Prototyping a Biomedical Ontology Recommender Service

Clement Jonquet[†]*, Nigam H. Shah*[†]* *and Mark A. Musen*

Center for Biomedical Informatics Research, Stanford University, CA 94305, USA

ABSTRACT

As the use of ontologies for annotation of biomedical datasets rises, a common question researchers face is that of identifying which ontologies are relevant to annotate their datasets. The number and variety of biomedical ontologies is now quite large and it is cumbersome for a scientist to figure out which ontology to (re)use in their annotation tasks. In this paper we describe an early version of an ontology recommender service, which informs the user of the most appropriate ontologies relevant for their given dataset. We provide results to illustrate that situation. The recommender service uses a semantic annotation based approach and scores the ontologies according to those annotations. The prototype service can recommend ontologies from UMLS and the NCBO BioPortal and is accessible at http://bioontology.org/tools.html

1 INTRODUCTION

Biomedical ontologies are widely used to design information retrieval systems, to facilitate the interoperability between different repositories, and to develop systems that parse, annotate or index available biomedical data resources. Biomedical researchers use ontologies and terminologies to structure and annotate their data with ontology concepts for better data integration and translational discoveries. However, the number and variety of biomedical ontologies is now large enough that it becomes cumbersome for a scientist to figure out which ontology to (re)use in their annotation tasks. Often, the scientist does not know about any ontologies or does not know about new versions or new ontologies that might be appropriate for his/her application. Members of NCBO often get requests for suggesting the 'optimal ontology' for a certain domain; a task made difficult by the large number of ontologies available. For the scientist, the process to choose a set of ontologies to use is oftentimes a hard, manual and time consuming task. The consequences of getting that fundamental task wrong in the scientist's work plan could be very bad. For example, the scientist could: (i) miss a number of relevant ontologies, and would have to reannotate the dataset; (ii) start designing a new ontology instead of re-using a standard and shared one; (iii) miss insights that would be achieved by using the right ontologies to link his/her data with other datsets.

To facilitate the task of selecting the appropriate ontologies to use in an annotation task, we prototyped an ontology recommender service, which – given textual metadata describing elements of a dataset – will recommend the appropriate ontologies to annotate and tag the given dataset. The ontology recommender uses a method based on annotations to score the appropriate ontologies. An annotation maps elements of a dataset (e.g. papers, GEO experiments, Clinical Trial records etc) to ontology concepts and declares: *this data element "is associated with" this concept*. The National Center for Biomedical Ontology (NCBO) has developed the Open Biomedical Annotator (OBA) web service (Jonquet et al 2009), to annotate data elements with ontology concepts – based on their textual descriptions – using one of the largest available set of biomedical ontologies (Combination of ontologies in the Unified Medical Language System (UMLS) Metathesaurus and the NCBO BioPortal)[1]. The biomedical community can use the annotator service to tag datasets automatically with ontology terms[2]. Internally, NCBO uses the annotation workflow to index biomedical data resources with ontology concepts (Shah et al 2009).

In this paper, we show how the annotator service can be used to implement an ontology recommender service by aggregating all the annotations done with the same ontologies. The information provided by the recommender service is simplified and customized to facilitate the task of choosing the right ontology to use. The biomedical ontology recommender service is deployed as a Representational state transfer (REST) web service in order to be embedded in automatic workflows. It can be also be used through a user interface.

2 METHODS

The recommender service accepts biomedical text data as input and suggests the most appropriate ontologies relevant for the given data. The annotations used to generate the rec-

[†] These authors contributed equally.

* To whom correspondence should be addressed

[1] The NCBO BioPortal (Musen et al 2008) is a web repository of biomedical ontologies. Users can browse, search, and comment (social web) ontologies both online and via a web services application programming interface. The UMLS Metathesaurus is a collection of concepts, terms and their relationships from various controlled vocabularies.

[2] http://gminer.mcw.edu/

ommendation are produced by the annotator service described hereafter.

2.1 The Open Biomedical Annotator

To facilitate the annotation of biomedical datasets, NCBO developed the Open Biomedical Annotator (OBA), a web service that processes the textual metadata of records in public datasets in order to annotate those records with biomedical ontology concepts. For a given chunk of text the annotator will assign ontology concepts as annotations and return them to the users. The OBA service's workflow is composed of 2 main steps. First, direct annotations are created from raw text based on syntactic *concept recognition* based on a dictionary compiled from terms (concept names and synonyms) pulled from the ontologies. Second, different *semantic expansion components* leverage the semantics in ontologies (e.g., is_a relations and mappings) to create additional annotations.

