Collaborative patterns in large-scale ontology development projects

Stanford Center for Biomedical Informatics Research

Sean Falconer
Collaborative patterns in large-scale ontology development projects

Stanford Center for Biomedical Informatics Research

Sean Falconer
Collaborative patterns in large-scale ontology development projects

Stanford Center for Biomedical Informatics Research

Sean Falconer
Background
Ontology Development

The Recent Past

Mostly developed by individuals
Ontology Development
Current Situation

• Ontologies are too big, too complex
• Tens of thousands of classes
  – Gene ontology, NCI Thesaurus, SNOMED
Collaborative Development

• Many people *working together*

• Variety of workflows
  – Formal review
    • E.g., NCI Thesaurus
  – Open suggestion, closed editing
    • E.g., Gene ontology
  – Crowdsourcing
    • E.g., ICD-11
Tool Support

- NeON Toolkit
- Collaborative Protégé
- Semantic wikis
- Knoodl
CKC Challenge

- Goal: assess the **state of the art** for tools that support **collaborative ontology development**

Still very much in their **infancy**

No single tool likely to fulfill requirements

Requirements **poorly understood**
My Research Focus
Need to understand existing projects

• What are the similarities and differences in terms of the roles participants fulfill?
• What characteristics distinguish the role of a contributor?
• Is there a relationship between collaborative ontology development and other collaborative projects? (E.g., OSS and Wikipedia)
• Can a deeper understanding of these projects and roles inform tool and ontology development?
Outline

• The Data: Three Large-Scale Biomedical Ontologies
• Collaborative Protégé
• Change Analysis Tool
• Discovering Collaborative Patterns
  – User behavior
  – User roles
  – Collaboration and changes
• Discussion
The Data: Three Large-Scale Biomedical Ontologies

• Studied three diverse projects:
  – National Cancer Institute’s Thesaurus (NCI Thesaurus)
  – WHO International Classification of Disease, revision 11
  – Biomedical Resources Ontology (BRO)

• Vary in size, development stage, and collaborative workflow
The NCI Thesaurus

- Reference vocabulary for clinical care, translational and basic research, and cancer biology
- Over 80,000 concept classes
- Been in development for years
- Workflow:
  - Edited by a multi-disciplinary team
  - Lead editor accepts/rejects changes
International Classification of Disease (ICD-11)

• Standard diagnostic classification used to encode information relevant to epidemiology, health management, and clinical use
• More than 20,000 concept classes
• Revision 11 only recent
• Workflow:
  – Web-based process, potentially thousands of contributors
  – Crowdsourcing model
Biomedical Resources Ontology (BRO)

• Enabling technology for describing biomedical resource types, areas of research, and resource activity

• Workflow:
  – Small group of distributed editors
  – Use web-based interface to edit and collaborate
  – Discussions used to reach consensus
Outline

• The Data: Three Large-Scale Biomedical Ontologies

• **Collaborative Protégé**

• Change Analysis Tool

• Discovering Collaborative Patterns
  – User behavior
  – User roles
  – Collaboration and changes

• Discussion
Collaborative Protégé

• Supports *client-server* model
• Changes are *recorded* and *tracked*
• Users can hold *discussions*, *chat* and provide *notes*
Collaborative Protégé
Collaborative Protégé

- Data is stored in the Change and Annotation Ontology (ChAO)
- Classes represent changes and annotations
- Changes and notes are stored as instances
- Example: Class_Created

