System requirements for BioCreative IV IAT task

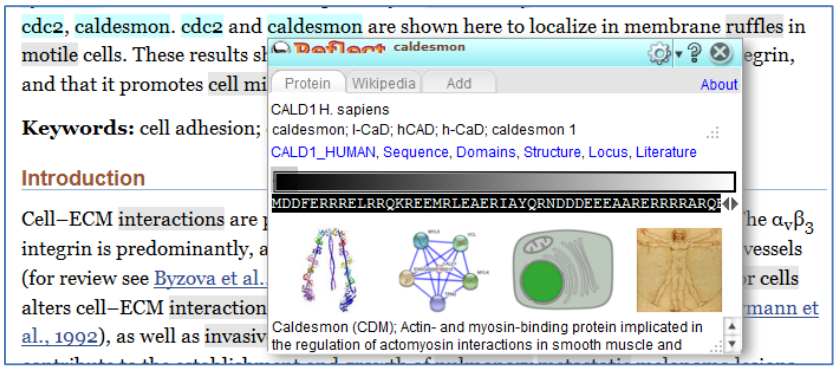
Prepared by: Cecilia Arighi, Sherry Matis and Phoebe Roberts; with input from the User Advisory Group (http://www.biocreative.org/#committee)

We are asking for web-based text mining systems that allow the user to modify text mining results. Based on User Advisory Group (UAG) discussions, we have come up with a list of desired functionalities that already exist in working tools (see below). Note that the examples are supplied to clarify the intent of the functionality and are not meant to be comprehensive.

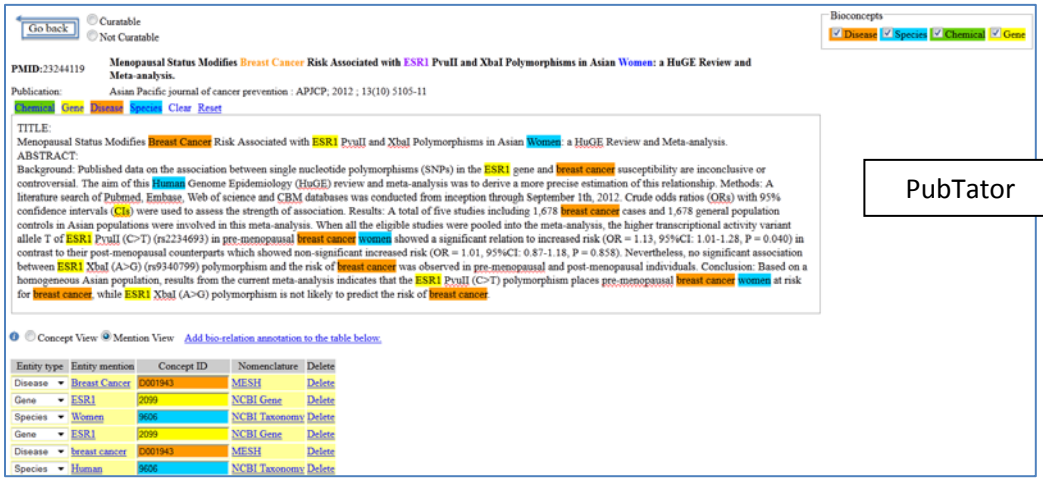
Mandatory functionalities:

a. System should be compatible with the most commonly used web browsers: Firefox, Chrome, Safari and Explorer.

