Ciccarese et al. Journal of Biomedical Semantics 2012, 3(Suppl 1):S1
http://www.jbiomedsem.com/supplements/3/S1/S1

JOURNAL OF
BIOMEDICAL SEMANTICS

PROCEEDINGS Open Access

Open semantic annotation of scientific publications using DOMEO

Paolo Ciccarese[1], Marco Ocana[2], Tim Clark[1,2,3*]

From Bio-Ontologies 2011
Vienna, Austria. 15-16 July 2011

* Correspondence:
tim_clark@harvard.edu
[1]Harvard Medical School and Massachusetts General Hospital, Boston MA, USA

Abstract

Background: Our group has developed a useful shared software framework for performing, versioning, sharing and viewing Web annotations of a number of kinds, using an open representation model.

Methods: The Domeo Annotation Tool was developed in tandem with this open model, the Annotation Ontology (AO). Development of both the Annotation Framework and the open model was driven by requirements of several different types of alpha users, including bench scientists and biomedical curators from university research labs, online scientific communities, publishing and pharmaceutical companies.
Several use cases were incrementally implemented by the toolkit. These use cases in biomedical communications include personal note-taking, group document annotation, semantic tagging, claim-evidence-context extraction, reagent tagging, and curation of textmining results from entity extraction algorithms.

Results: We report on the Domeo user interface here. Domeo has been deployed in beta release as part of the NIH Neuroscience Information Framework (NIF, http://www.neuinfo.org) and is scheduled for production deployment in the NIF's next full release.
Future papers will describe other aspects of this work in detail, including Annotation Framework Services and components for integrating with external textmining services, such as the NCBO Annotator web service, and with other textmining applications using the Apache UIMA framework.

Background

Annotation is a fundamental activity in biomedical research and in scholarship generally. It associates a commentary or formal judgment (textual comment, revision, citation, classification, or other related object) to a target of annotation, such as a text or image. It can be created for personal use, as in note-taking and personal classification of documents. Or it can be addressed to an audience beyond its creator, as in shared commentary on documents, reviewing, citation, and tagging.

On the Web, a target or subject of annotation necessarily means a digital artifact – that is, a Web document, or more technically, an "information resource" [1]. The predicate or content of an annotation will be typically either a discourse about, or metadata about, the target. At times, an annotation can be as simple as a highlight. Multiple systems and technical approaches are used today for annotating information resources on the Web and for viewing the annotations in context.

Shared document annotation such as that available in Utopia (http://getutopia.com) [2] or Mendeley (http://www.mendeley.com) [3], are becoming increasingly popular in certain communities. There is also a growing interest in, and recognition of the importance of, using information extraction algorithms to perform semantic tagging of biomedical publications [4].

Biological textmining algorithm performance trials are organized annually by the Biocreative group (http://www.biocreative.org/) [5-11]. The National Center for Biomedical Ontology (NCBO) offers ontology-driven term extraction on biomedical text as a core service. Several academic groups are active developers of biomedical textmining applications.

Biocuration, or biomedical database resource annotation, is highly useful and prevalent in biomedical research, for example, the annotation of resources such as Wormbase [12] or Flybase [13-15] with terms from the Gene Ontology [16-18], or annotating a UniProtKB/Swiss-Prot entry to reflect newly discovered information about a new protein family in *A. thaliana* [19].

Every professional scientist does a very simple kind of annotation on a routine basis: citation of the bibliographic metadata of supporting evidence for the papers' assertions. Publication of results in any peer-reviewed publication is impossible without it.

In whatever context it occurs, annotation is a key element of the process of doing science, as it supports the "virtual witnessing" process, characterized by Shapin as being fundamental to the scientific method since the first scientific journals and books began to be published in the 17th century [20].

Web-based annotation is done today in multiple ways using various technical approaches and annotation representation formats. This is understandable considering the current transitional state of scientific publishing between print media and the Web. Furthermore, design decisions made in the early architecture of the Web itself (links embedded in the page) made development of unified tools and representations of Web annotation quite difficult before tooling and infrastructure of the semantic web became available.

Now is a good time to resolve this situation.