<table>
<thead>
<tr>
<th>Action</th>
<th>Description</th>
<th>Entity</th>
<th>Created</th>
</tr>
</thead>
<tbody>
<tr>
<td>Composite Change</td>
<td>Create class Warmth -- Apply to: Warmth</td>
<td>... Ontology entity on which the change...</td>
<td>01/07/2010 14:34:49 EST</td>
</tr>
<tr>
<td>Class Created</td>
<td>Class Create: Warmth</td>
<td>... Warmth</td>
<td>01/07/2010 14:34:45 EST</td>
</tr>
<tr>
<td>Individual Added</td>
<td>Instance Added: Warmth (instance of: owlClass)</td>
<td>... Warmth</td>
<td>01/07/2010 14:34:45 EST</td>
</tr>
<tr>
<td>Subclass Added</td>
<td>Subclass Added: Warmth (added to: owl:Thing)</td>
<td>... Warmth</td>
<td>01/07/2010 14:34:45 EST</td>
</tr>
<tr>
<td>Superclass Added</td>
<td>Superclass Added: 'Property or Attribute' (added to: Warmth)</td>
<td>... Properties or Attributes</td>
<td>01/07/2010 14:34:47 EST</td>
</tr>
<tr>
<td>Subclass Added</td>
<td>Subclass Added: Warmth (added to: 'Property or Attribute')</td>
<td>... Warmth</td>
<td>01/07/2010 14:34:47 EST</td>
</tr>
<tr>
<td>Composite Change</td>
<td>Remove superclass owl:Thing from Warmth -- Apply to: <a href="http://ncicb.nci.nih">http://ncicb.nci.nih</a></td>
<td>... Warmth</td>
<td>01/07/2010 14:34:47 EST</td>
</tr>
<tr>
<td>Annotation Added</td>
<td>Annotation added: code: 'C85500' to class: Warmth</td>
<td>... Warmth</td>
<td>01/07/2010 14:34:48 EST</td>
</tr>
<tr>
<td>Annotation Added</td>
<td>Annotation added: rdfs:label: 'Warmth' to class: Warmth</td>
<td>... Warmth</td>
<td>01/07/2010 14:34:48 EST</td>
</tr>
<tr>
<td>Annotation Added</td>
<td>Annotation added: Preferred_Name: 'Warmth' to class: Warmth</td>
<td>... Warmth</td>
<td>01/07/2010 14:34:48 EST</td>
</tr>
<tr>
<td>Annotation Added</td>
<td>Annotation added: FULL_SYN: '&lt;term-name&gt;Warmth&lt;/term-name&gt;'</td>
<td>... Warmth</td>
<td>01/07/2010 14:34:48 EST</td>
</tr>
</tbody>
</table>
Motivation reminder ...

• Need to **understand** existing projects

• Establish user **roles**, contributor **characteristics**

• Inform **tool** and **ontology** design
Putting it together

Analyzing changes and annotations
Changes and Notes

• NCI Thesaurus:
  – October 5\textsuperscript{th}, 2009 to April 14\textsuperscript{th}, 2010
  – Six baseline releases
  – 43,702 changes by 10 authors

• ICD-11:
  – November 2009 to May 2010
  – 14,554 changes, 4,768 notes by 19 authors

• BRO:
  – February to March 2010
  – 762 changes, 373 notes by 5 authors
Lots of data

• How do we analyze the data?

• Collaborative Protégé not useful for exploring and asking analytical questions
Outline

• The Data: Three Large-Scale Biomedical Ontologies
• Collaborative Protégé
• **Change Analysis Tool**
• Discovering Collaborative Patterns
  – User behavior
  – User roles
  – Collaboration and changes
• Discussion
Change Analysis Tool

Developed **Change Analysis Plugin** for Protégé
Main screen

Supports exploration of changes by term
Graph view

Changes by authors on a per month basis
Dependency view

CONCEPTS

- V 'Mental and behavioural disorders' (419)
- VI 'Diseases of the nervous system' (2) (1459)
- VI 'Diseases of the eye and adnexa' (3) (2257)
- VII 'Diseases of the ear and mastoid process' (8)
- X 'Diseases of the respiratory system' (44)
- XI 'Diseases of the digestive system' (9) (1537)
- XII 'Diseases of the skin' (26) (7298)
- XIII 'Diseases of the musculoskeletal system and connective tissue' (419)
- XIV 'Diseases of the genitourinary system' (419)
- XV 'Injury, poisoning and certain other consequences of external causes' (4)
- XVI 'Certain conditions originating in the perinatal period' (4)
- XVII 'Congenital malformations, deformations and accidents in utero' (4)
- XVIII 'Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified' (201)

CHANGES (9)

<table>
<thead>
<tr>
<th>Action</th>
<th>Description</th>
<th>Author</th>
<th>Created</th>
</tr>
</thead>
<tbody>
<tr>
<td>Composite Change</td>
<td>Moved class <a href="http://who.int/">http://who.int/</a>... Linda Best</td>
<td>Linda Best</td>
<td>04/28/2010 07:12:20 PDT</td>
</tr>
<tr>
<td>Composite Change</td>
<td>Moved class <a href="http://who.int/">http://who.int/</a>... Linda Best</td>
<td>Linda Best</td>
<td>04/29/2010 07:13:02 PDT</td>
</tr>
<tr>
<td>Composite Change</td>
<td>Moved class <a href="http://who.int/">http://who.int/</a>... Linda Best</td>
<td>Linda Best</td>
<td>04/29/2010 07:13:06 PDT</td>
</tr>
<tr>
<td>Composite Change</td>
<td>Moved class <a href="http://who.int/">http://who.int/</a>... Francesco Gongolo</td>
<td>Francesco Gongolo</td>
<td>04/29/2010 07:34:48 PDT</td>
</tr>
<tr>
<td>Composite Change</td>
<td>Moved class <a href="http://who.int/">http://who.int/</a>... Francesco Gongolo</td>
<td>Francesco Gongolo</td>
<td>04/29/2010 07:35:01 PDT</td>
</tr>
<tr>
<td>Composite Change</td>
<td>Moved class <a href="http://who.int/">http://who.int/</a>... Francesco Gongolo</td>
<td>Francesco Gongolo</td>
<td>04/29/2010 07:35:07 PDT</td>
</tr>
<tr>
<td>Composite Change</td>
<td>Moved class <a href="http://who.int/">http://who.int/</a>... Linda Best</td>
<td>Linda Best</td>
<td>04/30/2010 03:51:54 PDT</td>
</tr>
</tbody>
</table>
Implicit dependencies