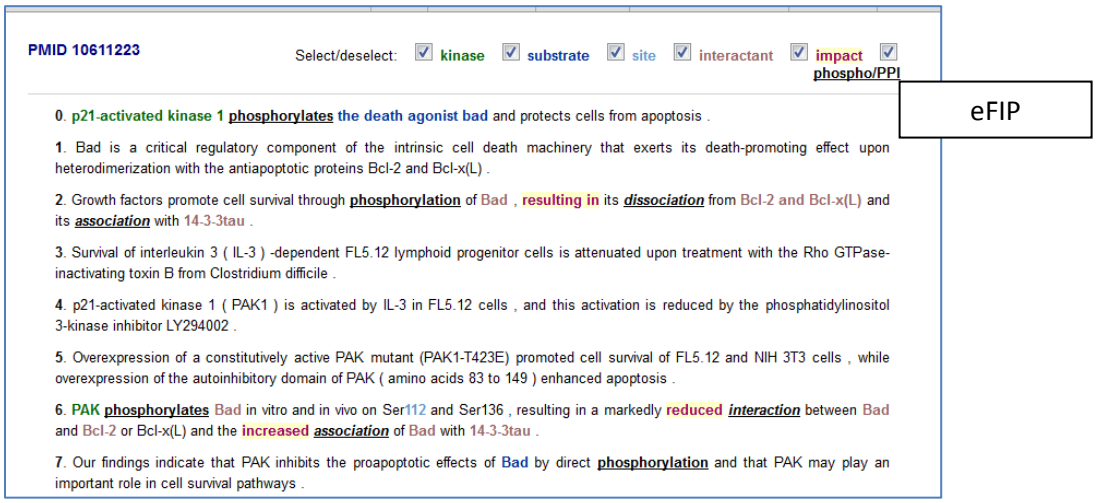
b. System should highlight entities and relationships (if applicable) relevant to the annotation task. We encourage color coding of entity types and links to relevant data sources when appropriate. E.g. Reflect (http://reflect.embl.de/) and PubTator (http://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/PubTator/) for entities. eFIP (http://proteininformationresource.org/pirwww/iprolink/eFIP.shtml) as example of entities and relations color coded.



Reflect

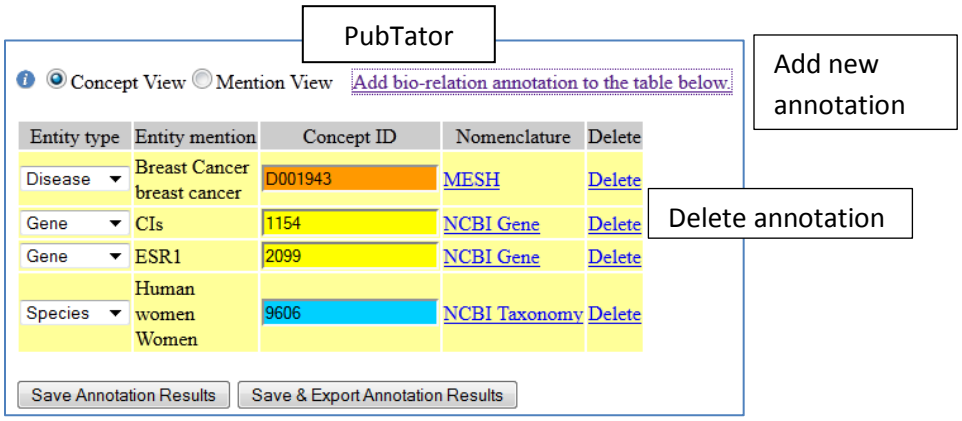


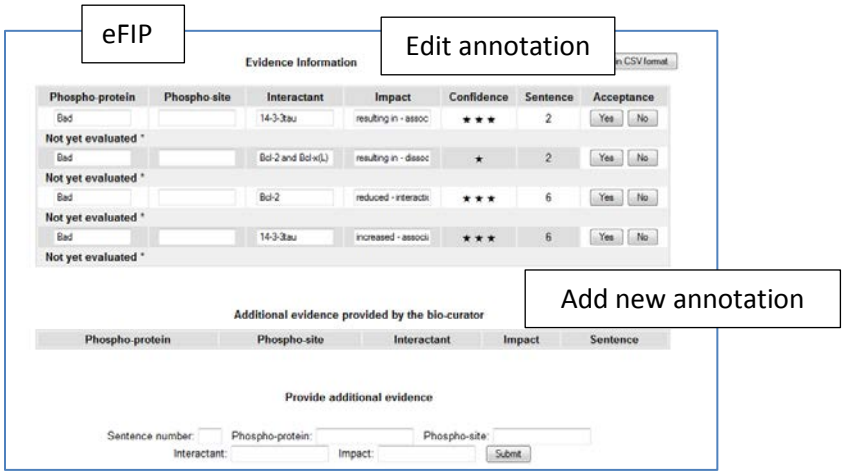
PubTator



c. User should be able to edit the text mining results by correcting errors or adding missing information, and should be able to export the corrected data.

