Help will be provided for this task: Ontology-Based Annotator Web Service

Clement Jonquet, Mark A. Musen, Nigam H. Shah

Help will be provided for this task: Ontology-Based Annotator Web Service

Clement Jonquet, Mark A. Musen, Nigam H. Shah

Stanford Center for Biomedical Informatics Research
Stanford University School of Medicine
Medical School Office Building, Room X-215
251 Campus Drive, Stanford, CA 94305-5479 USA
{jonquet, musen, nigam}@stanford.edu

Abstract. Semantic annotation is part of the vision for the semantic web. Ontologies are required for this task, and although they are in common use, there is a lack of annotation tools for users that are convenient, simple to use and easily integrated into their processes. This paper presents an *ontology-based annotator web service* methodology that can annotate a piece of text with ontology concepts and return annotations in OWL. Currently, the annotation workflow is based on syntactic concept recognition (using concept names and synonyms) and on a set of semantic expansion algorithms that leverage the semantics in ontologies (e.g., is_a relations). The paper also describes an implementation of this service for life sciences and biomedicine. Our *biomedical annotator* service uses one of the largest available set of publicly available terminologies and ontologies. We used it to create an index of open biomedical resources. Both the deployed web service and a user interface can be accessed at (http://www.bioontology.org/tools.html).

Keywords: ontologies, ontology-based/semantic annotation, web service, concept recognition, biomedical ontologies, semantic expansion, semantic web service, service-oriented architecture.

1 Introduction

One of the requirements of the semantic web is that web content must be semantically described using ontologies. Since 2002, the call for papers for the International Semantic Web Conference has contained the following sentence: *Authors of accepted papers will be required to provide semantic annotations for the abstract of their submission for the Semantic Web* (prior to 2005, "mark up" was used instead of annotations). However, this sentence is followed by the following parenthetical remark: *help will be provided for this task*. This indicates that, in general, annotation is not an easy task. In the case of the ISWC call for papers, it needs to be done manually, with the contribution of the authors. It cannot be done entirely automatically when papers are submitted. The challenges posed by semantic annotation [1] mean that today's web content is still mainly composed of unstructured text that is not re-usable by software agents or semantic engines.

Ontologies and terminologies already exist in several eScience domains and can be used to enrich the web content data description. As a result, semantically annotating web data using ontologies is becoming an important task. For example, in the biomedical domain, the variety of publicly available data is already enormous and is expanding very fast. This expansion means that researchers now face a hurdle to extracting the data they need from the large amount of data that is available. Biomedical researchers have turned to ontologies and terminologies to describe their data and turn it into structured and formalized knowledge [3]. For instance, the Gene Ontology is widely used to describe the molecular functions, cellular locations, and biological processes of gene products. The Gene Ontology allows integrating these descriptions across several databases. As another example, when a new PubMed[1] citation is created, its title and abstract are indexed with MeSH[2] terms. However, besides Gene Ontology annotations and PubMed indexing most available biomedical data are unstructured and rarely described with ontology concepts.

In their study, Uren et al. define semantic annotation as the process that formally identifies concepts and relations between concepts in documents [4]. In this paper, *annotating* refers to the process of describing data with ontology concepts; an annotation is a layer of meta-information on data that says: *this data deals with this concept*. Explicitly annotating data with ontology concepts is still not a common practice for several reasons:

- The numbers of relevant ontologies are increasing and getting access to all of them may become cumbersome because of different formats, locations or application programming interfaces (APIs);
- Users do not always know the structure of an ontology's content or how to use it in order to do the annotation themselves;
- Annotating data using ontologies is often a boring additional task without immediate reward for the user.

Both inspired by the semantic web and the importance of service-oriented computing, we present a web service methodology that allows users to utilize available ontologies for annotating their data automatically. The *Ontology-Based Annotator* (OBA) web service automatically processes a piece of raw text to annotate (or tag) it with relevant ontology concepts and return the annotations. The OBA web service is an *ontology-based service* that uses ontologies to produce a new output.[3] Plus, the service delivers semantic web enabled results, because it creates annotations that are semantically described from raw text. The service workflow is composed of two main steps: (i) the concept recognition step that syntactically identifies concepts from their names or synonymous terms;[4] (ii) the semantic expansion step, where the first set of annotations is expanded using the

[1] PubMed (www.ncbi.nlm.nih.gov/PubMed/) is the standard literature database of biomedicine.
[2] MeSH (Medical Subject Headings) is a biomedical controlled vocabulary created and updated by the United States National Library of Medicine (NLM).
[3] Ontology-based medical search engines are a good example of ontology-based services [5,6,7].
[4] A *concept* is unique in an ontology (class). A *term* is a particular string form that identifies a concept. Usually, a concept has several terms (e.g., name, synonyms, label).

knowledge from one or several ontologies (e.g., is_a relations), to a larger set of annotations according to different expansion algorithms. The OBA web service distinguishes itself from what has been previously reported for several reasons:

1. It is an automatic web service that can be integrated in current workflow and used by software agents,
2. It leverages ontologies to create new annotations,
3. It is described by a specific service model ontology and returns annotations as an OWL ontology populated with annotations as instances,
4. It has access to multiple ontologies (biomedical ontologies in our application).