Author B makes a change

Author A makes a change
Implicit dependencies

Ontology structure changes make-up of social network
Outline

• The Data: Three Large-Scale Biomedical Ontologies
• Collaborative Protégé
• Change Analysis Tool
• Discovering Collaborative Patterns
  – User behavior
  – User roles
  – Collaboration and changes
• Discussion
Discovering Collaborative Patterns
Chaos Into Order

• Taking inventory:
  – Lots of data (NCI Thesaurus, ICD-11 and BRO)
  – Changes and notes in a format we understand
  – Change Analysis plugin for exploring
Change Analysis Plugin

• By using the various views, we could start to explore and answer questions

• We explored the data from the three projects
• We were able to make important observations about user behavior
Where do users make their changes?

Author filters allow us to explore this

Author A – restricted to one hierarchy

- 'ICD Categories' (90/26347)
  - I 'Certain infectious and parasitic diseases' (0/2700)
  - II Neoplasms (0/1337)
  - III 'Diseases of the blood and blood-forming organs'
  - IV 'Endocrine, nutritional and metabolic diseases'
  - IX 'Diseases of the circulatory system'
  - V 'Mental and behavioural disorders'
  - VI 'Diseases of the nervous system'
  - VII 'Diseases of the eye and adnexa' (90/2257)
  - VIII 'Diseases of the ear and mastoid process'
  - X 'Diseases of the respiratory system'
  - XI 'Diseases of the digestive system'
  - XII 'Diseases of the skin'

Author B – changes throughout

- 'ICD Categories' (1/27) (2002/26347)
  - I 'Certain infectious and parasitic diseases' (1/15) (681,
  - II Neoplasms (65/1337)
  - III 'Diseases of the blood and blood-forming organs'
  - IV 'Endocrine, nutritional and metabolic diseases'
  - IX 'Diseases of the circulatory system'
  - V 'Mental and behavioural disorders'
  - VI 'Diseases of the nervous system'
  - VII 'Diseases of the eye and adnexa'
  - VIII 'Diseases of the ear and mastoid process'
  - X 'Diseases of the respiratory system'
  - XI 'Diseases of the digestive system'
  - XII 'Diseases of the skin' (1/26) (249/7298)
Do authors make overlapping changes?

• Yes, but some do this more often
Do authors make certain types of changes?

• It appeared that some authors made certain types of changes
  – Lots of deletions versus additions and edits

• Some authors made many changes while others made very few
  – E.g., One BRO author made 17 changes while another made 368
Identifying user roles

• Based on observations, were there similar user roles across the three projects?

• Represented authors as feature vectors
• Features were numerical representations of our observed characteristics

Feature examples:
• $C_{\text{del}}$ – ratio of deletion changes to all changes
• $C_{\text{add}}$ – ratio of term addition changes to all changes
• $D$ – average hierarchy depth of a change
• $M$ – ratio of multi-author changes
Clustering authors

- Used repeated K-means to cluster authors based on similarity between their feature vectors

- After measuring cluster compactness and quality, best result was with five clusters
Analyzing distinguishing characteristics

• What makes each cluster unique?
• Why is one author in one particular cluster and not another one?

Measured the **statistical significance** of each feature across all five clusters
Example

<table>
<thead>
<tr>
<th>Feature</th>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
<th>Cluster 4</th>
<th>Cluster 5</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deletion</td>
<td>0.068</td>
<td><strong>0.436</strong></td>
<td>0.0858</td>
<td>0.040</td>
<td>0.077</td>
<td>0.023</td>
</tr>
</tbody>
</table>

The table shows the means for different features across various clusters, with the highest mean for Deletion in Cluster 2.
<table>
<thead>
<tr>
<th>Cluster</th>
<th>Role</th>
<th>Characteristics</th>
<th>Primary activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 1</td>
<td>Ontology expert</td>
<td>Highly central, makes changes over multiple hierarchies, involved in fewer leaf changes than domain experts, but performs a lot of movement changes in the hierarchy.</td>
<td>Organizational</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>Content manager</td>
<td>Edits mostly in one sub-hierarchy, low centrality, performs few movement changes, but a high number of deletions</td>
<td>Hierarchy clean-up</td>
</tr>
<tr>
<td>Cluster 3</td>
<td>Domain expert</td>
<td>Edits mostly deep within one hierarchy, low centrality, few moves, but lots of concept additions</td>
<td>Content creation</td>
</tr>
<tr>
<td>Cluster 4</td>
<td>Central domain expert</td>
<td>Edits are restricted primarily to one sub-hierarchy, very central, edits high up in hierarchy.</td>
<td>Management and content creation of a specific area of the ontology</td>
</tr>
<tr>
<td>Cluster 5</td>
<td>Content editor</td>
<td>Highly central, makes changes over multiple hierarchies, lots of leaf changes, higher number of property changes.</td>
<td>Editing of existing content</td>
</tr>
</tbody>
</table>
## Summary of roles