The annotation workflow is parameterized to enable selection of ontologies from one of the largest available set of biomedical ontologies. We have implemented the service using the 98 English ontologies in UMLS 2008AA and a subset of the BioPortal ontologies (92 at the moment of writing). Those ontologies offer a dictionary of 3,582,434 concepts and 7,024,618 terms. The annotator returns annotations in several formats like tab delimited, XML and RDF/OWL. Annotations are scored according to the context (e.g. Title, summary, description) from which they have been generated and returned to the user.

The score is a number assigned to an annotation and reflects the accuracy of the annotation. The higher the score is the better the annotation is. The scoring algorithm gives a specific weight to an annotation according to the context of the annotation as well as the matched term. For instance, an annotation done by matching a concept's preferred name will be given a higher weight than an annotation done by matching a concept's synonym or than an annotation done with an parent level 3 (ancestor) term in the is_a hierarchy. The final score for an annotation is the sum of all the weights corresponding to the annotations done with that same concept for a certain piece of text. The weights used by the scoring algorithm are described in Table 1.

Table 1. Annotation weights per context

Annotation context	Weights
direct annotation done with the concept preferred name	10
direct annotation done with a concept synonym	8
expanded annotation done with a mapping	7
expanded annotation done with a direct parent concept (parent level 1),	8
expanded annotation done with an ancestor (level < 3)	7
idem (level < 5)	6
idem (level < 7)	5
idem (level < 15)	3
idem (level >= 15)	1

2.2 The ontology scoring method

The recommender service first uses the Open Biomedical Annotator service to annotate the user supplied text with the ontologies available. The user can either choose to use all the UMLS Metathesaurus ontologies or all the BioPortal ones. In the second step the recommender service sorts the ontologies based on the sum of the scores of the annotations generated with concepts from a particular ontology.

2.3 Example

Consider for example the text "*Melanoma is a malignant tumor of melanocytes which are found predominantly in skin but also in the bowel and the eye*". This sentence upon annotation with OBA will generate direct annotations (i.e., string matching with dictionary) with concepts such as:[3]

- `NCI/C0025201`, *Melanocyte* in NCI Thesaurus {10}
- `NCI/C0025202`, *Melanoma* in NCI Thesaurus {10};
- `NCI/C0027651`, *Neoplasm* (synonym of *Tumor*) in NCI Thesaurus {8};
- `39228/DOID:1909`, *Melanoma* in Human Disease {10};

The is_a closure expansion will generate the annotations:

- `39228/DOID:191`, *Melanocytic neoplasm*, direct parent (level 1) of *Melanoma* in Human Disease {8};
- `39228/DOID:0000818`, *cell proliferation disease*, grand-parent (level 2) of *Melanoma* in Human Disease {8};
- `NCI/C0027651`, *Neoplasms* in NCI Thesaurus, grand-grand-parent (level 3) of `NCI/C0025202` in NCI Thesaurus {7};

The mapping expansion will generate annotations such as:

- `FMA/C0025201`, *Melanocyte* in Foundational Model of Anatomy, concept mapped to `NCI/C0025201` in UMLS {7}.

The scores of the annotations with `NCI/C0025201` and `NCI/C0025202` will be 10 where as the score of the annotation with `NCI/C0027651` will be 15 (8+7) as the annotation was generated twice both because of a synonym and because of a descendent. The final score computed as the sums of the annotations scores per ontology will be:

[3] Weights of the annotations are detailed between brackets {w}. UMLS Metathesaurus concepts are identified by ABbreviated Source name (SAB) and Concept Unique Identifier (CUI). NCBO BioPortal concepts are identified by Uniform Resource Identifier (URI).

- NCI Thesaurus (NCI): 50
- Human Disease (39228): 26
- Foundational Model of Anatomy (FMA): 7

Figure 1 shows the results for the example text in the recommender service user interface. (note that the results presented in Figure 1 use all the UMLS ontologies and therefore contain much more annotations than the ones presented in the previous example. SNOMED-CT is the highest scored ontology in the results shown in the figure. Exact parameters used for the Annotator service call are described in the Appendix).