We have deployed an application of the OBA web service for the biomedical domain. As stated above, the need for such a service in the life sciences and biomedicine is critical. Our previous experience [8,9] has allowed us to identify appropriate use cases for our service such as annotating pathology samples, high-throughput datasets and clinical trial reports. Our prototype uses 47 publicly available terminologies and ontologies from the Unified Medical Language System (UMLS) and the National Center for Biomedical Ontology (NCBO) repository. It uses *mgrep*, a concept recognizer with a high degree of accuracy (>95%) in recognizing disease names [10]. The service is deployed as a SOAP (Simple Object Access Protocol) web service as well as a RESTful (REpresentational State Transfer) web service. An OWL (Web Ontology Language) ontology defines the service model. We have used the annotator service internally to process several biomedical resources and have constructed an ontology-based index that allows a user to search for biomedical data annotated with ontology concepts [2].

2 Ontology-based annotator description and design

Automatically annotating a user's data for the semantic web poses technical challenges such as scalability, accuracy, and maintenance. For example, the number of ontologies available for use is large and, many ontologies change often and overlap as well. The ontologies are not in the same format and do not always provide APIs to query them. Users need services that abstract ontology formats or access mechanisms and matches with the *service oriented computing* principles [11]. The OBA is such a service. This section presents the ontology-based annotating workflow as well as the service model.

2.1. Ontology-based annotator web service workflow

Fig. 1 describes the OBA web service workflow, which is composed of two main steps. First, the user's 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. It is constructed by accessing several ontologies and dumping all concept names or other string forms, called terms, such as synonyms or labels that syntactically identify concepts. The choice of the set of ontologies used to create the dictionary depends of the domain for which the OBA web service is deployed. The tool recognizes concepts by using string matching and syntax based techniques such as stemming, spelling, or morphological forms. The output is a set of *direct annotations*.

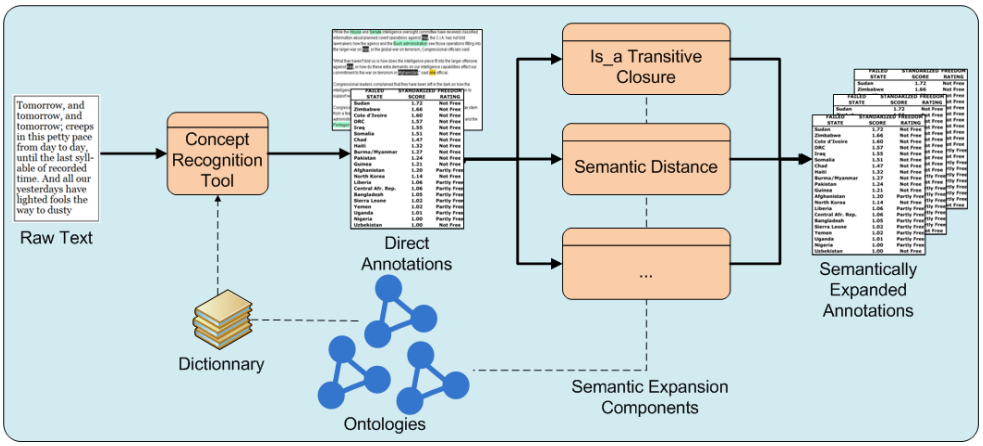

Fig. 1. OBA web service workflow. First, direct annotations are created from raw text according to a dictionary that use terms from a set of ontologies. Second, different components expand the first set of annotations using ontologies semantics.

Second, this primary set of annotations serves as input for the *semantic expansion components*. These components enhance the annotations extracted from the first step using the semantics of one or more ontologies. There may be different types of expansion components, such as:

- An *is_a transitive closure* component that traverses an ontology parent-child hierarchy up to the root to create new annotations with parent concepts. For instance, if a data is annotated with a concept from the National Cancer Institute Thesaurus (NCIT), 'melanoma,' the tool generates a new annotation with the concept 'neoplasm' because the NCIT provides the knowledge that 'melanoma' is_a 'neoplasm.' The component assumes that a document dealing with a child concept also deals with the parent concept(s).
- A *semantic distance* component creates new annotations by obtaining related concepts according to a given semantic distance in an ontology [12,13]. For example, a semantic distance algorithm can evaluate at 1 the distance between siblings in an ontology and 4 the distance between cousins (i.e., child of parents sibling). A corresponding semantic distance component, configured to return all the concepts that are at the maximum distance 2 will return all the siblings of a concept but not the cousins.
- An *ontology mapping* component creates new annotations based on existing mappings between different ontologies [14].

The OBA web service is designed in manner that allows multiple semantic expansion components to be plugged-in, selected and parameterized by a user when requesting the service. These components may also be composed to create new component. The OBA *service level agreement* depends on the selected components as each consumes resources

at a different level. For example, the is_a transitive closure takes a long time to process, even when using a pre-computed hierarchy table.

As an output of the second step, several sets of *semantically expanded annotations* are extracted and returned, along with direct annotations, to the user. The system may classify this final set of annotations by accounting for the frequency with which a concept has been identified directly by the concept recognizer (i.e., number of occurrence of a concept in a text) or by semantic expansion components (i.e., number of time an annotation was founded by expansion).[5]

2.2. Ontology-based annotator web service model

In this section we define the OBA web service model i.e., what the service returns to the user: the objects as well as their relations and the constraints that applies. Fig. 2 describes the OBA web service model as a Unified Modelling Language (UML) class diagram. The model is defined by a set of six objects:

