A Methodology to Migrate the Gene Ontology to a Description Logic Environment using DAML+OIL

Chris Wroe, Robert Stevens, Carole Goble
University of Manchester, UK
Michael Ashburner
EBI, Hinxton, UK

http://gong.man.ac.uk

THE UNIVERSITY of MANCHESTER

Gene Ontology Next Generation Project (GONG)

- Demonstrate the utility of finer grained concept descriptions in DAML+OIL
- Develop methodologies and tools to support the process

http://gong.man.ac.uk

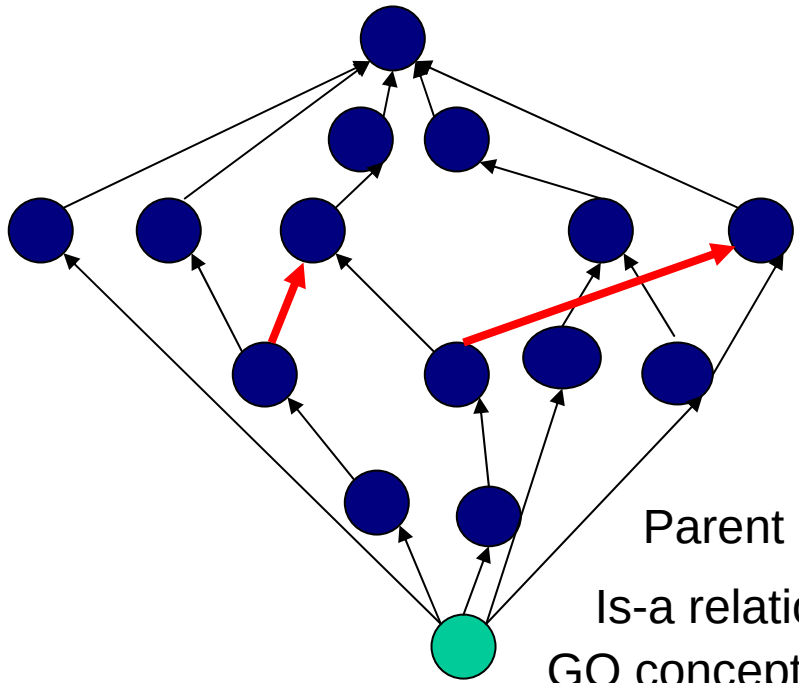
THE UNIVERSITY of MANCHESTER

Translating theory into practice

- Gene Ontology provides a service to the model organism database community
- Description logic (DL) is a technology born out of computer science research
- DAML+OIL becoming a standard ontology interchange language underpinned by DL
- Adopted by W3C and will soon become Ontology Web Language (OWL)

GONG - proof of concept

- Maintaining an exhaustive is-a structure



http://gong.man.ac.uk

THE UNIVERSITY of MANCHESTER

Example: heparin biosynthesis

Axis 1:

Chemicals

[*chemical*] biosynthesis (GO:0009058)

[i] *carbohydrate* biosynthesis (GO:0016051)

[i] *aminoglycan* biosynthesis (GO:0006023)

[i] *heparin* biosynthesis (GO:0030210)

Example: heparin biosynthesis

Axis 1:

Chemicals

[*chemical*] biosynthesis (GO:0009058)

[i] *carbohydrate* biosynthesis (GO:0016051)

[i] *aminoglycan* biosynthesis (GO:0006023)

.....

[i] *heparin* biosynthesis (GO:0030210)

Axis 2:

Process

[i] heparin *metabolism* (GO:0030202)

[i] heparin *biosynthesis* (GO:0030210)

http://gong.man.ac.uk

THE UNIVERSITY of MANCHESTER

Example: heparin biosynthesis

Axis 1:

Chemicals

[*chemical*] biosynthesis (GO:0009058)

[i] *carbohydrate* biosynthesis (GO:0016051)

[i] *aminoglycan* biosynthesis (GO:0006023)

[i] glycosaminoglycan biosynthesis (GO:0006024)

[i] *heparin* biosynthesis (GO:0030210)

Axis 2:

Process

[i] heparin *metabolism* (GO:0030202)

[i] heparin *biosynthesis* (GO:0030210)

Is this important?