- Ability to review curation decisions and revise if necessary. Systems such as PubTator (see snapshot below and previous example) and eFIP (once logged in) offer this functionality.





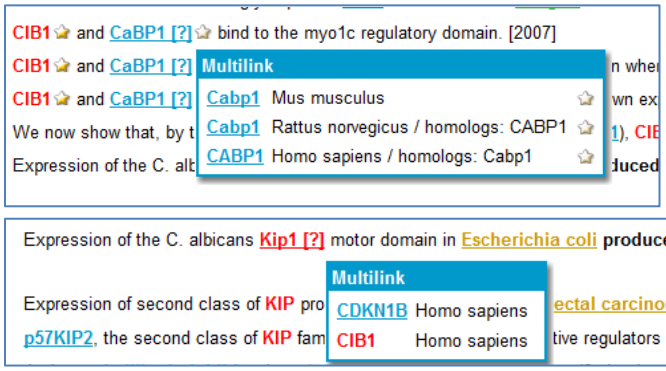
d. Use standard input and output formats; if possible support more than one format type: *Input*: For biomedical literature: document ID (e.g. PubMed IDs, or PubMed central IDs) and/or document text (e.g. text, html, pdf or XML). For other types of input use adopted standards when possible, which should be consulted with user community.

   Output: Export the results at least in tab-delimited and XML formats. We also encourage teams to adopt the BioC format (XML-based) described in the interoperability Track I (http://www.biocreative.org/tasks/biocreative-iv/track-1-interoperability/)

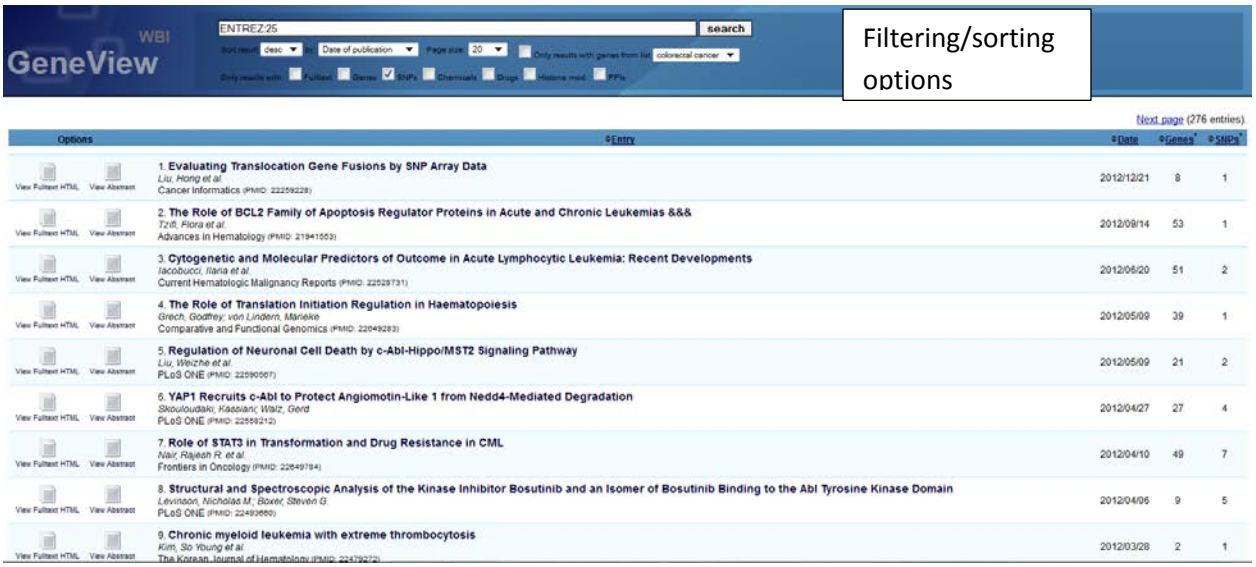
Other strongly desired functionalities:

a. Full-text processing. Minimally to be able to process PubMed Central open access articles. This was one of the top selected features requested by UAG. However, we know that some curation activities are abstract-based, and whether to enable full text processing will be left to your discretion.