- `ResultBean` is the main object returned by the OBA web service. It is a representation of an OBA result. `text` refers to the piece of text originally sent to the service, while `name` identifies the result. `contextNb` is a constant that defines the number of different types of context possible for this result (see below) i.e., the number (`x`) of components selected for the annotation, + 1 (for direct annotations). The properties `annotations`, `dictionary`, `ontologies` and `statistics` are defined hereafter. A `ResultBean` provides functions to export its content in different form for the user. These functions correspond to the four different calls a user can make via the OBA web service API. The `toText()`, `toTabDelimited()` and `toXML()` functions return the result content without any formal semantics. `toOWL()` returns the content in a structured and semantically rich form (see section 3).
- `AnnotationBean` is a representation of one annotation. `conceptID` globally identifies (e.g., URI) the unique annotating concept that forms the annotation. `conceptTerms` specifies all the possible terms for the annotating concept that have been used in the dictionary. `context` asserts the context information for the annotation.
- `ContextBean` has information about the context in which an annotation was created. For example, it could be a direct annotation or could have been made by a semantic expansion component. `contextName` is a keyword (e.g., `DIRECT`, `ISA_CLOSURE`) which identifies the type of annotation. `conceptID` identifies the concept from which an annotation with a particular context was derived. For instance, if the annotation was produced by the is_a transitive closure component, `name` is `ISA_CLOSURE` and `conceptID` identifies the child concept from which the is_a relation was used to produce this annotation. `ContextBean` objects are strongly related to semantic expansion components in the sense that they specify from which component an annotation has been produced. Therefore, they may be adapted to fit with specific components added to the OBA web service workflow.

[5] Information retrieval techniques (frequency, inverse document frequency) can also be used here.

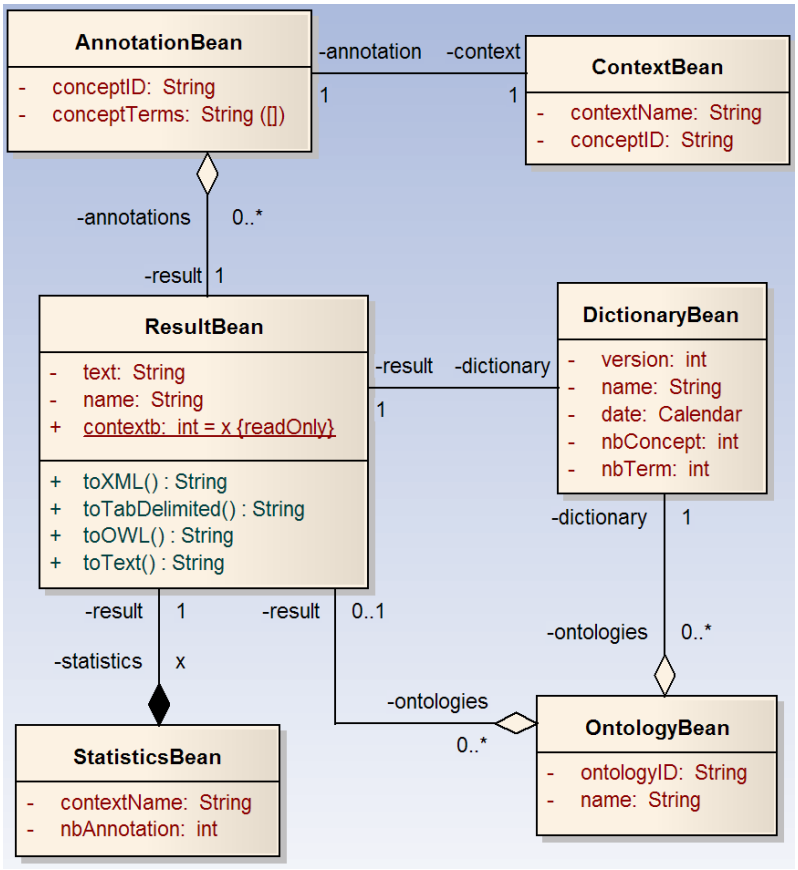

Fig. 2. OBA web service UML model.

- `DictionaryBean` contains the metadata (not the content) of the dictionary used for a result. `version`, `name`, and `date` identify the dictionary on the server side and give information about its content. `nbConcept` and `nbTerm` represent the number of concepts and terms in a dictionary and `ontologies` specifies the set of ontologies used in it. Dictionary versioning is strongly linked to the evolution of the ontologies used. Each time ontologies change, the dictionary is updated. All the dictionary information may be useful for comparing results of the OBA web service on time.
- `OntologyBean` is a representation of an ontology. To keep the model simple, we provide only two key pieces of information: `ontologyID` and `name`. If the ontology was specifically involved in creating annotations returned as a result of an OBA web service call, the ontologies is associated to a `result`.

- `StatisticsBean` contains information on the number of annotations done for a given context. The `contextName` keyword identifies the type of context and `nbAnnotation` is the number of annotations of this type.

The model introduced in this section presents the simple semantics of the OBA web service. Most of the constraints can be represented by beans in a Web Services Description Language (WSDL) description. For example, data types used and most of the cardinalities are straightforward and can be expressed through a WSDL description. However, some aspects of the model semantics cannot be expressed by a simple WSDL description. Data type restrictions are an example, as when `version` must be a positive integer, and `conceptID` must be an URI. Similarly, it is not simple to express that an ontology involved in a specific result must be part of the ontologies that form the dictionary of this result. There is also the question of how to represent different `ContextBean` (or `OntologyBean`) objects with inheritance. These examples require a higher semantic description for the OBA web service to allow users and software agents to manipulate service results.