- Missing is-a not noticed by users
- BUT… improves fidelity of DB record retrieval.
  - Asking for gene products involved in 'glycosaminoglycan biosynthesis' will lead to an additional result:

O94923 SPTr ISS - D-glucuronyl C5-epimerase (Fragment)

How can DAML+OIL support the task?

- Step 0. Translate to DAML+OIL syntax
  - Provided by OilEd (ontology editing tool)

http://oiled.man.ac.uk

GO RDF	DAML+OIL
<go:term>	<daml:Class>
<go:isa>	<daml:subClassOf><daml:Class>
<go:part-of>	<daml:subClassOf><daml:Restriction> <daml:onProperty><daml:ObjectProperty rdf:resource="go:part-of"/> <daml:hasClass><daml:Class>

DAML+OIL definitions for metabolism concepts

- heparin biosynthesis
  - **class** heparin biosynthesis **defined**
    subClassOf biosynthesis
    restriction onProperty acts_on **hasClass** heparin
    (acts_on is unique)
  - Paraphrase: biosynthesis which acts solely on heparin
- glycosaminoglycan biosynthesis
  - class glycosaminoglycan biosynthesis defined
    subClassOf biosynthesis
    restriction onProperty acts_on hasClass **glycosaminoglycan**

A chemical ontology

- Mapped chemical concepts to MeSH (using UMLS tools/ API e.g. Norm).
- Created a DAML+OIL ontology from a subset of the MeSH chemical taxonomy
- Provides the following information:

carbohydrates
[i] polysaccharides
[i] glycosaminogylcans
[i] heparin

Paraphrased reasoning process

- heparin biosynthesis
  - **class** heparin biosynthesis **defined**
    subClassOf biosynthesis
    restriction onProperty acts_on **hasClass** **heparin**

- glycosaminoglycan biosynthesis
  - class glycosaminoglycan biosynthesis defined
    subClassOf biosynthesis
    restriction onProperty acts_on hasClass **glycosaminoglycan**

Is-a

Inferring a new is-a link

- heparin biosynthesis
  - **class** heparin biosynthesis **defined**
    subClassOf biosynthesis
    restriction onProperty acts_on **hasClass heparin**

Is-a

- glycosaminoglycan biosynthesis
  - class glycosaminoglycan biosynthesis defined
    subClassOf biosynthesis
    restriction onProperty acts_on hasClass **glycosaminoglycan**

Is-a

Output

- OilEd API reports additional inferred is-a relationships.
  E.g.
  heparin biosynthesis has new is-a parent
  glycosaminoglycan biosynthesis
- Report sent to GO editorial team for comment.
- They makes changes to GO if appropriate and sends back queries