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Role</th>
<th>Characteristics</th>
<th>Primary activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 1</td>
<td>Ontology expert</td>
<td>Highly central, makes changes over multiple hierarchies, involved in fewer leaf changes than domain experts, but performs a lot of movement changes in the hierarchy.</td>
<td>Organizational</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>Content manager</td>
<td>Edits mostly in one sub-hierarchy, low centrality, performs few movement changes, but a high number of deletions</td>
<td>Hierarchy clean-up</td>
</tr>
<tr>
<td>Cluster 3</td>
<td>Domain expert</td>
<td>Edits mostly deep within one hierarchy, low centrality, few moves, but lots of concept additions</td>
<td>Content creation</td>
</tr>
<tr>
<td>Cluster 4</td>
<td>Central domain expert</td>
<td>Edits are restricted primarily to one sub-hierarchy, very central, edits high up in hierarchy.</td>
<td>Management and content creation of a specific area of the ontology</td>
</tr>
<tr>
<td>Cluster 5</td>
<td>Content editor</td>
<td>Highly central, makes changes over multiple hierarchies, lots of leaf changes, higher number of property changes.</td>
<td>Editing of existing content</td>
</tr>
<tr>
<td>Cluster</td>
<td>Role</td>
<td>Characteristics</td>
<td>Primary activity</td>
</tr>
<tr>
<td>---------</td>
<td>--------------------------</td>
<td>---------------------------------------------------------------------------------</td>
<td>------------------------------------------------------</td>
</tr>
<tr>
<td>Cluster 1</td>
<td>Ontology expert</td>
<td>Highly central, makes changes over multiple hierarchies, involved in fewer leaf changes than domain experts, but performs a lot of movement changes in the hierarchy.</td>
<td>Organizational</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>Content manager</td>
<td>Edits mostly in one sub-hierarchy, low centrality, performs few movement changes, but a high number of deletions</td>
<td>Hierarchy clean-up</td>
</tr>
<tr>
<td>Cluster 3</td>
<td>Domain expert</td>
<td>Edits mostly deep within one hierarchy, low centrality, few moves, but lots of concept additions</td>
<td>Content creation</td>
</tr>
<tr>
<td>Cluster 4</td>
<td>Central domain expert</td>
<td>Edits are restricted primarily to one sub-hierarchy, very central, edits high up in hierarchy.</td>
<td>Management and content creation of a specific area of the ontology</td>
</tr>
<tr>
<td>Cluster 5</td>
<td>Content editor</td>
<td>Highly central, makes changes over multiple hierarchies, lots of leaf changes, higher number of property changes.</td>
<td>Editing of existing content</td>
</tr>
</tbody>
</table>
What can the notes tell us?
Analyzing notes

Using notes from ICD-11 and BRO, we attempted to answer:

Q1: Is there a relationship between changes and author discussions with respect to a specific ontology term?

Q2: Do people who make a lot of changes also participate in a lot of discussion?
Results

Q1: Compared change and note activity across all terms
   – Positive correlation between change and note activity in both projects

Q2: Compared specific author changes to specific author notes
   – Positive correlation in ICD-11 but not in BRO
Outline

• The Data: Three Large-Scale Biomedical Ontologies
• Collaborative Protégé
• Change Analysis Tool
• Discovering Collaborative Patterns
  – User behavior
  – User roles
  – Collaboration and changes
• Discussion
Discussion

• Roles indicated:
  – domain experts work primarily within one hierarchy
  – ontology experts and content editors work over multiple hierarchies
  – important implication for tool design

• Change data for domain experts points to topic areas of ontology
  – potentially useful for ontology modularization
Relationship to OSS

- Similar roles discovered in OSS (Xu et. al)
  - Project leaders, Core developer, Co-developers and Active users

- OSS has strong relationship between mailing list contributions and development activity (Bird et. al)
  - similar to our analysis with notes and changes
Conclusions

• Collaborative ontology development is still very new
• Little is known about how existing projects function
• Our results indicate that there are clearly discernible roles
• Future: hope to get more data to continue to generalize results
Acknowledgements

Thanks to:

– Tania Tudorache
– Natasha Noy
– Mark Musen
Thank you