Community-based ontology development, alignment, and evaluation

Natasha Noy Stanford Center for Biomedical Informatics Research Stanford University

Community-based Ontology ... Everything



Evaluation and Reuse

Mapping and Alignment

How Large Can the "Community" be?



Can be just a few distributed users





Collaborative Ontology Development

• Collaborative

- Several users contribute to a single developing ontology
- There are mechanisms to carry out discussions and to reach consensus

Ontologies

- From simple taxonomies
- To expressive OWL ontologies

Use Cases in Biomedical Domain

- Gene Ontology (GO)
- NCI Thesaurus
- BiomedGT
- OBI, BIRNLex

The Gene Ontology

Terminology for consistent description of gene products



- 3 full-time curators have access to edit GO
- Anyone in the community can submit an issue or request

The NCI Thesaurus

Reference ontology for cancer biology, translational science, and clinical oncology



- ~20 full-time editors making changes
- Changes are not immediately visible
- I "lead editor" who approves the changes, and assigns new tasks

BiomedGT

An "open" version of the NCI Thesaurus



- Several different roles
- Capabilities depend on the user's role
- Anyone in the community can contribute suggestions; small number of curators

Other Biomedical Projects

• OBI, BIRNLex, RadLex

- tightly-knit community of developers (20-30)
- most actively participate in discussions
- I or 2 editors make changes to the ontology

Dimensions of Collaborative Workflows

- Ontology size
 - from 100s to 10,000s of concepts
- Size of the community
 - Contributors (in some form): from 2-3 to dozens
 - Editors: from 1-2 to 20
- Control mechanisms
 - Variety of roles
 - Gatekeepers, etc.
 - Client-server editing

- Discussion tools
 - mailing lists, message boards
 - face-to-face meetings, telecons
- Synchronization and editing mechanisms
 - CVS, SVN

Tool Requirements

Tools for discussion and reaching consensus

- annotate components and, maybe, changes
- have as an integral part of the development process
- Context for discussions on modeling decisions
- Record of changes and associated discussions and controversies
- Provenance and trust
 - support concept histories
 - have ways establish trust and credibility
- Personalized views of an ontology
 - based on user's role and tasks
 - based on user's level of expertise
 - based on user's trust network

Tool Requirements (cont'd)

- Support for personal and shared spaces
- Access control
 - fine-grained control for editing and viewing rights
- User roles
- Flexible workflow support
 - configurable workflows
 - workflow-execution coupled with ontology development

Bringing Collaboration to Protégé

Protégé

- is an open-source ontology editor
 - developed at Stanford, in collaboration with University of Manchester
 - with more than 100,000 registered users
- has dozens of plugins for
 - visualization
 - inference
 - import and export
 - natural-language processing
 - ... (things we don't know exist)

Collaborative Protégé

 An extension of the Protégé to support collaborative development



Features of Collaborative Protégé

- Support for:
 - annotating ontology components and changes in the ontology
 - discussion threads
 - proposals and voting
 - searching and filtering
 - defining users, groups, policies
- Distributed with Protégé installation

http://protege.stanford.edu/doc/collab-protege/

Annotations

- Annotations are linked to a specific ontology component:
 - class
 - property
 - individual
 - ontology as a whole
 - ontology change
 - ... others

- There are different types of annotations:
 - question
 - comment
 - proposal
- Users may filter and search annotations based on different criteria

