(Towards) The Last Annotation Tool


U. Colorado School of Medicine
LIMSI-CNRS
US National Library of Medicine
U. Paris-Sorbonne
U. Potsdam

http://compbio.ucdenver.edu/Hunter_lab/Cohen
Neves and Leser (2012): Multiple good annotation tools are available

• …so, why has the community spent $500,000 building new ones in the past 5 years?
• ...and can we do it just one more time, and be done for a while?
• Hypothesis: None of the existing tools has a sufficiently broad set of capabilities

• So: Figure out what that set is, build it, and we’re done (for a while)
### Textual annotation tools: Why build, why reuse?

<table>
<thead>
<tr>
<th>Incentives to build</th>
<th>Incentives to reuse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Don’t have to learn</td>
<td>Community of maintainers</td>
</tr>
<tr>
<td>Don’t have to modify task</td>
<td>Be productive sooner</td>
</tr>
<tr>
<td>Overhead</td>
<td>Already beta-tested</td>
</tr>
<tr>
<td>Fun</td>
<td></td>
</tr>
<tr>
<td>Tangible deliverable</td>
<td></td>
</tr>
<tr>
<td>LPU</td>
<td></td>
</tr>
<tr>
<td>Your format definitely supported</td>
<td></td>
</tr>
</tbody>
</table>
What’s different about annotation in biomedicine?

• Frequent need to accommodate large, deep ontologies
• Longer texts than are common in other domains
• Frequent multi-model annotation tasks
• Frequent combination with light annotation tasks
Approach (abstract)

A. Find out what features people have been building
B. Find out what else they might want
C. Figure out how to prioritize affordances
Approach (concrete)

1. Draft paper
2. Literature review
3. Design a survey
4. Delphi: BLAHmuc participants; authors of corpus construction papers; …
5. Write requirements specification
One way to carry out the Delphi procedure

- Which of these functionalities would be more important to you?
  A. Annotate portions of a word
  B. Discontinuous annotations

- Which of these functionalities would be more important to you?
  A. Discontinuous annotations
  B. Unicode support

- Which of these functionalities would be more important to you?
  A. Link to Wikipedia
  B. Unicode support
<table>
<thead>
<tr>
<th>Claim</th>
<th>How to test it</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequent need to accommodate <strong>large, deep ontologies</strong></td>
<td>Sample 20 each biomedical and other projects—determine size of ontologies</td>
</tr>
<tr>
<td><strong>Longer texts</strong> than are common in other domains</td>
<td>Sample 20 each biomedical and other projects; necessary to specify scientific/clinical?</td>
</tr>
<tr>
<td>Frequent <strong>multi-model annotation tasks</strong></td>
<td>Need to differentiate between multiple « ontologies » and multiple « models »</td>
</tr>
<tr>
<td>Frequent combination with <strong>light annotation tasks</strong></td>
<td>Need to « normalize » across papers</td>
</tr>
</tbody>
</table>
Should we be designing a tool, or an architecture?

• Tool plusses:
  – Tangible deliverable
  – Far shorter learning curve for potential users
  – May make problems more obvious

• Tool minusses:
  – Easy to « design in » limitations from the beginning

• Architecture plusses:
  – Far more extensible
  – Easier for a community to pick up development

• Architecture minusses:
  – Adopted less often than tools
  – Probably need to understand what the tool would need to provide before we can design an architecture
Schedule

• Monday: Symposium
• Tuesday:
  – Assemble list of extant tools to examine
    • Does it need to be sub-classed by task?
  – Design feature matrix
    • If it gets boring, can this be outsourced?
  – Populate same
    • If it gets boring, can this be outsourced?
• Wednesday: How to find out what else people want?
• Next week: design Delphi procedure.
• Next month: Delphi. Survey Monkey? Email? …
Acknowledgments

- BioHackathon 2016 participants
- BLAHmuc organizers

Me, Rick Osborne, Prabha Yadav, Natalya Panteleyeva, Negacy Hailu, Irina Grichtchenko, Ivo Georgiev (not pictured: Lisa Ensign)