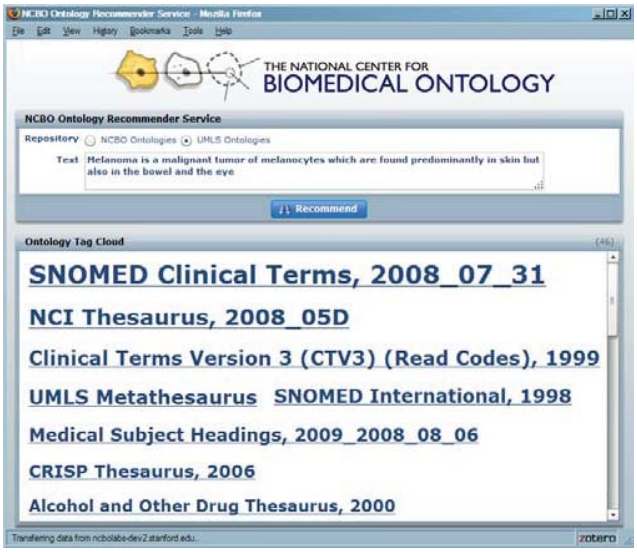

Figure 1. Ontology recommender web service user interface. The user can select the repository of ontologies to use (UMLS/NCBO) and enter the text to recommend. A tag cloud is generated in which the score of an ontology is represented by the size of its name in the cloud.

3 RESULTS AND DISCUSSION

In order to illustrate the importance of appropriate ontology recommendation, we present results obtained with the ontology recommender on three different types of biomedical datasets about the same topic:

- **Literature:** Top 10 articles from PubMed (PMID) obtained with the query "melanoma AND skin disease" for which the fields title and abstract have been concatenated;
- **Clinical:** Top 10 Clinicaltrials.gov trials (NCT)) obtained with the query "melanoma AND skin disease" for which the fields title, purpose, condition and intervention have been concatenated;
- **High throughput experiments:** Top 10 Gene Expression Omnibus datasets (GDS) obtained with the query "melanoma" for which the fields title and summary have been concatenated;

For each of these types of "datasets" a scientist could ask the question: which ontologies are relevant to annotate (or "tag") the elements. The ontology recommender service results for each set are presented respectively in Table 2, Table 3 and Table 4. Only the top 10 ontologies returned by the service are presented here with their score and the number of annotations that contributed in their scores. In these examples, only BioPortal ontologies have been used.

On examining the results, the following observations stand out: 1) the recommender service gives a high score to big ontologies (such as NCI Thesaurus).[4] Indeed, because of the high number of concepts in those ontologies they are more appropriate to fully markup or tag the textual descriptions submitted by the user. 2) Some ontologies are recommended high in the results, regardless of the source of the underlying textual-data (e.g., NCI Thesaurus, Human disease). 3) Some ontologies appear only with a specific type of data (e.g., anatomy ontologies for clinical trials) which illustrate the importance of appropriate recommendation. We note that almost all anatomy ontologies show up including model organism ones. 4) An ontology may have a better overall score than another one even if the numbers of annotations from it are small (e.g. Human disease ontology). This finding illustrates the importance of scoring as well as the annotation context weights.

We also note that the results of the recommender are dependent on the accuracy of the Open Biomedical Annotator, which uses lexical matching for concept recognition and the limitations that go with it (Jonquet et al 2009).

Table 2. Recommender results for PubMed articles

Ontology name	Annotations	Score
NCI Thesaurus	2047	448223
Human disease	246	87214
Galen	802	61754
Experimental Factor Ontology	185	23228
RadLex	252	21310
Human phenotype ontology	114	17699
Phenotypic quality	227	15390
Units of measurement	230	14833
Mouse pathology	55	14749
Suggested Ontology for Pharmacogenomics	226	13944

Table 3. Recommender results for clinical trials

Ontology name	Annotations	Score
NCI Thesaurus	1141	123649
Human developmental anatomy, timed version	292	76836
Human disease	333	43161

[4] This is also illustrated with Figure 1 with ontologies from UMLS.

