ Y chromosome
Man —*has-chromosome* 1→ X chromosome
Man —*has-chromosome* 44→ autosome

MANCHESTER 1824

The University of Manchester

OWL represents classes of instances



MANCHESTER 1824

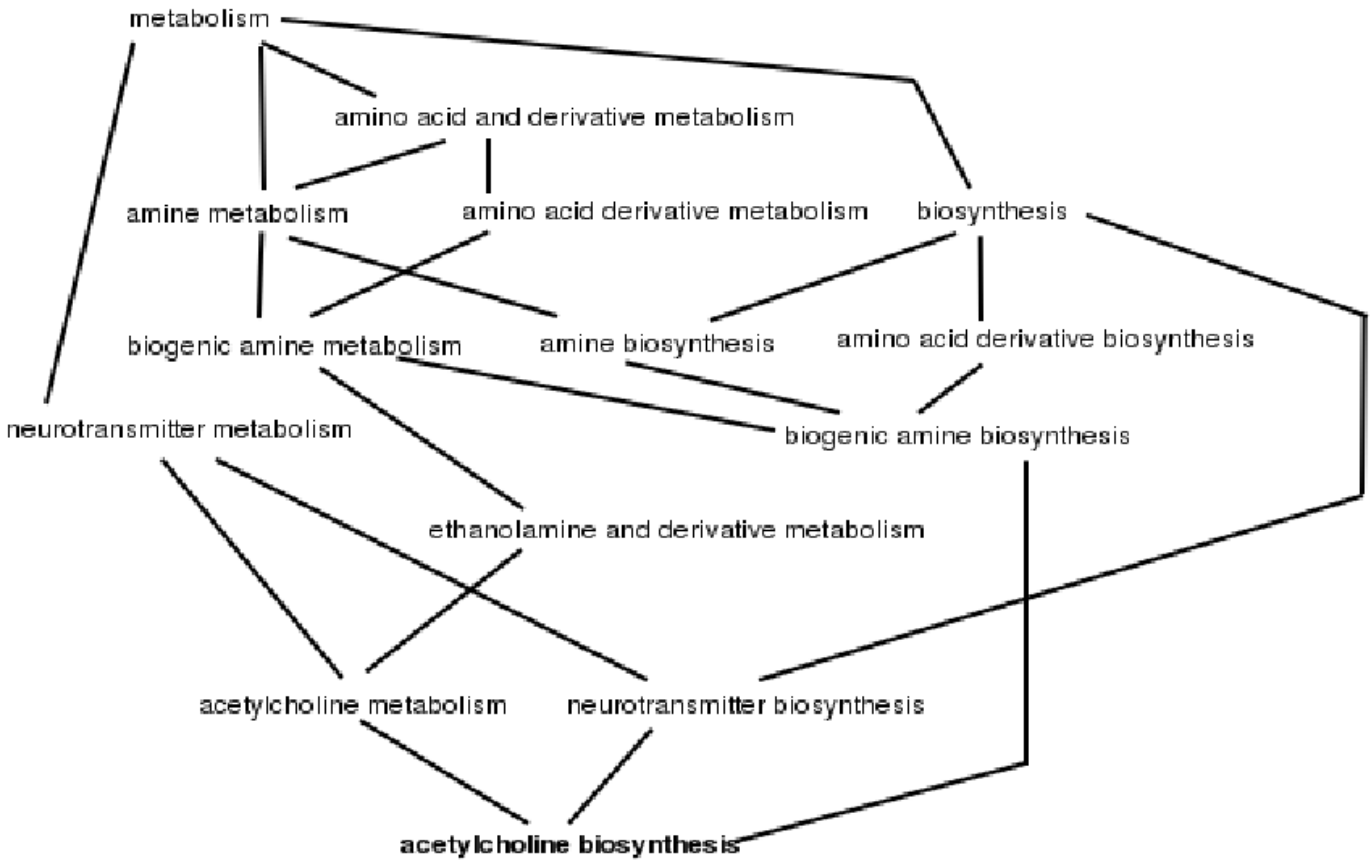
The University of Manchester

Migrating to OWL

- Often don't start from scratch
- In life sciences, many ontologies in DAG form
- Migrating towards full, explicit semantics
- *In situ* untangling
- Piecemeal dissection of terms and generation of OWL descriptions

