Simple, ontology-based representation of biomedical statements through fine-granular entity tagging and new web standards

Matthias Samwald [1,2] and Holger Stenzhorn [1,3]

[1] Digital Enterprise Research Institute (DERI), Galway, Ireland
[2] Konrad Lorenz Institute for Evolution and Cognition Research, Altenberg, Austria
[3] Institute of Medical Biometry and Medical Informatics (IMBI) - University Medical Center Freiburg, Germany

ABSTRACT

The number of web applications that enable end-users (such as biomedical researchers) to employ ontologies and Semantic Web technologies for formulating, publishing and finding information on the web in a practical manner is surprisingly small. In this paper we present a prototype of the *aTag* system which aims to drastically lower the entry barriers to the biomedical Semantic Web.

aTags ('associative tags') are short snippets of XHTML+RDFa with embedded RDF/OWL based on the SIOC vocabulary and domain ontologies and taxonomies (OBO ontologies and DBpedia). The structure of the embedded RDF/OWL is decidedly simple: a very short piece of human-readable text that is 'tagged' with relevant ontological entities. An *aTag generator* can be easily added to any web browser and allows researchers to quickly generate aTags out of key statements from web pages, such as PubMed abstracts. The resulting aTags can be embedded anywhere on the web, for example on blogs, wikis, or biomedical databases. We demonstrate how the resulting statements that are distributed over the web can be searched, visualized and aggregated with Semantic Web / Linked Data tools, and discuss how aTags can be used to answer practically relevant biomedical questions even though their structure is very simple.

The aTag project is carried out in cooperation with the BioRDF task force of the Semantic Web for Health Care and Life Science Interest Group of the World Wide Web Consortium (W3C).

1 INTRODUCTION

Significant progress has been made in the field of biomedical ontologies and Semantic Web technologies in recent years. The OBO Foundry (Smith et al. 2007) and the NCBO BioPortal (Musen et al. 2008) make a vast array of biomedical ontologies available to the public. Large biomedical knowledge bases have been created using RDF/OWL[1], such as the Neurocommons Knowledge Base (Ruttenberg et al. 2009), which currently contains over 400 million statements. The Linked Data community[2] created a global network of RDF/OWL data that consists of billions of statement. However, there is still a widely recognized lack of applications that empower end-users to access, add and link to this structured, ontology-based information on the web. While many current biomedical RDF/OWL resources offer a wealth of valuable information, their structures are often so complex and heterogeneous that it is hard for developers and users to make practical use of them.

aTags ('associative tags') offer a simple means of capturing biomedical statements in RDF/OWL format and publishing them anywhere on the web. aTags are short snippets of XHTML+RDFa with embedded RDF/OWL, and are based on the SIOC vocabulary and domain ontologies / taxonomies (OBO ontologies[3] and DBpedia). Below we will first describe how end-users can interact with aTags and surrounding tools, followed by a description of the underlying technologies and ontologies, and further discussion. A prototype of the software described here is accessible at *http://hcls.deri.org/atag*

2 END-USER EXPERIENCE

Note: High-resolution versions of the screenshots in this paper are available at *http://hcls.deri.org/atag/screenshots/*

2.1 Creating aTags

A user can simply start creating aTags by adding the *aTag bookmarklet* to his or her web browser (the bookmarklet is a bookmark with embedded Javascript that calls the aTag generator). Then, a user can navigate to any web page (such as an abstract of an article on PubMed), highlight a relevant statement in the text, and click on the bookmarklet. The aTag generator captures the highlighted statement and allows the user to add and refine semantic 'tags' (entities from OBO ontologies and DBpedia), as exemplified in Fig. 1. These tags capture the key entities mentioned in the statement in a machine-readable format, interlinking it with existing ontologies and Linked Data resources.