Xenopus anatomy and development	157	35740
Experimental Factor Ontology	188	29997
Galen	445	22652
Foundational Model of Anatomy	122	20416
Nci anatomy after fix	69	19285
Mouse adult gross anatomy	82	19229
Medaka fish anatomy and development	70	19084

Table 4. Recommender results for GEO series

Ontology name	Annotations	Score
NCI Thesaurus	571	41072
Human disease	202	30308
Galen	239	11811
Experimental Factor Ontology	112	8749
Human developmental anatomy, timed version	305	7756
Human phenotype ontology	52	6002
Mouse pathology	44	5848
Zebrafish anatomy and development	112	4746
Mosquito gross anatomy	87	3597
RadLex	86	3319

4 RELATED WORK

Alani et al address the problem of ontology search, i.e. finding ontologies from an ontology repository that are relevant to the user's query (Alani et al 2007). They examine the case when users search for ontologies relevant to a particular *topic* (e.g., an ontology about anatomy). In their approach, when looking for ontologies on a particular topic (e.g., anatomy), they retrieve, from the Web, a collection of terms that represent the given domain (e.g., terms such as body, brain, skin, etc. for anatomy). The terms are then used to expand the user query and search existing ontologies. Their results show an improvement in retrieval results by 113%, compared to the tools (e.g. Swoogle) that search only for the user query terms and consider only class and property names. Our approach is quite different and we do not search for the best ontologies for "anatomy", but aim to inform the user about what ontologies might be worth considering for annotating or tagging the data elements under consideration.

5 CONCLUSION AND FUTURE WORK

We have presented a prototype of a biomedical ontology recommender service, which – based on given textual descriptions of element of a dataset – informs the user of the most appropriate ontologies to annotate the dataset. This approach, to the best of our knowledge, is unique for ontology recommendation. Our approach uses both (i) a syntactic concept recognition step (string matching with a dictionary) and (ii) a semantic expansion step, which utilizes the knowledge in ontologies to generate new annotations. The ontology recommender service can recommend ontologies from over 190 biomedical ontologies and terminologies contained the UMLS and the NCBO BioPortal (Musen et al 2008).

In the future, we envision several directions for further work on the recommender service:

- Investigating different scoring methods that will support different kinds of recommendation scenarios. A good feature would be to be able to recommend very specific small ontologies. One way to do that would be to use the size of the ontologies to normalize the score.
- Allowing parameterized scoring methods i.e., users can customize weights given to each context.

Currently, we are in the process of evaluating the results and the utility of the recommender service.

6 APPENDIX

Parameters to give to the Open Biomedical Annotator to get annotations used by the ontology recommender service (non specified parameters are set to default values):

```
withDefaultStopWords = true
minTermSize = 4
localSemanticTypeIDs = T000 (for UMLS repository) | T999 (for NCBO repository)
levelMax = 5
activateMapping = true
```

ACKNOWLEDGEMENTS

This work is supported by NIH grant U54 HG004028 in support of the National Center for Biomedical Ontology. We acknowledge the assistance of Manhong Dai and Fan Meng (NCIBI) and Chris Callendar (Univ. of Victoria, CA).

REFERENCES

Shah NH, Jonquet C, Chiang AP, Butte AJ, Chen R, Musen MA., Ontology-driven Indexing of Public Datasets for Translational Bioinformatics, *BMC Bioinformatics*, Vol. 10, February 2009.

Jonquet C, Shah NH, Musen MA, The Open Biomedical Annotator, *AMIA Summit on Translational Bioinformatics,* p. 56-60, March 2009, San Francisco, CA, USA.

Musen MA, Shah NH, Noy N, Dai B, Dorf M, Griffith N, Buntrock JD, Jonquet C, Montegut MJ, Rubin DL, BioPortal: Ontologies and Data Resources with the Click of a Mouse, *AMIA 2008 Annual Symposium, Demonstrations*, p.1223-1224, Washington DC, USA, November 2008

Olivier Bodenreider' The Unified Medical Language System (UMLS): integrating biomedical terminology, *Nucleic Acids Research*, Vol. 32, Database issue, 2004

Alani, H., Noy, N., Shah, N., Shadbolt, N. and Musen, M. (2007) Searching Ontologies Based on Content: Experiments in the Biomedical Domain. In: *The Fourth International Conference on Knowledge Capture (K-Cap)*, Whistler, BC, Canada.