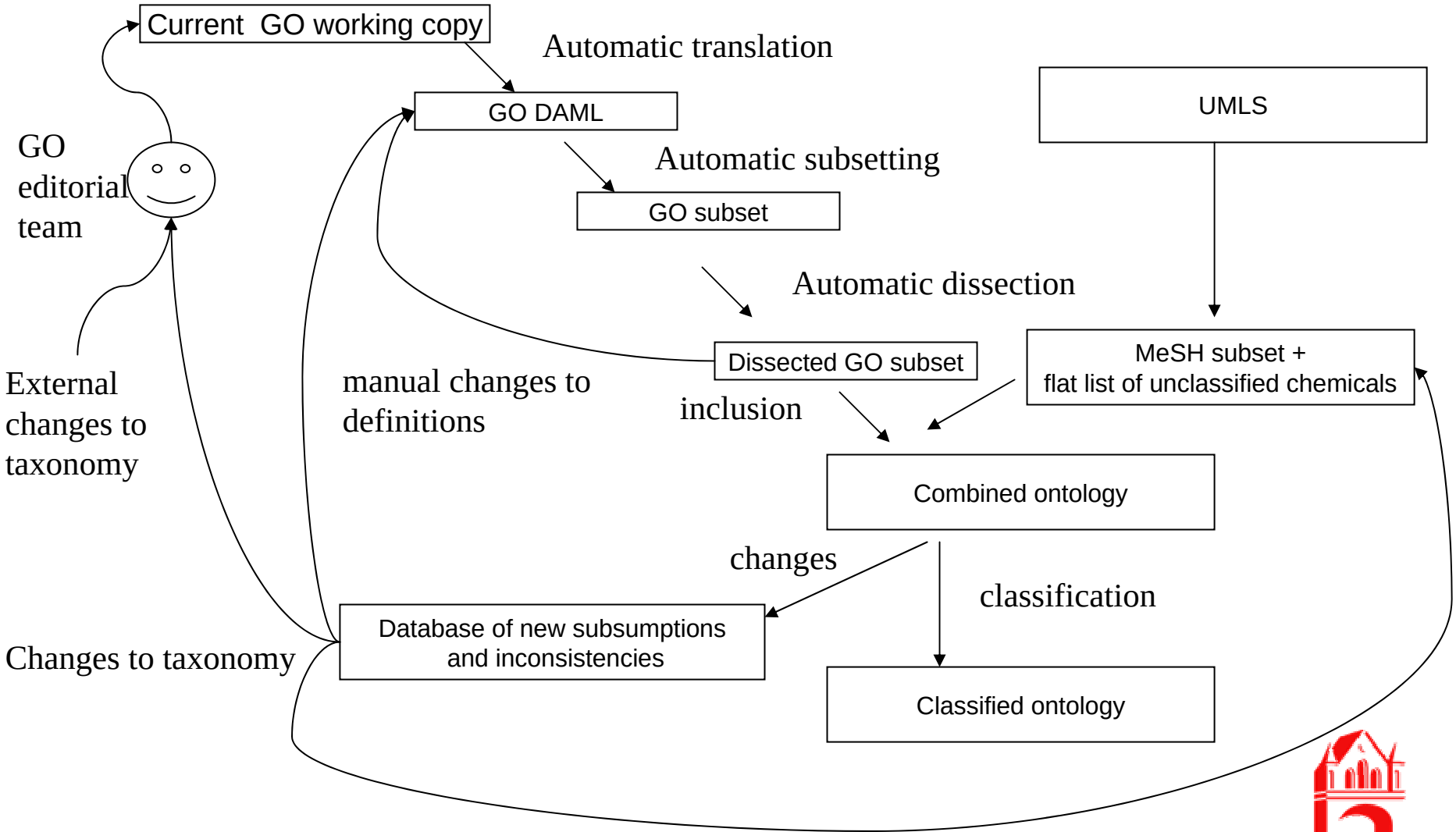
Results

- Carbohydrate metabolism ~250 concepts
  - 22 additional is-a links 17 of which now in GO
- Amino acid metabolism ~ 250 concepts
  - Further 17 additional is-a links now in GO
- GO team will be reviewing results for metabolism as a whole once we have the tools to support the process
- Useful results come from even a partial coverage

Build a practical environment

- Tools needed for:
  - Creating DAML+OIL definitions
  - Tracking changes
  - Reporting reasoning results
  - Viewing definitions

