NCBO Annotator: Semantic Annotation of Biomedical Data

Clement Jonquet, Nigam H. Shah, Cherie H. Youn, Mark A. Musen
Stanford Center for Biomedical Informatics Research
Medical School Office Building, Room X-215
251 Campus Drive, Stanford, CA 94305-5479 USA
{jonquet, nigam, cyoun, musen}@stanford.edu

Chris Callendar, Margaret-Anne Storey
University of Victoria
Department of Computer Science
PO Box 3055, STN CSC
Victoria, B.C. Canada V8W 3P6
chriscallendar@gmail.com, mstorey@uvic.ca

ABSTRACT
The National Center for Biomedical Ontology Annotator is an ontology-based web service for annotation of textual biomedical data with biomedical ontology concepts. The biomedical community can use the Annotator service to tag datasets automatically with concepts from more than 200 ontologies coming from the two most important set of biomedical ontology & terminology repositories: the UMLS Metathesaurus and NCBO BioPortal. Through annotation (or tagging) of datasets with ontology concepts, unstructured free-text data becomes structured and standardized. Such annotations contribute to create a biomedical semantic web that facilitates translational scientific discoveries by integrating annotated data.

Keywords
biomedical ontologies, ontology-based annotation, automatic annotation, semantic annotation, web service, concept recognition, semantic expansion, named entity recognition.

1. ANNOTATION & SEMANTIC WEB
One of the requirements of the semantic web is that web content must be semantically described using ontologies. Semantic annotation is the process that formally identifies concepts and relations between concepts in documents [1]. In this paper, *annotating* refers to the process of describing data with ontology concepts; an annotation is a meta-information that says: *these data is about (or deal with) this concept.* The challenges posed by semantic annotation [2] mean that today's web content is still often composed of unstructured text that is not re-usable by software agents or semantic engines. Furthermore, ontologies and terminologies already exist in several eScience domains and can be used to enrich the web content data description. However, explicitly annotating data with ontology concepts is still not a common practice for several reasons: (i) annotation often needs to be done manually either by expert curators or directly by the authors of the data; (ii) the number of ontologies available for use is large and ontologies change often and frequently overlap; (iii) users do not always know the structure of an ontology's content or how to use the ontology to do the annotation themselves; (iv) annotation can be a boring additional task without immediate reward for the user. Therefore, users need to annotate their data using automatic, easy to use, fast and accurate services that can be integrated into their processes.

One mechanism of achieving automatic ontology-based annotation is to use natural language processing based concept recognizers or named entity recognition tools, to identify the related concepts in the textual metadata describing a data. Once concepts have been identified, relations and mappings between those concepts can be used to expand the set of annotations.

8th International Semantic Web Conference (ISWC 2009) Posters and Demonstrations, October 25-29 2009, Washington DC, USA.

2. BIOMEDICAL CONTEXT
The range of publicly available biomedical data is enormous and is expanding fast which means that researchers now face a hurdle to extracting the data they need. The biomedical community has invested many efforts in text/data mining techniques to process text metadata. However, many translational discoveries that could be made by mining biomedical resources are hampered because most resources typically do not use standard terminologies and ontologies to annotate their elements (i.e., experimental data sets, diagnoses, samples, experimental conditions, clinical-trial descriptions, and papers). This annotation process cannot be easily automated and often requires expert curators. Plus, even if there is a profusion of tools for semantic annotation (e.g., SemanticHacker, OpenCalais, KIM, OntoMat, Magpie) the biomedical domain still lacks easy-to-use systems that facilitate the use of biomedical ontologies for annotation. In this paper, we present a web service that allows scientists to utilize most of the public biomedical ontologies for annotating their datasets automatically [3]. The Annotator web service is publicly available and can be used by the community to tag their own data.[1]

3. NCBO ANNOTATOR
The Annotator workflow is composed of two main steps (Fig. 1). First, the user's free text is given as input to *a concept recognition tool* along with a dictionary. The dictionary (or lexicon) is a list of strings that identifies ontology concepts. The dictionary is constructed by accessing ontologies and pooling all concept names or other string forms (synonyms, labels) that syntactically identify concepts. The Annotator uses Mgrep[2] [4] to recognize concepts by using string matching on the dictionary. This primary set of direct annotations serves as input for the *semantic expansion components*, which expanse the annotations extracted from the first step using the knowledge represented in one or more ontologies. For example:

• An *is_a transitive closure* component traverses an ontology parent-child hierarchy to create new annotations with parent concepts of the concepts involved in direct annotations. For example, if data are directly annotated with the concept `melanoma` from NCI Thesaurus, this component can generate new annotations with concepts `skin tumor` and `neoplasms` because NCI Thesaurus provides the knowledge that `melanoma` *is_a* `skin tumor` and `skin tumor` *is_a* `neoplasms`. The maximum level in the hierarchy to use is parameterizable i.e., until which ancestor the annotation should be expanded.

[1] http://bioportal.bioontology.org/annotate

[2] Mgrep (NCIBI at the Univ. f Michigan) has a high degree of accuracy (>95% precision) in recognizing disease names. We are working to make other concept recognizers pluggable in the Annotator workflow.