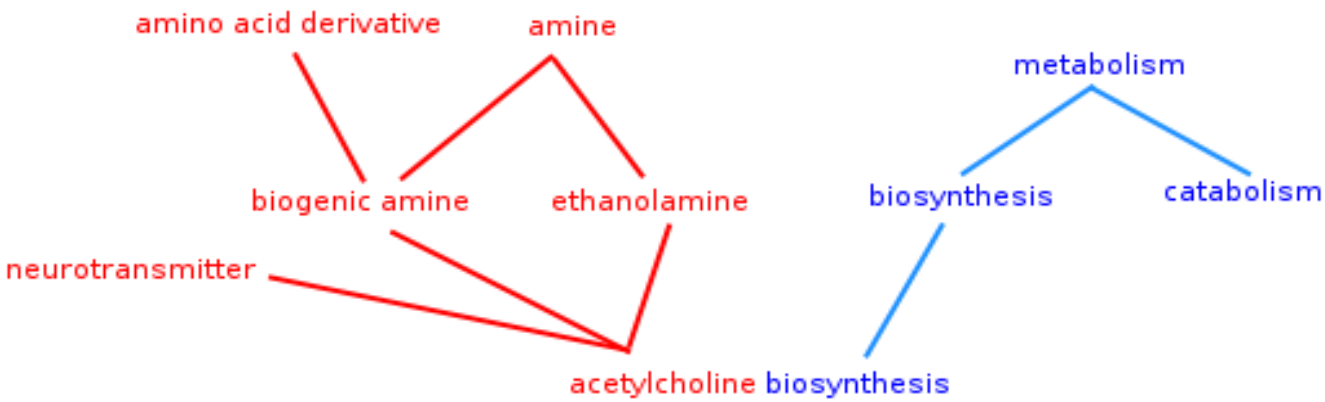
ep 3: GO moving to step 4

- Gene Ontology “acetylcholine biosynthesis”:



A Dissection

- Gene Ontology “acetylcholine biosynthesis” dissected:



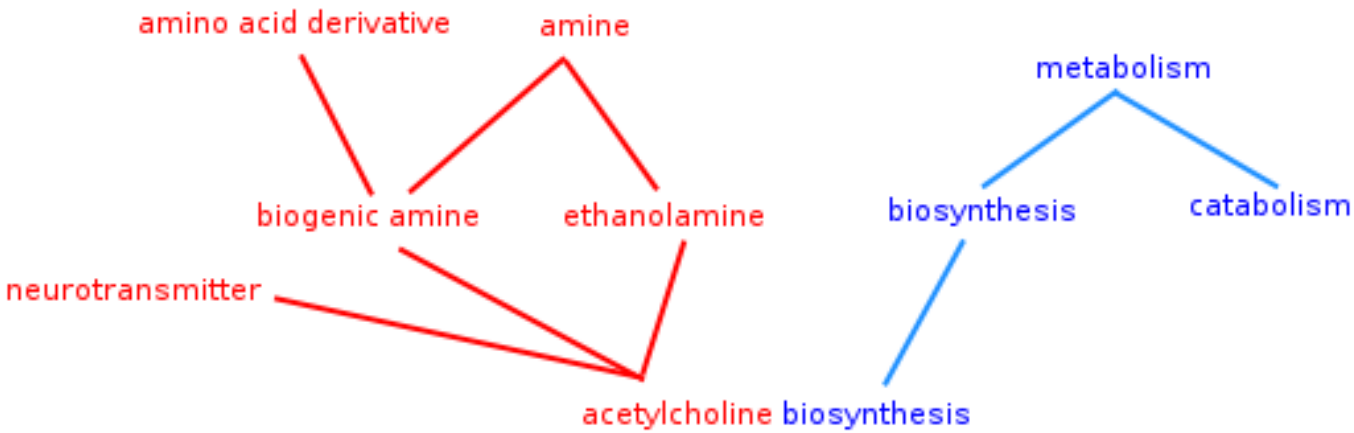
Generating OWL

- Gene Ontology “acetylcholine biosynthesis” dissected in OWL DL (Abstract syntax):

Class (acetylcholine biosynthesis complete

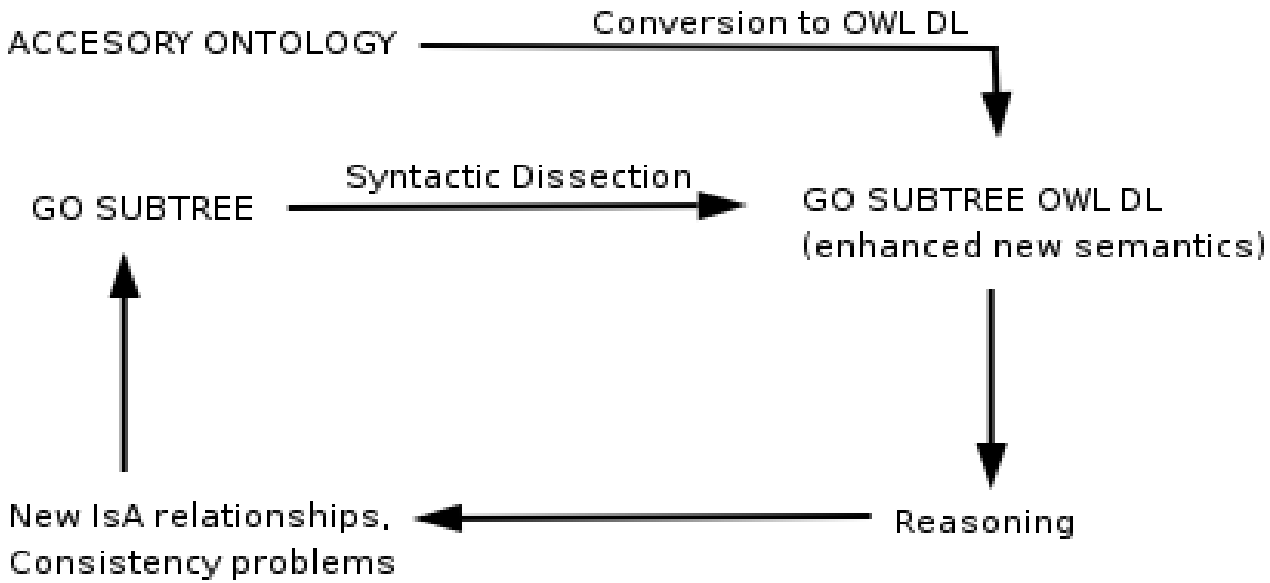
restriction (actsOn someValuesFrom (acetylcholine)))