3 Semantically described results

The OBA web service provides a semantic layer that allows a formal description of the service model and results. We have developed an OWL ontology that formalizes the OBA web service model (section 2.2) and implements constraints and restrictions on it. For example, the fact that an ontology involved in a result must be part of the ontologies that form the result's dictionary can easily be expressed by a SWRL rule (Semantic Web Rule Language) [15]:

```
ResultBean(?r)^DictionaryBean(?d)^OntologyBean(?o)
   ^dictionary(?r,?d)^ontologies(?r,?o)->ontologies(?d,?o)
```

The OBA web service ontology uses OWL to semantically define objects returned by the service. When the OBA server compiles a result for a user, it populates its ontology with instances and directly returns the description or its URI location, and the instances as a specific OWL file. The couple ontology/instances works well for this type of application as it allows a clean separation of the data and the semantics defining the data. For example, the following annotation states that the concept `PATO:0000051` (morphology), from the `Unit` ontology, is an annotation included in the result `OBA_RESULTBEAN_d45dc8`. It is an `ISA_CLOSURE` annotation derived from the concept `PATO:0000918` (i.e., the `Unit` ontology provides the knowledge `PATO:0000918` is_a `PATO:0000051`). The OWL representation of this annotation is:

```
<AnnotationBean rdf:about="#OBA_AnnotationBean_624">
   <conceptID rdf:datatype="http://www.w3...#string">
      Unit:::PATO:0000051 </conceptID>
   <conceptTerms rdf:datatype="http://www.w3...#string">
      morphology </conceptTerms>
   <result rdf:resource="#OBA_RESULTBEAN_d45dc8"/>
```

```
   <context>
      <ContextBean rdf:about="#OBA_ContextBean_624">
         <contextName rdf:datatype="http://www.w3...#string">
            ISA_CLOSURE </contextName>
         <conceptID rdf:datatype="http://www.w3...#string">
            Unit:::PATO:0000918 </conceptID>
         <annotation rdf:resource="#OBA_AnnotationBean_624"/>
      </ContextBean>
   </context>
</AnnotationBean>
```

Parsing the OWL file with an OWL enabled API such as Jena (http://jena.sourceforge.net/) or the Protégé-OWL API (http://protege.stanford.edu/) enables a human user or a software agent to fully utilize the content of the OBA web service result and eventually integrate several results together. In addition, the entire semantic web technology stack (SPARQL, RDF tools, etc.) can be used to process the result as it is returned in a standard format. For example, a user can query OWL results to get all annotations done directly with or derived from one `conceptID`.

4 Application: a biomedical annotator service

The emergence of information and communication technologies has drastically changed biomedical research processes. Experimental data and results are easy to share and repurpose thanks to web APIs enabling connections to large databases. As a consequence, the biomedical data available in the public domain is now diverse and growing rapidly. This expansion has motivated research on data integration, and the research community agrees that ontologies are essential for data integration to occur [16,17,18,19,20]. However, although the variety of biomedical data is very large (from experimental data sets in repositories, to records of disease associations of gene products in mutation databases, to entries of clinical-trial descriptions, to published papers, and so on), it often contains free text meta-data information entered by the researcher who created it. This situation creates a challenge of producing consistent terminology or ontology labels for each element in the public biomedical resources. Such labels would enable the identification of all related elements at a given level of granularity. One mechanism of achieving consistency would be to map text metadata describing the resource element to ontology concepts allowing formulation of refined or coarse search criteria. Creating ontology-based annotations from this metadata will enable users to formulate flexible searches for biomedical data [21,6,8]. In this section, we present a prototype of the OBA web service that can address some of the biomedical community needs.

4.1. Implemented prototype

The National Center for Biomedical Ontology (NCBO) [3] maintains *BioPortal*, a web application for accessing biomedical ontologies. BioPortal contains a large collection of ontologies, such as Gene Ontology, the National Cancer Institute Thesaurus, and the

International Classification of Diseases, in different formats (OBO, OWL, etc.). Users can browse and search this repository of ontologies both online and via a web services API. BioPortal ontologies along with the Unified Medical Language System (UMLS)[6] terminologies provide a key element for our *biomedical annotator service*: one of the largest set of biomedical terminologies and ontologies publicly available for our dictionary.

The complex task in building the dictionary is mainly the access to the ontologies. As previously stated, ontologies are spread all over the web and are defined with different languages that use different conventions. For example, the Open-Biomedical Ontology (OBO) format provides a specific field for synonyms facilitating the dumping of all the terms that define a concept. This is not the case for OWL ontologies for which a specialist has to decide the relevant property (e.g., `rdfs:label`, `skos:prefLabel`, `skos:altLabel`) that represent a term for a concept. Additionally, the dictionary is a representation of the content of several ontologies at a given time. It has to follow the ontologies evolution. For instance, if a concept is removed from an ontology, the corresponding terms in the dictionary needs also to be removed. In our prototype the complexity of accessing ontologies to create and update the dictionary is significantly reduced by the use of BioPortal and UMLS APIs. Both APIs provide means to help us construct the list of terms that identify concepts in their ontologies. The dictionary is automatically (re)built by accessing these APIs.

In the first step of the workflow presented in section 2.1, our biomedical annotator service uses a selection of 47 biomedical ontologies that give a dictionary of 793,681 unique concepts and 2130700 terms. The current prototype uses *mgrep*, a concept recognizer developed by the National Center for Integrative Biomedical Informatics. Mgrep is a tool that implements a novel radix tree based data-structure that enables fast and efficient matching of text against a set of dictionary terms. Mgrep was parameterized to match all the possible concepts.[7] This tool recently reported a degree of accuracy >95% in recognizing disease names [10]. We have conducted [22] a comparative evaluation of this tool with the gold standard in the biomedical community, MetaMap [17]. Mgrep is more precise at recognizing concepts, is more scalable, and can be configured to use third party dictionaries. As a result, unlike MetaMap, mgrep is not tied to the UMLS structure. Note that the prototype is designed to allow us to plug-in other concept recognizers. In the second step of the workflow, the biomedical annotator uses only an is_a transitive closure component to expand the annotations created by mgrep. Both BioPortal and UMLS APIs provide a function to get is_a hierarchy relations (i.e., for a given concept, which are the parents concepts) allowing us to pre-compute a hierarchy table.[8]

---

[6] UMLS (http://umlsinfo.nlm.nih.gov/) is a reference in biomedicine for controlled vocabularies. In our prototype, UMLS is available as a local database.