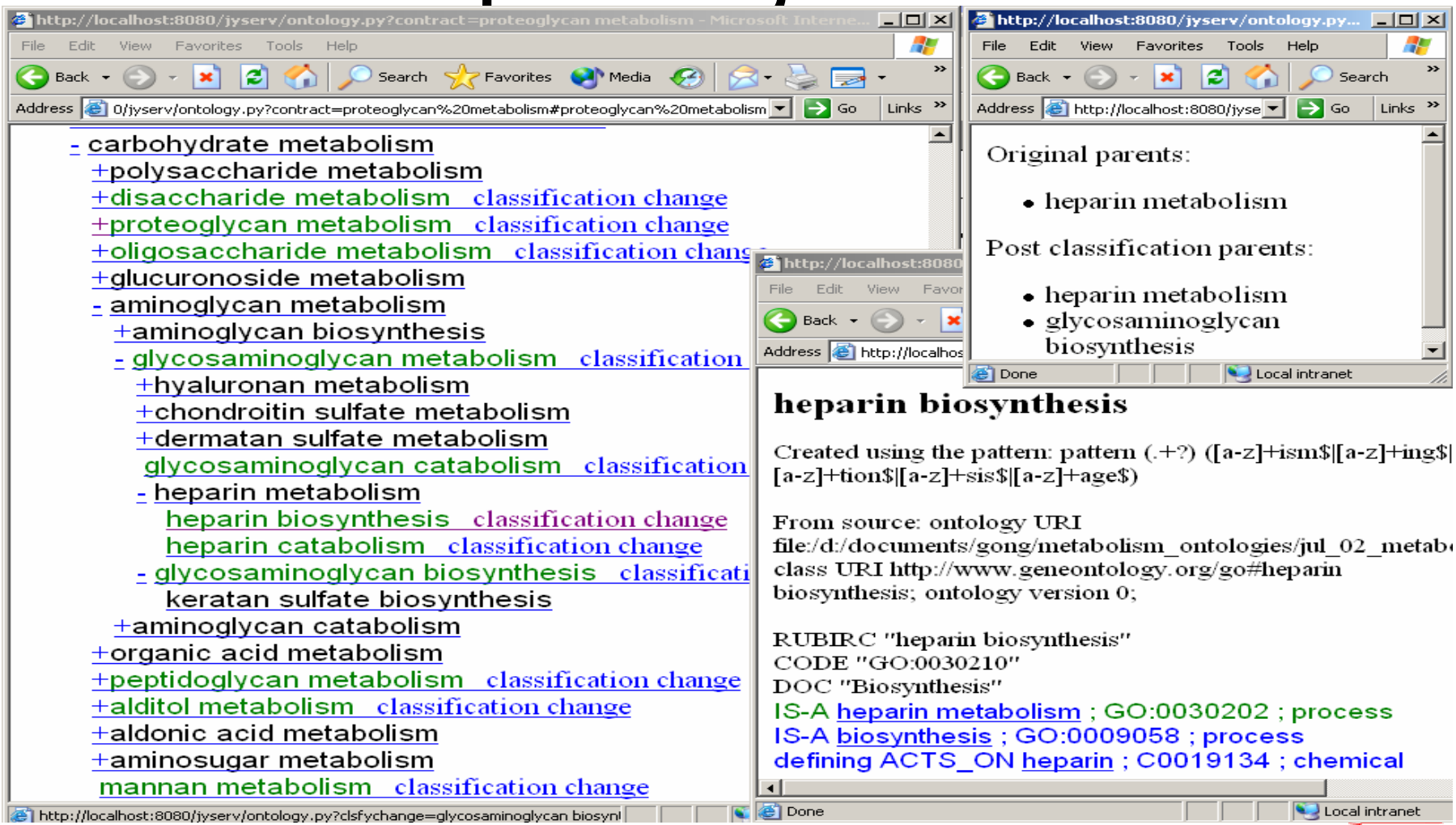
Multi-step workflow



http://gong.man.ac.uk

THE UNIVERSITY of MANCHESTER

Reporting tools



http://gong.man.ac.uk

THE UNIVERSITY of MANCHESTER

Using Amigo to view definitions

Conclusion & future plans

- Description logic approach is useful
- Beginning to build a practical environment
- Extend quantity and detail of DAML+OIL definitions
  - ~2000 complete but un-validated metabolism definitions
  - ~2000 partial enzyme definitions
- Continue prototyping and developing tools to support the process.

Acknowledgments

- Jane Lomax and Midori Harris of the GO editorial team for help and advice and responding to the suggested changes
- UMLS and MeSH which provided valuable resources for chemical information
- Sean Bechhofer for development on OilEd

- Project funded as a subcontract of the DARPA DAML programme