b. Interactive disambiguation of domain entities. E.g. in iHOP (see example http://www.ihop-net.org/UniPub/iHOP/gs/95608.html?IN=1) a question mark is added next to the entity for disambiguation of both names and species to select correct one.
   Example:

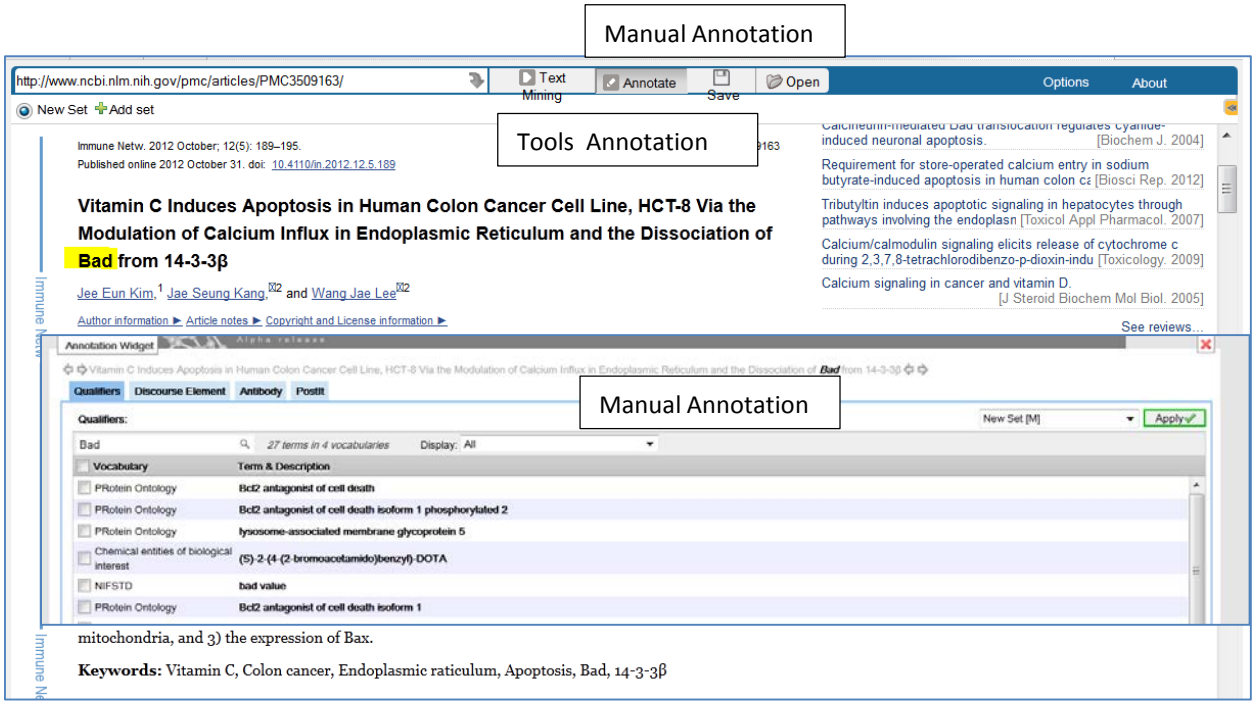


c. Ability to filter/sort the results according to different criteria; rank results based on what is more relevant to the user. E.g., when searching for a gene in GeneView (http://bc3.informatik.hu-berlin.de) you can sort results by relevancy, mentions of SNPs, or date of publication. In addition, you can filter and display only those that contain certain information (such as drugs, PPIs, histone modification, etc).



Additional Functionalities:

d. On/off for text mining tool, allowing manual annotation in off mode
e.g. the Domeo Annotation tool includes both ontology matching and manual annotation by selecting the desired term and linking it to the corresponding ontology term (a list of which is suggested by the tool when the term is highlighted).



e. Record time, be able to record time of curation session for each user (need log in as well)
f. Load curation suggestions or warnings for display during curation

- If curator adds GO molecular function “transcription factor”, systems suggests GO celular component (CC) “nucleus” and warns if CC “extracellular space” is entered

g. Upload gene list or ontology term list for focused curation