[7] If a text contains the string "cutaneous melanoma," two annotations are generated: one with 'melanoma' one with 'cutaneous melanoma' because the dictionary contains the two terms.

[8] The is_a transitive closure component can also be parameterized to return detailed results (i.e., track of the concept from which an `ISA_CLOSURE` annotation has been derived).

The service is deployed as a SOAP web service as well as a RESTful web service. The current WSDL and OWL ontology of the biomedical annotator service are available online (http://www.bioontology.org/tools.html) and the prototype can be accessed programmatically. A user friendly web interface is also provided. Current response times are ~20-25 seconds for 500 words. Nevertheless, some technical improvements are being considered: (i) keeping the dictionary loaded into memory between service calls (mgrep constraint); (ii) loading the pre-computed hierarchy table into memory.

We evaluated our biomedical annotator for the purpose of annotating a wide range of public biomedical datasets. For example, we annotated a set of 1,050,000 PubMed citations (title, abstract and other metadata), creating 174,840,027 annotations (18% `DIRECT`, 82% `ISA_CLOSURE`). By considering only one unique annotation done in several contexts of the same citation (e.g., if an annotation was found both in the title and in the abstract, we count just one), we obtained an average of 160 annotating concepts by citation. Approximately 99% of our test set was annotated (with at least 1 concept), demonstrating the service's utility.

In the current prototype, `conceptIDs` and `ontologyIDs` are ad-hoc usable identifiers rather than URIs. In this implementation, we have to abstract on both BioPortal and UMLS APIs that do not have the same representation and constraints for a concept. For example, in UMLS a concept may exist in several terminologies while this is not allowed in BioPortal where ontology mappings are encouraged. In the future, BioPortal will integrate all the UMLS ontologies and provide URIs for concepts and ontologies.

4.2. Use cases enabled and potential impact of the biomedical annotator service

In a recent study, we described the prototype implementation of OBR, an *open biomedical resources* index. OBR index was constructed using the our annotator service workflow and is directly queriable in the NCBO BioPortal [2]. OBR's objective is to perform offline annotation of a large number of biomedical resources and to provide an up-to-date index of annotated resource elements. There are five biomedical resources in the current prototype. The OBR index keeps track of the structures of elements it has annotated i.e., from which part of the element (e.g., title, description) an annotation has been produced. The system allows a user to search for various biomedical data related to a specific ontology concept (in one place) greatly enhancing the value of the ontology repository. For example, searching for "melanoma" in BioPortal returns, among others the concept `DOID:1909` from the human disease ontology. A user can access the 13 ArrayExpress experiments, the 673 clinical trials, the 960 articles in PubMed, or the 10 Gene Expression Omnibus datasets related to that concept that have been annotated in the OBR index. The OBR and OBA together create a system that allows a user to make a request of the annotator and get back the biomedical data related to a piece of text. The system may be useful, for example, when a researcher is writing a paper. He can annotate his article abstract and then use the system to find PubMed citations related to it.

They are many use cases for the annotator service we propose. For example, the Stanford Tissue Microarray Database (STMD), annotates the tissue samples in the database with concepts from the National Cancer Institute Thesaurus [9]. STMD designers

are potential users of the OBA web service for the following reasons: (i) annotating with ontologies is not their domain of expertise, (ii) integrating an ad-hoc annotation tool into STMD processes is not a good software practice, nor it is a stable and scalable solution over time, (iii) STMD designers may be interested by annotating their samples with more than one terminology. As a second example, NCBO collaborators from Univ. of California, San Francisco have to create annotations for HIV/AIDS clinical trials in order to provide an ontology-driven web application for visualizing, understanding, and analyzing the trials. As a third example, within the context of the Ontology Development Information Extraction project, biomedical informaticians led by a group at the Univ. of Pittsburgh are developing a set of tools to assist researchers with 'extracting meaning and codifying medical documents.' These groups are potential users of the annotator service.

5 Discussion and related work

Semantic annotation is an important research topic [1]. Tools vary along with the types of documents to annotate (e.g., image [23]). For an overview and comparison of semantic annotation tools the reader may refer to the study of Uren et al. [4]. The authors compare a large number of tools and the two main annotation frameworks Annotea [24] and CREAM [25] with seven requirements. They show that these requirements are hardly addressed by current solutions. The requirements and how the OBA web service is compatible with them are as follows:

☑ *Standard formats. Using standard formats is preferred, wherever possible.* The OBA web service returns OWL data.

☑ *User centered/collaborative design. To provide knowledge workers with easy to use interfaces that simplify the annotation process and place it in the context of their everyday work.* The choice of a service-oriented architecture fulfills this requirement.

☑ *Ontology support (multiple ontologies and evolution). Annotation tools need to be able to support multiple ontologies. Systems will have to cope with changes made to ontologies over time.* Our biomedical annotator supports multiple ontology formats (OBO, OWL, UMLS) and provides a means for versioning the dictionary (as explained before). This characteristic supports ontology evolution.