Collaborative Protégé Interface

≪	Thesaurus F	Protégé 2 Collaborativa	ojects/nci/0712e/Thesaurus.pprj, OWL / RDF Files)
<u>F</u> ile <u>E</u> dit <u>P</u> roject <u>O</u> WL <u>R</u> easoning <u>C</u> ode	<u>T</u> ools <u>W</u> indow <u>!</u>		Collaborative
008 406 22 44	? D P <	Panel	Tabs Protégé
🔶 Metadata (Thesaurus.owl) 👘 😑 OWLCIasses	Properties C	hanges	Collaboration
SUBCLASS EXPLORER	CLASS EDITOR	+ - C	🖓 🚺 🔞 Annotations 🖗 Changes 🦪 Discussion Threads 🕅 Search 🐵 Chat
For Project: 🖶 Thesaurus	For Class: 🛑 Gen	e_Product (instance of owl:Class) [Inferred]	View
Asserted Hierarchy 🥸 😫 🧙	📑 📀 🍫 🔽	م 🖵 📃	Go Go Comment
🛑 owl:Thing 🔼	Property	Value	Discussion Threads
A rdfs:Class	rdfs:comment		This term is clearly mis-assigned to the cell adhesion molecules. SwissProt indicates that this is a nuclear protein in
Abnormal_Cell_Kind	code	C26548	Integrinal term does this go with?
Activity_Kind	DEFINITION	<def-source>NCI</def-source> <def-definition>A</def-definition>	
Biological Process Kind		determined by genomic sequences.	V Gene Products Discussion Thread
Chemicals_and_Drugs_Kind	🗖 DesignNote	In NCI Thesaurus: includes Proteins, Functional RNAs,	🔻 🍓 GO Biological ProcessI've been examining the depths of the GO heirarchy for a way to represent the proteins in
Chemotherapy_Regimen_Kind		Protein Complexes, and Riboprotein Complexes	— UPDATEOur next meeting is scheduled for Thursday March 13 at 10:30 am. (After that we should be moving
Diagnostic_and_Prognostic_Factors_Kind		Product <term-group>PT<td>p I'll try to look at the GO slim before our next meeting - s</td></term-group>	p I'll try to look at the GO slim before our next meeting - s
EO_Anatomy_Kind		> <term-source>NCI</term-source>	
EO_Findings_and_Disorders_Kind	FULL_SYN	<pre><term-name>Genome Encoded EntitySY</term-name></pre>	
Equipment_Kind		<term-source>NCI</term-source>	
	NCI_META_CUI	CL055397	Details Annotations
🕨 🛑 Gene	🗖 Preferred_Na	Gene Product	
🔻 🦲 Gene_Product_Kind	🗖 rdfs:label	Gene Product	GO Biological Process
🔻 😑 Gene_Product 🤕 🔪	🗖 Semantic_Type	Classification	I've been examining the depths of the GO heirarchy for a way to represent the proteins in the Gene Product branch by
Functional_RNA	🗖 Synonym	Gene Product	even that needs a little trimming. There are some surprises that I didn't anticipate as well (i.e. Proteolysis is not in the
hnRNA	🗖 Synonym	Genome Encoded Entity	Catabolic Process branch). I have put together a heirarchy based on this analysis and we can go over it at our next
			to look at it.)
Protein Fragme ha			
Protein_or_R			
🔻 🥚 Molecular_Abnormality annot	ations		
🔻 😑 Molecular_Abnormality 📃 👻			
	•		Created A Part of Madified Part of of
Gene_Product 💌 🛤 💼 🗣 🔏 🚴	👲 🔅 🕒	😑 💿 Logic View 🔿 Properties Vi	iew
			Annotation

details

Chat



Architecture



Changes and Annotations



But wait, there is more...



Support For Flexible Workflows

- Goal:
 - automatically generate tools that support customtailored workflows
 - cover a wide range of existing workflows

Custom-Tailored Workflows



Workflow Ontology



Community-based Ontology ... Everything



Evaluation and Reuse

Mapping and Alignment

NCBO BioPortal

- The National Center for Biomedical Ontology (<u>http://bioontology.org</u>) is developing BioPortal, an opensource repository of ontologies, terminologies, and thesauri of importance in biomedicine.
- An early version of BioPortal is accessible at http://bioportal.bioontology.org. An alpha version of the next release is at http://alpha.bioontology.org/
- Users can access the BioPortal content interactively via Web browsers or programmatically via Web services.