SubClassOf (acetylcholine biosynthesis biosynthesis)



GONG Workflow

- The GONG workflow:

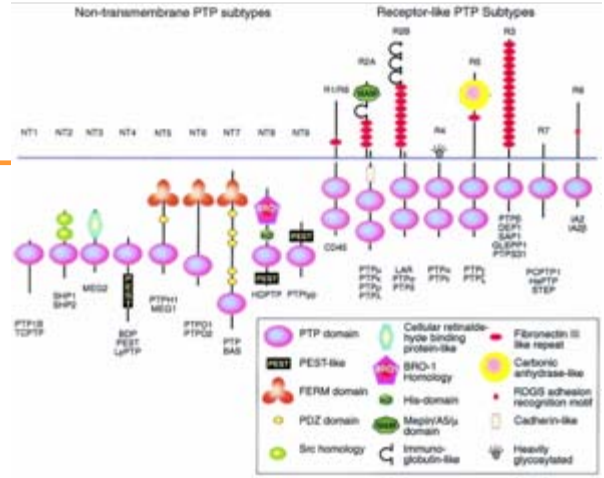


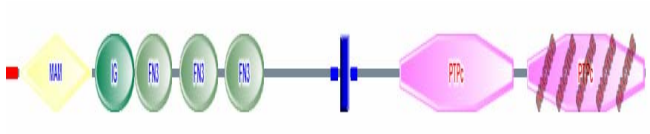
GONG results

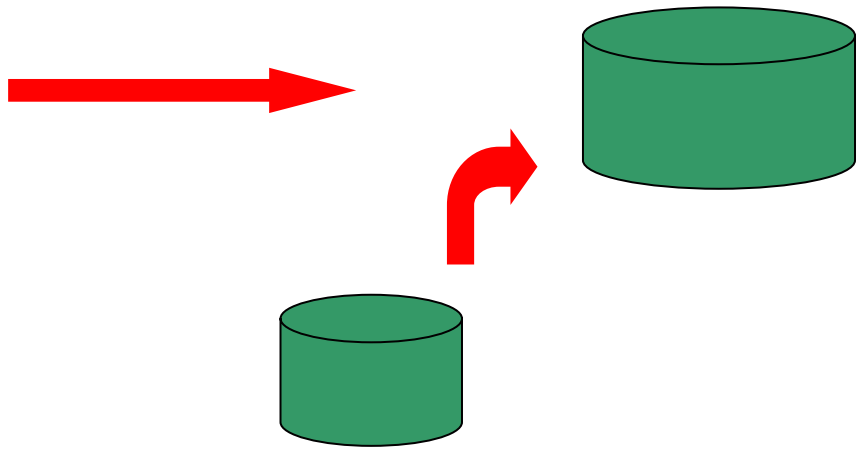
Go Area	Changed Terms	Accepted change
Binding	17%	5%
Transport activity	21%	8%

Classifying Proteins

```
>uniprot|Q15262|PTPK_HUMAN Receptor-type protein-tyrosine
      phosphatase kappa precursor (EC 3.1.3.48) (R-PTP-
      kappa).
MDTTAAAALPAFVALLLLSPWPLLGSAQGQFSAGGCTFDDGPGACDYHQDLYDDFEWVHV
SAQEPHYLPPEMPQGSYMIVDSSDHDPGEKARLQLPTMKENDTHCIDFSYLLYSQKGLNP
GTLNILVRVNKGPLANPIWNVTGFTGRDWLRAELAVSSFWPNEYQVIFEAEVSGGRSGYI
AIDDIQVLSYPCDKSPHFLRLGDVEVNAGQNATFQCIATGRDAVHNKLWLQRRNGEDIPV
      ……..
```







MANCHESTER 1824

The University of Manchester

Summary

- In the spectrum of “ontology” each point has its uses
- One can migrate towards the formal end
- Has benefits in machine reasoning and maintenance
- Has costs in effort
- OWL loses user view of DAG
- But can transform back again
- OWL can simple tree to beastly complexity
- Perhaps the best of both worlds

GONG extended results

Go area	Captured terms	Changed terms	Accepted Changes by Mikel	Accepted Changed
Binding	98%	17%	8%	5%
Transporter activity	94%	21%	11%	8%