☑ *Support of heterogeneous document formats. Dealing with multiple document formats is a prerequisite for integrating annotation into existing work practices.* The OBA web service deals with the most basic format available: free text. Most documents can be converted into free text (or have free text extracted from them). For the moment, it is up to the user to keep track of the part of the document from where an annotation has been derived, as we are doing in our OBR project. In the future, the OBA web service should be able to process directly structured documents.

The two following requirements do not apply to the OBA web service:

☑ *Document evolution (document and annotation consistency). What should happen to the annotations on a document when it is revised.*

☑ *Annotation storage. The semantic web model assumes that annotations will be stored separately from the original document.* As a service, the OBA stores neither annotations

nor annotated text. OBA users have to manage this information. Note that in our OBR index, we have implemented wrappers that automatically keep resource contents (i.e., documents) and annotations up-to-date. They are stored separately.

☑ *Automation.* This requirement intrinsically fulfilled by the OBA web service.

The OBR-OBA use case presented before is similar to the approach of the SemanticHacker project (http://semantichacker.com). SemanticHacker enables 'semantic discovery' by providing users with a 'semantic signature,' which is a weighted representation of the concepts contained in a piece of text. It is based on Wikipedia content and links back to the related Wikipedia articles. The OBA web service differs principally from SemanticHacker in that it uses biomedical ontologies.

In the biomedical domain, automatic annotation or indexing of biomedical resources is an important topic. A number of publicly available concept recognizers identify concepts from ontologies or terminologies in a piece of text. For examples, see IndexFinder [26], MetaMap [17], SAPHIRE [27], and mgrep [10]. MetaMap is the gold standard for evaluating these tools. The main challenges for these tools are openness to various terminologies or ontologies, accuracy, and scalability. Most of the time, they deal only with UMLS. Our choice of mgrep was motivated by its high accuracy and its facility to consume a simple text based dictionary. The fact that it is not tied to a particular ontology structure was beneficial [22].

More similarly, CONANN [28] is an online biomedical concept annotator. It takes a source phrase, identifies potential matching concepts and phrases in a domain-specific thesaurus (UMLS), uses an incremental filter to remove candidate phrases using a variation of inverse document frequency, and maps the source phrase to the best matching concepts. CONANN provides a speed advantage and a better recognition precision when compared to MetaMap. CONANN aims to identify the best possible matches, whereas mgrep in the OBA web service identifies the greatest number of concepts. Maximizing the number of annotations for the OBR index enables the extraction of new knowledge. For example, it enables data-driven ontology alignment by mapping annotations together e.g., if a large number of resource elements have been annotated with two concepts it may means that this two concepts are related [8]. One feature of CONANN not currently implemented in the OBA web service is the use of term frequency to order/filter results. Note that if CONANN becomes freely available, and open to third party dictionaries, we can also use it for the concept recognition step fairly easily.

Many research projects leverage ontology annotations such as biomedical data search engines. MedicoPort [7] uses UMLS semantics to expand user queries. Moskovitch et al. [6] use ontologies for annotation (concept based search) and demonstrate the importance of the context (context-sensitive search) when annotating structured documents. They use UMLS based is_a relations for their semantic expansion. HealthCyberMap [5] uses ontologies and semantic distances for visualizing biomedical resources information. Finally, Essie [28] shows that a judicious combination of exploiting document structure, phrase searching, and concept based query expansion is a useful approach for information retrieval. Essie also leverages context and frequency of occurrence. Most of these tools

are limited to UMLS. This limitation gives our biomedical annotator a significant advantage.

Khelif et al. [30] present a similar work to the OBR project. They have annotated the GeneRIF resource using GATE [31], a natural language processing framework, for concept recognition. They used UMLS and Galen as ontologies. GATE allows them to extract not only concepts from text, but also relations. Their system returns and stores annotations in RDF and uses a software based on conceptual graphs, called Corese [32], to do semantic expansion (is-a relation, external rules, reasoning, and views on the annotations). The OBR index contains more content (resources annotated and ontologies used) but it may be useful in the future to see how to reuse GATE to extract relations and Corese as a semantic expansion component.

In contrast with classic web service, *semantic web services* (SWS) [33] are semantically described with ontologies. They use ontologies to create a knowledge-level model of information describing and supporting what the service accepts, does and returns. They also enable automated understanding of their functionalities. More generally, the semantic layer helps service discovery, invocation, composition, or interoperation. It enables reasoning about services, planning compositions of services, and automating their use by software agents. For example, web service discovery should be based on the semantic match between a description of a service demand, and a description of a service offer both available in a semantically competent platform/registry [34]. The two main current frameworks for SWS development are proposed by the OWL-based Web Service Ontology, (OWLS) [35] and Web Service Modeling Ontology (WSMO) [36] working groups. The semantic description of 'what is returned by the service' is the first step to deploying the OBA web service as SWS. It does not yet provide a semantic description of what the service does using a framework such as OWL-S or WSMO.

6 Conclusion and future work

Annotations play a crucial role in the emergence of the semantic web. The need to switch from the current web to semantically rich content annotated using ontologies has been clearly identified. Meeting this need requires services (usable by humans and agents) that can be integrated into web processes. We have presented a service for semantic annotation. Our service methodology leverages ontologies to create annotations of raw text and returns them using semantic web standards. We have also described an implementation of this model for the biomedical domain. Our biomedical annotator is distinguished from previous work in the biomedical domain because (i) it is clearly positioned as a service-oriented tool; and (ii) it has access to a large dictionary, which is composed of UMLS and NCBO ontologies. Evaluation has demonstrated accuracy (>95% for the concept recognizer) and utility (160 annotating concepts for one PubMed citation). The service workflow is currently involved in a project within NCBO to annotate a large number of open biomedical resources. The comparison, done in section 5, with the requirements of [4] shows that our semantic service-oriented approach makes sense and distinguishes our tool from what exists or is currently proposed in the community.