The BioPortal Ontology Repository

- Open repository of ontologies in biomedicine
- Each ontology is described by a set of metadata
- Ontologies in different formats
 - OWL, RDF(S)
 - OBO
 - Protégé frames
- BioPortal technology
 - open-source
 - domain-independent

Home Browse	Search Proje	ects Mapp	oings				
Ontology Name	Format	Version	Author	Date	Related Projects	Reviews	
African Traditional Medicine	OBOF	1.0.1			0 Related Projects	<u>See All 0</u> <u>Reviews</u>	Explore
Amino Acid	OWL-FULL	1.2			0 Related Projects	See All 0 Reviews	Explore
Amphibian gross anatomy	OBOF	1.7			0 Related Projects	See All 0 Reviews	Explore
Animal natural history and life history	PROTEGE	See Remote Site			0 Related Projects	See All 0 Reviews	Exploring Unavailable
Basic Vertebrate Anatomy	OWL-FULL	1.1			0 Related Projects	See All 0 Reviews	Explore
Biological imaging methods	OBOF	1.1			0 Related Projects	See All 0 Reviews	Explore
Biological process	OBOF	1.204			0 Related Projects	See All 0 Reviews	Explore
Biomedical Resource Ontology	OWL-LITE	1.1			0 Related Projects	See All 0 Reviews	Explore
BIRNLex	OWL-DL	1.3.1			0 Related Projects	See All 0 Reviews	Explore
BRENDA tissue / enzyme source	OBOF	1.295			0 Related Projects	See All 0 Reviews	Explore
<u>C. elegans</u> development	OBOF	1.3			0 Related Projects	<u>See All 0</u> <u>Reviews</u>	Explore
C. elegans gross anatomy	OBOF	See Remote Site			1 Related Projects	See All 1 Reviews	Exploring Unavailable
<u>C. elegans</u> phenotype	OBOF	See Remote Site			0 Related Projects	See All 0 Reviews	Exploring Unavailable

Ontology Navigation in BioPortal



Major Function: Ontology Assessment



Which ontology is appropriate for my task?

• Sources of information for the answer:

- Ontology metadata
 - usually provided by authors
- Computable metrics
 - can be provided by the tools
- Community-based evaluation
 - provided by other users of the ontologies

Ontology Metadata

- Provides answers to pertinent questions about the content and provenance of the ontology in the library:
 - What is the domain covered by an ontology?
 - What are the key classes and concepts?
 - Who developed the ontology?
 - What is the policy for maintenance and distribution?
 - What is the format of the ontology (syntax, language, tools used to build it, etc.)

Computable ontology metrics

Logical consistency checking

- e.g., use a reasoner to determine if an ontology is consistent
- Structural consistency based on meta-properties
 - e.g., OntoClean
- Rules of thumbs and heuristics
- Statistical information
 - number of classes and properties
 - connectedness, fan-out, etc.
 - cycles

Not all useful metrics are computable

- Many aspects of ontology quality are subjective
- The most useful information for the user selecting an ontology:
 - who used an ontology for a similar task and how well did it work?

Some Ontology Metrics are Subjective

- What is a "good" feature in some setting, can be a "bad" feature in another setting
- High level of axiomatization:
 - good if you want to perform reasoning
 - can be bad because of the high computational and cognitive cost if you don't need the axioms
- Organizing anatomy concepts based primarily on their structure rather than function
 - can be good if you need to understand which organs a wound goes through
 - not appropriate if you need to understand spread of disease

Community-Based Evaluation



- The only people who know the answer to these question are
 - (maybe) ontology authors
 - other users of the ontology
- Allow users to provide ratings for ontologies

Open-Rating Systems

Open publishing system

- anyone can publish content
 - (Semantic) Web is inherently an open publishing system
 - many ontology libraries allow contributions by anyone
- Closed Rating system
 - only a group of "editors" can provide ratings
 - Open directory
 - Yahoo! directory
- Open Rating system
 - anyone can publish reviews and ratings
 - Amazon reviews



Other Editions: Kindle Edition

Excerpt - page 6: "... 1 The Semantic Web The Semantic Web is the application of advanced knowledge technologies ..."

Surprise me! See a random page in this book.



