Toward Distributed Development of Ontologies

Mark A. Musen
Stanford University
Manual, top–down development of ontologies is the norm

- Yahoo! taxonomy of Web pages
- Most clinical terminologies
- All the worlds’ upper-level ontologies
  - SUMO
  - Open CYC
  - HL7 RIM
The Open Directory Project

- Started by Netscape in 1999
- "Editors" volunteer to develop categorizations of their domains of expertise
- There are explicit expectations of editors and review of their work
- At present, thousands of editors, yet only 0.5 FTE of paid management!
Gene Ontology Consortium

• Outstanding example of community-driven ontology development
• Small group of volunteer developers work tirelessly to expand their work
• A new version of GO is released every 30 minutes!
• But there are lots of tensions in the GO community
Prologue

— Carole Goble

Two households, both alike in dignity,
In fair genomics, where we lay our scene,
(One, comforted by its logic’s rigour,
Claims ontology for the realm of pure,
The other, with blessed scientist’s vigour,
Acts hastily on models that endure),
From ancient grudge break to new mutiny,
When “being” drives a fly-man to blaspheme.
From forth the fatal loins of these two foes
Researchers to unlock the book of life;
Whole misadventured piteous overthrows
Can with their work bury their clans’ strife.
The fruitful passage of their GO-mark’d love,
And the continuance of their studies sage,
Which, united, yield ontologies undreamed-of,
Is now the hours' traffic of our stage;
The which if you with patient ears attend,
What here shall miss, our toil shall strive to mend.

Based on an idea by Shakespeare
A Portion of the OBO Library
Moving from cottage industry to the industrial age

- There must be widely available **tools** that are open-source, that are easy to use, and that adhere to knowledge representation standards.
- There must be a large user **user community** of developers who use the tools and who can provide feedback to one another and to the core team of tool builders.
- There must be an **organizational structure** to provide a modicum of coordination and quality control.
- If all these pieces are in place, people can do great work from their cottages!
The NCI Thesaurus
NCI Thesaurus Production Environment with History

Lead Editor TDE
- Work Manager Client
- Editing Application
- Conflict Detection/Resolution
- DB Schema
  - Master NCI Baseline
  - Master History

Work List Generation

Change Set

External Testing

NCI Thesaurus Test DTS Servers

Production Release

Release

Hx

Validation

Hx

Baseline

Classification

Conflict Detection and Resolution

NCI Thesaurus Editing Environment

Individual Editors' TDE
- Workflow Client
- Editing Application
- DB Schema
  - Current NCI Baseline
  - Local History

Schema

NCI Thesaurus Production DTS Servers

Schema

Schema

Deployed (Real)

Integrative, Experimental, Collaborative

Sustainable, End-to-End Production Process

NCI Thesaurus Today:
Three Critical Pieces
Egyptian Crop Pests
A thousand flowers are blooming!

- Ontologies are being developed by interested groups from every sector of academia, industry, and government—often “on the cheap”
- Many of these ontologies have been proven to be extraordinarily useful to wide communities
- Many of these same ontologies have been shown to have structural flaws
- We finally are at the stage where we have open-source tools and standard representation languages that can help us to create durable and maintainable ontologies with rich semantic content
Research Questions

• How do we measure the “quality” of ontologies developed via industrial engineering and by grass-roots efforts?
• Can we perform controlled studies to compare the cost-effectiveness of top–down versus bottom–up approaches?
• Can we evaluate the cost-effectiveness of alternative management structures to facilitate distributed ontology development?
• Can we develop correlates of peer review to ensure open quality control of ontologies created in a distributed fashion?