Future work will concentrate on three main issues that will determine the OBA web service success in processing annotations:

- *Concept recognition*. The choice of the mgrep tool is not fixed. Better concept recognizers can easily be plugged-in. It would be interesting to recognize more than concept from text, but also relations [37,38].
- *Ontologies.* We wish to allow users to choose (and eventually propose) any ontology desired for the annotation process (i.e., customize the dictionary on demand). We will also propose a set of 'topic oriented' dictionaries which will simplify the use of the service and will abstract on the ontology layer.
- *Semantic expansion components.* We are currently working on a semantic distance component [39] and a component that will extract concept similarity based on the current version of the OBR index (data-driven ontology alignment). Plus, in the future, we want to develop and re-use other components and still allow the user to select and parameter the one(s) to use when querying the service.

The OBA web service also has role as a testbed for evaluating semantic web services for NCBO. The evaluation of the OBA web service semantic approach (i.e., the couple ontology/instances) may enable more semantics for NCBO services as it may be a solution to add a semantic layer to the current NCBO SOAP and RESTful services.

Acknowledgements

This work is supported by the National Center for Biomedical Computing (NCBC)/ National Institute of Health roadmap initiative; NIH grant U54 HG004028. We also acknowledge the assistance of Manhong Dai and Fan Meng at University of Michigan.

References

1. Handschuh, S., Staab, S., eds.: Annotation for the Semantic Web. Vol. 96 of Frontiers in Artificial Intelligence and Applications. IOS Press (2003)
2. Jonquet, C., Musen, M.A., Shah, N.H.: A System for Ontology-Based Annotation of Biomedical Data. In: International Workshop on Data Integration in the Life Sciences. Vol. 5109 of LNBI, Evry, France, Springer (Jun 2008) 144–152
3. Rubin, D.L., Lewis, S.E., Mungall, C.J., Misra, S., Westerfield, M., Ashburner, M., Sim, I., Hute, C.G., Solbrig, H., Storey, M.A., Smith, B., Day-Richter, J., Noy, N.F., Musen, M.A.: National Center for Biomedical Ontology: Advancing Biomedicine through Structured Organization of Scientific Knowledge. OMICS A Journal of Integrative Biology 10(2) (Jun 2006) 185–198
4. Uren, V., Cimiano, P., Iria, J., Handschuh, S., Vargas-Vera, M., Motta, E., Ciravegna, F.: 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
5. Kamel-Boulos, M.N.: A first look at HealthCyberMap medical semantic subject search engine. Technology and Health Care 12 (2004) 33–41
6. Moskovitch, R., Martins, S.B., Behiri, E., Weiss, A., Shahar, Y.: A Comparative Evaluation of Full-text, Concept-based, and Context-sensitive Search. American Medical Informatics Association 14(2) (Mar-Apr 2007) 164–174

7. Can, A.B., Baykal, N.: MedicoPort: A medical search engine for all. Computer Methods and Programs in Biomedicine 86(1) (Apr 2007) 73–86
8. Shah, N.H., Rubin, D.L., Supekar, K.S., Musen, M.A.: Ontology-based Annotation and Query of Tissue Microarray Data. In: American Medical Informatics Association Annual Symposium, Washington DC., USA (Nov 2006) 709–713
9. Marinelli, R.J., Montgomery, K., Liu, C.L., Shah, N.H., Prapong, W., Nitzberg, M., Zachariah, Z.K., Sherlock, G.J., Natkunam, Y., West, R.B., van de Rijn, M., Brown, P.O., Ball, C.A.: The Stanford Tissue Microarray Database. Nucleic Acids Research 36 (2008) 871–877
10. Xuan, W., Dai, M., Mirel, B., Athey, B., Watson, S.J., Meng, F.: Interactive Medline Search Engine Utilizing Biomedical Concepts and Data Integration. In: BioLINK SIG: Linking Literature, Information and Knowledge for Biology, Vienna, Austria (Jul 2007) 55–58
11. Singh, M.P., Huhns, M.N.: Service-Oriented Computing, Semantics, Processes, Agents. John Wiley & Sons (2005)
12. Lee, W.J., Raschid, L., Srinivasan, P., Shah, N.H., Rubin, D., Noy, N.: Using Annotations from Controlled Vocabularies to Find Meaningful Associations. In Cohen-Boulakia, S., Tannen, V., eds.: 4th International Workshop Data Integration in the Life Sciences. Vol. 4544 of LNCS, Philadelphia, PA, USA, Springer (Jun 2007) 264–279
13. Caviedesa, J.E., Cimino, J.J.: Towards the development of a conceptual distance metric for the UMLS. Biomedical Informatics 37(2) (Apr 2004) 77–85
14. Euzenat, J., Shvaiko, P.: Ontology matching. Springer, Berlin Heidelberg, DE (2007)
15. O'Connor, M., Knublauch, H., Tu, S.W., Grosof, B.N., Dean, M., Grosso, W.E., Musen, M.A.: Supporting Rule System Interoperability on the Semantic Web with SWRL. In: 4th International Semantic Web Conference. Vol. 3729 of LNCS, Galway, Ireland, Springer (Nov 2005) 974–986
16. Wache, H., Vogele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H., H¨S.: Ontology-Based Integration of Information -A Survey of Existing Approaches. In: Workshop on Ontologies and Information Sharing, Seattle, WA, USA (Aug 2001) 108–117
17. Aronson, A.R.: Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: American Medical Informatics Association Annual Symposium, Washington, DC, USA (Nov 2001) 17–21
18. Bodenreider, O., Stevens, R.: Bio-ontologies: Current Trends and Future Directions. Briefings in Bioinformatics 7(3) (Aug 2006) 256–274
19. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., Consortium, T.O., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.A., Scheuermann, R.H., Shah, N.H., Whetzel, P.L., Lewis, S.: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology 25(11) (Nov 2007) 1251–1255
20. Bao, J., Hu, Z., Caragea, D., Reecy, J., Honavar, V.G.: A Tool for Collaborative Construction of Large Biological Ontologies. In: 4th InternationalWorkshop on Biological Data Management, Krakov, Poland, IEEE Press (Sep 2006) 191–195
21. Spasic, I., Ananiadou, S., McNaught, J., Kumar, A.: Text mining and ontologies in biomedicine: making sense of raw text. Brief Bioinformatics 6(3) (2005) 239–251
22. Bhatia, N., Shah, N.H., Rubin, D., Chiang, A.P., Musen, M.A.: Comparing Concept Recognizers for Ontology-Based Indexing: MGREP vs. MetaMap. In: Submission to American Medical Informatics Association Annual Symposium, Washington DC, USA (Nov 2008)
23. Hollink, L., Schreiber, G., Wielemaker, J., Wielinga, B.: Semantic Annotation of Image Collections. In: Knowledge Markup and Semantic Annotation Work., Sanibel, FL (Oct 2003)