Applying Open Ratings to Ontologies

Broject Name: Badlley	Institution: RSNA	
	People:	
Description: HadLex is a controlled terminology for radiology.	Home Page: http://radlex.org	
Ontologies Used		
RadLex		
★★☆☆☆ Degree Of Formality ★★☆☆☆ Documentation And Support ★★★☆☆ Usability ★★★☆☆ Domain Coverage ★★★☆☆ Quality Of Content		

Home Browse Search Projects	Mappings NCI Thesaurus 🗵 FMA	Sign In Register
View Ontology Metadata	Physical anatomical entity Visualization Details Marginal Notes Mappings	Resources Annotations
Non-physical anatomical entity Physical anatomical entity Immaterial physical anatomical entity Material anatomical entity	Comment: Wrong superclass in Class/Details The superclass for this class in the Class detail is	ngriff at 01/31/08 17:09 wrong. <u>Reply</u>
Dimensional entity	New Thread	ngriff at 01/31/08 17:10

Reviewers Provide

- General review and rating
- Usage information
 - Which applications have successfully used the ontology?
 - What problems were encountered?
- Coverage
 - Does it cover the domain properly?
 - Are there major gaps?
 - Are some parts developed better than others?
- Concept-specific comments
 - Are there problems with specific concepts?
 - What alternative definitions should be used?

Conflicting Sources of Metadata

- Authors and users can contradict one another
 - Quality of documentation?

•

...

- References (e.g., positive and negative analyses of the ontology)
- Metadata schema must enable diversity of views on some metadata values

Representing Reviews and Ratings

Looks familiar?



Community-based Ontology ... Everything



Evaluation and Reuse

Mapping and Alignment

Multiple and Overlapping Ontologies

- Ontologies, vocabularies, and terminologies will inevitably overlap in coverage
- An ontology library can provide declarative mappings
 - found by the tools (efficient, but far from perfect)
 - specified by users (low throughput, but better quality)





Mappings in BioPortal

Home	Browse	Search	Projects	Mappings	NCI Thesa	iurus 🗵				-
Perineal Artery		Sural Artery								
Popliteal Artery		Visualization Details Marginal Notes Mappings Resources Annotations								
Azygos Articular Artery		New Point-to-Point Mapping								
	Inferior Internal Articular Artery		Mapping T	0			Source	Mapped By	Mapped On	
Superior External Articular Arten			external sural artery (Mouse adult gross anatomy)			NCICB	TerryHayamizu	04/23/08		
	Sural Artery S Posterior Tibial Artery		sural artery (Mouse adult gross anatomy)			NLM	SongmaoZhang	04/23/08		
Po			external sural artery (Mouse adult gross anatomy)			NLM	SongmaoZhang	04/23/08		
Posterior Tibial Artery Branch 🖗 📗		superficial sural artery (Mouse adult gross anatomy)			NLM	SongmaoZhang	04/23/08			
Pul	Imonary Arte	ry 🖘 ry Branch								
🛨 Radial Artery 🔗										
Retinal Artery Spiral Artery of the Endometrium										
Splenic Trabecular Artery										

Creating Point-to-Point Mappings

Home	Home Browse Search Pro		Projects	Mappings	NCI Thesaurus 🗵
			Poir	nt-to-Point	Mapping from Sural Artery to :
Ontology: Search: Searc	Galen artery h Can	icel	•		
RenalA RenalA RenalF RenalF RenalP RenalS RightG RightG RightG RightG SightG Superio Superio Superio	nteriorSegm nteriorSuper rtery oferiorSegm osteriorSegm uperiorSegn olicArtery astricArtery astroepiploid epaticArtery feriorPhrenid ulmonaryArt Artery orMesenterid orRectalArte uodenalArte tery	nentalArtery riorSegment RenalArtery mentalArtery nentalArter nentalArtery cArtery ery cArtery ery cArtery ry		NAMEDAR NAMEDAR	aduodenalArtery.SupraduodenalArtery rtery asSubclass nalArtery