Our group has developed a useful software framework for performing, versioning, sharing and viewing Web annotations of a number of kinds, using an open representation model, the Annotation Ontology (AO) [21,22]. This framework – and in particular the Domeo user interface component – were developed in tandem with AO. A number of different alpha users provided concrete and practical use cases, which were incrementally added to the toolkit. We report on the Domeo user interface here. Future papers will describe other aspects of this work in detail, including the integration with the Apache UIMA framework through the open source Apache Clerezza-AO plugin, and the Annotation Framework Services.

Results and discussion

The Domeo Annotation Toolkit and the Annotation Ontology

The Domeo Annotation Toolkit (http://annotationframework.org) is a collection of software components that provides a rich set of features including

(i) semantic annotation of online HTML and XML documents;

(ii) automated, semi-automated and fully manual annotation protocols;

(iii) structured, semi-structured and unstructured annotation types;

(iv) full provenance of annotation and curation records;

(v) selective sharing of annotations;

(vi) serialization of annotations in RDF/XML using the Annotation Ontology; and

(vii) enhanced searching of annotations by leveraging semantic inference.

Domeo is currently in beta release.

As of this writing we support installations at the Massachusetts General Hospital; and at the University of California at San Diego (UCSD), as part of the NIH Neuroscience Information Framework (NIF, http://www.neuinfo.org/) and the NIH Blueprint for Neuroscience Research (http://neuroscienceblueprint.nih.gov/).

A full production version of Domeo will be included in the next release of the Neuroscience Information Framework.

Using Domeo, registered users can create unstructured, semi-structured and fully structured or semantic annotation on Web documents using this framework. It does not matter whether or not the documents themselves are under update control of the annotator. Annotation can be kept private, shared within selected groups, or made public and therefore available to the entire Web. These access control features enable personal as well as collaborative use of the tool.

The Domeo Toolkit was developed in parallel with the Annotation Ontology (AO), an OWL ontology providing a model for creating 'stand-off' annotation anchored to online resources such as documents, images and databases and their fragments [21,22]. AO provides a robust set of methods for linking online resources, for example text in scientific publications, to ontological elements, with full representation of the annotation provenance.

Through AO, existing domain ontologies and vocabularies – in OWL [23] or SKOS [24] - can be utilized, out of the box, for creating extremely rich stores of metadata on web resources. In the bio-medical field, subjects for ontological structuring include biological processes, molecular functions, anatomical and cellular structures, tissue and cell types, chemical compounds, and biological entities such as genes and proteins. However, AO is not limited to the bio-medical domain and can be easily used in other scientific and non-scientific contexts. In fact, AO is already used by other projects focusing on biodiversity [25] and social tagging [26,27].

AO, by linking new scientific content to computationally defined terms and entity descriptors, can help establish semantic interoperability across the diverse masses of specialist science embodied in digital media – from journals, to wikis and blogs [28,29], to the growing world of web-based research "collaboratories" [30]. In biomedicine, semantic interoperability facilitates cross-species comparisons, pathway analysis, disease modeling, compact "mashups" for visualization, and the generation of new hypotheses through data integration and machine reasoning. Annotation can enrich the information content of web documents as well as contextualizing discussion about them. When annotation metadata take the form of controlled biomedical term sets, it can be used by software agents to enhance "strategic reading" [31,32].

While AO provides the model for encoding and sharing annotation in the convenient RDF (Resource Description Framework) format [33], it is still necessary to develop software applications allowing the users, in our case bio-medical scientists, to manually or semi-automatically create, share/publish, search and utilize annotation, and to manage algorithmically created annotation. As we strongly believe developing actual software is required to test the exchange model format against real use cases, we developed AO in parallel with AF.

In the following sections we describe some of the features of the current version of AF's user interface web component, the Domeo annotation tool.

The Domeo user interface

The Domeo user interface is an extensible web component enabling direct user-invoked annotation of online HTML/XHTML/XML documents. It was developed using a combination of Google Web Toolkit (GWT) and pure JavaScript. As the GWT code is eventually compiled into JavaScript, the final product consists of a JavaScript file to be imported by the hosting page. In order to work properly, the tool requires a series of services provided by the Domeo server component.

Domeo was designed to be part of the normal everyday workflow of scientists. It enables loading of online HTML documents that display as they would display when loaded directly in the browser. The task can be performed by copy and paste of the document URL in the Domeo address bar or, in a more efficient way, by using a Firefox plugin that adds the Domeo icon to the browser statusbar. With the plugin, when a user wishes to annotate the current web page, invoking the plugin via a single click, triggers re-loading of the page within the Domeo user interface.