24. Kahan, J., Koivunen, M.R., Prud'Hommeaux, E., Swick, R.R.: Annotea: an open RDF infrastructure for shared Web annotations. In: 10th international World Wide Web conference, Hong Kong (May 2001) 623–632
25. Handschuh, S., Staab, S., Studer, R.: Leveraging Metadata Creation for the Semantic Web with CREAM. In: 26th Annual German Conference on AI. Vol. 2821 of LNCS., Hamburg, Germany', Springer (2003)
26. Zou, Q., Chu, W.W., Morioka, C., Leazer, G.H., Kangarloo, H.: IndexFinder: A Method of Extracting Key Concepts from Clinical Texts for Indexing. In: American Medical Informatics Association Annual Symposium, Washington DC, USA (Nov 2003) 763–767
27. Hersh, W.R., Greenes, R.A.: SAPHIRE -an information retrieval system featuring concept matching, automatic indexing, probabilistic retrieval, and hierarchical relationships. Computers and Biomedical Research 23(5) (Oct 1990) 410–425
28. Reeve, L.H., Han, H.: CONANN: An Online Biomedical Concept Annotator. In Cohen-Boulakia, S., Tannen, V., eds.: 4th International Workshop Data Integration in the Life Sciences. Vol. 4544 of LNCS., Philadelphia, PA, USA, Springer (Jun 2007) 264–279
29. Ide, N.C., Loane, R.F., Demner-Fushman, D.: Essie: A Concept-based Search Engine for Structured Biomedical Text. American Medical Informatics Association 14(3) (2007) 253–263
30. Khelif, K., Dieng-Kuntz, R., Barbry, P.: An ontology-based approach to support text mining and information retrieval in the biological domain. Universal Computer Science, Special Issue on Ontologies and their Applications 13(12) (2007) 1881–1907
31. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: GATE: A framework and graphical development environment for robust NLP tools and applications. In: 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA (Jul 2002)
32. Corby, O., Dieng-Kuntz, R., Faron-Zucker, C., Gandon, F.: Searching the Semantic Web: Approximate Query Processing Based on Ontologies. IEEE Intelligent Systems 21(1) (Jan-Feb 2006) 20–27
33. Cabral, L., Domingue, J., Motta, E., Paynec, T., Hakimpour, F.: Approaches to Semantic Web services: an overview and comparisons. In: 1st European Semantic Web Symposium, Heraklion, Crete, Greece (May 2004)
34. Crubezy, M., Lu,W. ,Motta,E., Musen, M.A.:Configuring online problem-solving resources with the Internet Reasoning Service. Intelligent Systems 18(2) (Mar-Apr 2003) 34–42
35. Martin, D.L., Paolucci, M., McIlraith, S.A., Burstein, M.H., McDermott, D.V., McGuinness, D.L., Parsia, B., Payne, T.R., Sabou, M., Solanki, M., Srinivasan, N., Sycara, K.P.: Bringing semantics to Web services: The OWL-S approach. In: 1st International Workshop on Semantic Web Services and Web Process Composition. Vol. 3387 of LNCS, San Diego, CA, USA, Springer (Jul 2004) 26–42
36. Roman, D., Keller, U., Lausen, H., de Bruijn, J., Lara, R., Stollberg, M., Polleres, A., Feier, C., Bussler, C., Fensel, D.: Web Service Modeling Ontology. Applied Ontology 1(1) (2005) 77–106
37. Bundschus, M., Dejori, M., Stetter, M., Tresp, V., Kriegel, H.P.: Extraction of semantic biomedical relations from text using conditional random fields. BMC Bioinformatics 9 (Apr 2008) 207
38. Kim, J.D., Ohta, T., Tsujii1, J.: Corpus annotation for mining biomedical events from literature. BMC Bioinformatics 9 (Jan 2008) 10
39. Lee, W.N., andKaranjot Sundlass, N.H.S., Musen, M.A.: Comparison of Ontology-based Semantic-Similarity Measures. In: Submission to American Medical Informatics Association Annual Symposium, Washington DC, USA (Nov 2008)774