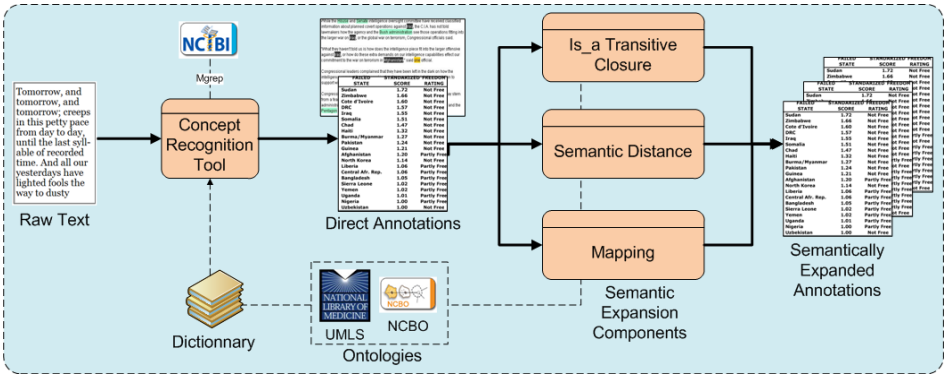

Fig. 1. NCBO Annotator workflow. First, direct annotations are created from raw text based on syntactic concept recognition according to a dictionary that use terms (concept names and synonyms) from both UMLS and NCBO BioPortal ontologies. Second, different components expand the first set of annotations step using the knowledge represented in one or more ontologies.

• An *ontology-mapping* component creates new annotations based on pre-existing mappings (available in UMLS & BioPortal) between ontologies. For example, if text is directly annotated with the concept `NCI/C0025202` (`melanoma` in NCI Thesaurus), this component can generate new annotations with concepts `SNOMEDCT/C0025202` (`melanoma` in SNOMED-CT) and `38865/DOID:1909` (`melanoma` Hunan disease) because a one-to-one mapping exists between those concepts. The type of mapping to use is parameterizable i.e., where does the mapping come from (UMLS CUI based, human created, etc.).

• A *semantic distance* component (still in development) will use semantic similarity measures (e.g., Rada, Resnik) between concepts to obtain related concepts and create new annotations.

We have implemented the Annotator web service using (at the time of writing) 207 biomedical ontologies & terminologies. Those ontologies offer a dictionary of 4,021,662 concepts and 7,637,125 terms. The Annotator web service is embedded in BioPortal (Fig. 2) [5]. Annotations, scored according to both their frequency and the context in which they have been generated (e.g., direct/indirect, is_a/mapping) and can be returned to the user in different formats (text, tab delimited, XML, or OWL).

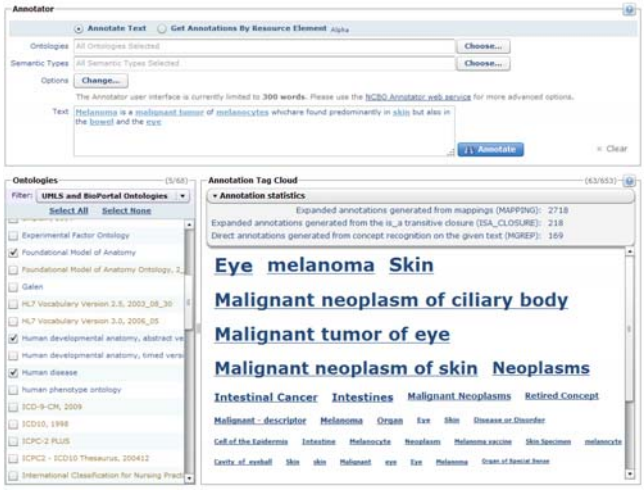

Fig. 2. NCBO Annotator user interface in BioPortal. A user adds the text to annotate in the text area and select the parameters to use for annotations (ontologies, UMLS semantic types, hierarchy level, mappings). A tag cloud of concepts is then generated to represent the annotations. The bigger a concept in the cloud is, the higher the score of the corresponding annotation is.

The Annotator provides a novel contribution to preexisting tools as it: (i) is clearly positioned as an easy to use pluggable service-oriented tool; (ii) leverages the knowledge embedded in ontologies (concept recognition & semantic expansion); (iii) is compliant with semantic web standards[3], and (iv) has access (and provides an abstract common access) to a large set of biomedical ontologies. The Annotator is currently used by NCBO to index biomedical resources and enhance information retrieval and data integration in the biomedical domain. In addition, it is being used and evaluated by eight external biomedical informatics groups.

We have not conducted detailed research as it is often done in the biomedical informatics community using a specific data resource and a specific ontology. Instead, the Annotator tries to address the real issue of semantic annotations of biomedical data on a large scale, for various resources and multiple ontologies in order to provide users with a service they can concretely use. Although the methodology is not domain dependent, this work is a good illustration of applied eScience research where semantic web technologies are used in practice.

This work is supported by the National Center for Biomedical Computing (NCBC)/ National Institute of Health roadmap initiative; NIH grant U54 HG004028.

4. REFERENCES

[1] Uren, V., et al., Semantic annotation for knowledge management: Requirements and a survey of the state of the art. Web Semantics: Science, Services and Agents on the WWW 4(1), (Jan 2006) 14–28

[2] Handschuh, S., Staab, S., eds.: Annotation for the Semantic Web. Vol. 96 of Frontiers in Artificial Intelligence and Applications. IOS Press (2003)

[3] Jonquet, C., Shah, N. H., Musen, M. A., The Open Biomedical Annotator, AMIA Summit on Translational Bioinformatics, p. 56-60, March 2009, San Francisco, USA.

[4] Dai, M., et al., An Efficient Solution for Mapping Free Text to Ontology Terms. AMIA Summit on Translational Bioinformatics, March 2008, San Francisco, USA.

[5] Noy, N. F. et al., BioPortal: Ontologies and Integrated Data Resources at the Click of a Mouse, Nucleic Acids Research, (May 2009), 37, web server issue.

[3] OWL annotations are described as instances of an OWL ontology: http://obs.bioontology.org/ontologies/NCBO_OBS_ontology.owl. This feature is available only in the prototype environment for now.