Create

Analyzing Mappings

- Filter					
User: Any					
Source: Any					
Filter Export Current Results To RDF					
	« Previous 1 2 3 4 5 6 7 8 9 15 16 Next »				
Concept	Maps To				
	Mouse adult gross anatomy : pelvis bone (1)				
	Mapped By TerryHayamizu				
NCI Thesaurus - Pelvic Bone ->	Mouse adult gross anatomy : hip bone (1)				
	Mapped By SongmaoZhang				
	Mouse adult gross anatomy : pelvic girdle bone (1)				
	Mapped By SongmaoZhang				
	Mouse adult gross anatomy - external sural artery (2)				
	Mapped By TerryHayamizu , SongmaoZhang				
	Mouse adult gross anatomy : sural artery (1)				
NCI Thesaurus : Sural Artery>	Mapped By SongmaoZhang				
	Mouse adult gross anatomy : superficial sural artery (1)				
	Mapped By SongmaoZhang				
	Mouse adult gross anatomy : spermatic vein (2)				
	Mause edult grees ensterny a testionic weis (4)				
NCI Thesaurus : Spermatic Vein>	Mapped By SongmaoZhang				
	Mouse adult gross anatomy : internal spermatic vein (1)				
	Mapped By SongmaoZhang				

Functionality for Mapping Support

- Enable users to
 - upload bulk mappings in a specified format
 - download mappings based on a selection criteria
 - define point-to-point mappings interactively
 - comment on mappings created by others
 - refine and discuss existing mappings

Mapping Metadata

- yes, metadata again...
- Users require a comprehensive set of metadata to augment mappings
 - mapping relationship
 - provenance (who created the mapping and when)
 - discussion and comments
 - application context
 - mapping dependency
 - algorithm used to create the mapping (configuration, parameters, etc.)
 - external references

Representing mappings



Current Mappings in BP

More than 30,000 mappings

- created manually
 - as part of concept definition (OBO xref, UMLS)
 - as a mapping (NCI Mouse-Human anatomy)
- created automatically
 - Using algorithms such as Prompt

Mappings as the Product of Community Contribution

- Mappings can contradict each other
 - application context may be different
 - trust is the key again
- Users can use the BioPortal framework to reach consensus on the mappings

Use of Mapping Repository

- Source of data for automatic algorithm
 - machine learning
 - algorithms that need a priori alignment
- Accessible through web services
 - can be used in other applications
- Used for annotating and browsing resources through ontology elements
- Use for finding "important" ontologies:
 - If everyone maps to NCI Thesaurus, it must be important

Community-based Ontology ... Everything

Pevelopment and Evolution Are we there yet?

> Mapping and Alignment

Research Challenges

- If we build it, will they come?
- How do we encourage users?
 - it is a paradigm shift, in some sense
 - so far, our users are asking for these features
 - collaborative ontology development is hot!
 - community-based mappings and evaluation is still new

One of the key issues: Trust

- Aggregation
- Meta-rankings: rating the raters
- Personalized and filtered views of the system: Web of Trust
- Online version of "word of mouth"
- Topic-specific trust



"On the Internet, nobody knows you're a dog."

Collaborative Workflows

- How do we develop a representation that is comprehensive enough to capture a wide variety of collaborative workflows?
- How do we make it simple and usable?
- What are the reusable workflow modules? Can we have a simple wizard that instantiates a workflow description?

Dynamics of Collaborative Ontology Development

- Properties and dynamics of the social networks that form as the result of collaborative ontology editing?
- What are the different types of users based on their activity in the editing process (e.g., those who lead the discussions vs those who do most of the editing)?
- What are the characteristics of the concepts that are discussed most actively?
- Do tools such as Collaborative Protégé make collaboration more efficient or hinder it?

Personalized Views of Ontologies

- How do we represent and compute personalized views
 - based on the areas of interest
 - based on the user's web of trust?
- How do you enable local editing of ontologies with subsequent integration in the ontology being developed by others?

Maintaining Metadata through Ontology Versioning



Potentially, any part of the description can change: author, language, domain, ... Mappings can become invalid, too... Reviews may no longer apply...

Acknowledgements

- Mark Musen
- Tania Tudorache
- Abraham Sebastian
- Nick Griffith
- Sherri de Coronado
- Tim Redmond

- Jennifer Vendetti
- Nigam Shah
- Benjamin Dai
- Michael Dorf
- Harith Alani
- R.V. Guha
- Holger Lewen