[1] http://www.w3.org/TR/owl-features/

[2] http://linkeddata.org/

[3] http://obofoundry.org

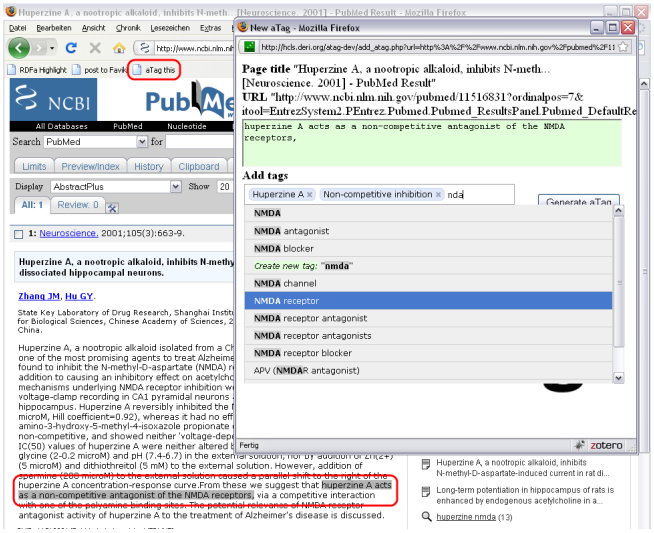

Fig. 1. Creating an aTag out of a PubMed abstract. The simple statement 'Huperzine A acts as a non-competitive antagonist of the NMDA receptor' is captured by tagging with the DBpedia entity 'Huperzine A', The OBO Gene Ontology entity 'receptor antagonist activity' and the DBpedia entity 'NMDA receptor'.

2.2 Publishing aTags

Since aTags are small fragments of HTML, they can be embedded into most current web resources and applications. Fig 2. shows some aTags that were published on a blog. In the current prototype, all aTags created with the aTag generator are also added to a central HTML document, the 'aTag pastebin'. Since the RDFa embedded within the HTML is also carried along when parts of web documents are copied and pasted to other locations, it is possible to re-arrange and merge datasets with the ease of re-arranging free text.

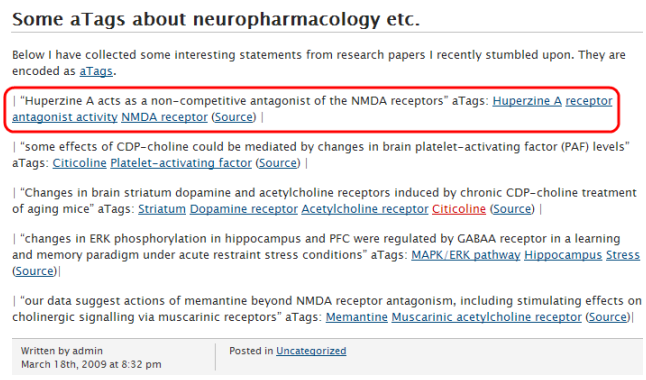

Fig. 2. aTags embedded on a personal blog

2.3 Finding and Querying aTags

Documents containing aTags can be found with Semantic Web search engines such as Sindice[4]. Collections of aTags that are located on a single webpage can be easily made accessible for generic faceted browsing with SIMILE Exhibit[5]. Since aTags distributed over the web use shared identifiers, aTags on different locations are interlinked with each other, was well as expressive domain ontologies and Linked Data resources.. This makes it possible to search, explore and query aTags in a very sophisticated manner. This is exemplified by Visinav[6] (Fig. 3), which crawls interlinked RDF/OWL on the web and allows faceted searching over the aggregated data from these distributed resources.

Fig. 3. Faceted browsing of aggregated data from aTags, DBpedia and OBO ontologies on the web, crawled and visualized by Visinav.

3 UNDERLYING TECHNOLOGIES

Currently, the entity tags suggested by the user come from a subset of the Open Biomedical Ontologies (OBO) and DBpedia, a vast repository of highly interlinked RDF/OWL derived from Wikipedia. aTags are short fragments of XHTML, where RDF statements are embedded according to the RDFa standard. Parts of the popular *Semantically Interlinked Online Communities* (SIOC) vocabulary[7] are re-used to represent the association between content and the entities from domain ontologies/taxonomies.

The aTags generator is based on an Ajax interface and server-side PHP code. Apache Solr[8] (which is based on Apache Lucene) running on an Apache Tomcat server is used for tag autocompletion of entity tags. Solr/Lucene allows for optimizing the relevance of suggested tags by

[4] http://sindice.com
[5] http://www.simile-widgets.org/exhibit/
[6] http://visinav.deri.org/atags/
[7] http://sioc-project.org
[8] http://lucene.apache.org/solr/

```
<http://hcls.deri.org/atag-data/pastebin.html#49ddfee65f7f4> a sioc:Item ;
        sioc:content "Huperzine A acts as a non-competitive antagonist of the NMDA receptors"@en ;
        sioc:topic  <http://dbpedia.org/resource/Huperzine_A> ,
                    <http://purl.org/obo/owl/GO#GO_0048019> ,
                    <http://dbpedia.org/resource/NMDA_receptor> ;
        rdfs:seeAlso <http://www.ncbi.nlm.nih.gov/pubmed/11516831> .
        sioc:ip_address "131.130.109.218"@en ;
```

Fig. 4. RDF/OWL statements embedded in an aTag HTML fragment, shown in *Turtle* syntax. URIs from OBO ontologies and DBpedia are used as 'tags'.

elaborate ranking based on contextual cues and ontological structures.

4 DISCUSSION

The design philosophy of the aTag system is based on the hypothesis that semi-structured data can be sufficient to tackle many realistic biomedical use-cases., and that simplistic data structures are easier to integrate, understand and use than more complex data structures. This hypothesis is motivated by experiences we made with creating, integrating and using large-scale ontology-based information repositories in recent years.

In the case of aTags, a set of ontological entities is used to describe biomedical statements without capturing the actual relationships between those entities. Preliminary experiments with aTags show that this incomplete information can still yield very useful results in realistic use cases, but further research is needed to determine the potential as well as the limitations of this approach.

4.1 Future directions

We are currently pursuing several threads of development, among them are:

- Adding a 'tag recommender' function to the aTag generator by integrating the Open Biomedical Annotator web service[9] (C Jonquet et al. 2009) of NCBO.
- Converting data from existing structured, biomedical databases (such as relationa databases) into aTags.
- Exposing automated text annotations from sentence-based NLP services such as Whatizit and iHop (Hoffmann und Valencia 2005) as aTags
- Integrating aTag widgets into popular web applications such as Wordpress and Drupal. This if facilitated by existing extensions of these applications for SIOC.
- Exploring the alignment of data represent with aTags to data represented with *SWAN* (Ciccarese et al. 2008), an ontology for the representation of scientific statements and discourse structures in RDF/OWL. SWAN has recently been aligned to SIOC as part of the W3C HCLS interest group work, and allows the description of agreement, disagreement etc. between scientific statements.
- Creating a mapping between OBO ontologies and DBpedia.
- Creating a specialized web crawler and graphical user interface for aTags that makes use of background information from ontologies that are used for tagging, e.g., for query expansion based on subsumption reasoning.

4.2 Related work

Faviki[10] is a social bookmarking application that allows users to tag web pages with entities from DBpedia. It is geared towards classical tagging of entire documents, rather than selected statements. *Loomp*[11] is centered around an extension of a popular web-based WYSIWYG editor and allows the addition of fine-granular semantic annotations during content authoring It is currently focused on the media industry and journalism.

ACKNOWLEDGEMENTS

Thanks to Andreas Harth (DERI Galway) for setting up the Visinav system. Thanks to Kei Cheung (Yale University) for organising the BioRDF task force of the HCLS IG. The work presented in this paper has been funded in part by a postdoctoral fellowship from the Konrad Lorenz Institute for Evolution and Cognition Research, Austria and by the Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion-2).

REFERENCES

Ciccarese,P. et al. (2008) The SWAN biomedical discourse ontology. *J Biomed Inform*, 41, 739-51.

Hoffmann,R. und Valencia,A. (2005) Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics*, 21 Suppl 2, ii252-8.

Jonquet,C. et al. (2009) The Open Biomedical Annotator. San Francisco, California, USA.

[9] http://bioontology.org/wiki/index.php/Annotator_Web_service

[10] http://www.faviki.com/

[11] http://www.loomp.org/

Musen,M. et al. (2008) BioPortal: Ontologies and Data Resources with the Click of a Mouse. *AMIA Annu Symp Proc*, 1223-4.

Ruttenberg,A. et al. (2009) Life sciences on the Semantic Web: the Neurocommons and beyond. *Brief Bioinform*, bbp004.

Smith,B. et al. (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol*, 25, 1251-5.