The Domeo server

A Domeo server can potentially be developed in any language or platform able to publish a web page accommodating the JavaScript of the Domeo web application. Our current Domeo services implementation runs currently relies on a Grails installation [34] and on a Java and Groovy [35] code base. MySQL (http://www.mysql.com/) supports the annotation repository. Communication between the server and the Javascript client is via JSON.

Domeo web services

Domeo is designed to access [36], ontologies and other automated markup facilities via web service calls. We currently access vocabulary lookup, selection and entity recognition services through the NCBO Bioportal and Annotator web services hosted at Stanford University. These services are called when annotation vocabularies are selected, and when textmining is specified for a document or document section. Clearly automatic term recognition in biology is an evolving and highly competitive area, as can be seen by the fact that international competitions between various algorithms are regularly organized – see for example [37-41]. Our web services strategy is designed to enable Domeo to track and adapt to these changes in technology over time, and to differences amongst algorithms in fitness for particular purposes.

We also provide bibliographic reference lookup and identification through PubMed web services hosted at the National Center for Biomedical Information (NCBI). When a web document is loaded into Domeo, the software attempts to recognize its source. If the source is known and the source structure is available to Domeo, all the references are parsed to find PubMed and/or PMC identifiers, which are used to instantiate the complete bibliographic metadata for the article itself and its cited references as internal object references. This citation network is by default, shared as a "public" annotation in Domeo. If a bibliographic record already exists as a Domeo object reference, it will not be re-extracted. Over time, a Domeo installation will build up a large network of citations in the areas of interest being curated.

As the citation network extraction is run server-side, it is possible to pre-fetch substantial citation networks as required, although this is not a standard feature in our current beta deployment.

Manual annotation with the Domeo user interface

Once the resource is loaded, the user may manually annotate the whole document, or sections of it, by selecting the desired portion of text and attaching a "topic", representing an instance of one of several available annotation types. The simplest annotation types that can be created through Domeo are:

- highlight: A **highlight** is the process of marking a fragment of a resource with some - usually visual - mechanism.
- note: A **note** is defined as a brief record of facts, topics, or thoughts, written down as an aid to memory. When using the term 'note' we here refer only to free text notes.
- semantic tag: A **semantic tag** or, in AO terms, a **qualifier,** expresses a relationship between the target of the annotation and a well-defined semantic or structured entity consisting of a term of a controlled vocabulary or ontology identified by a URI.

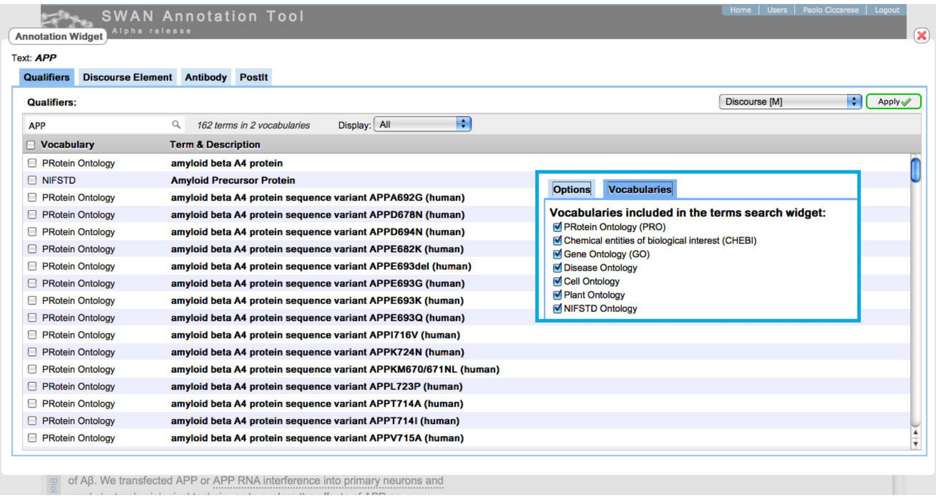
The annotation process is enabled by a user interface widget that allows choosing and detailing the desired annotation type. In the case of qualifiers, the widget allows automatic term recognition (entity identification) on selected text, via an external web service accessed through a software connector. The current beta version connects to the NCBO (National Center for Biomedical Ontology) Annotator REST web service for terms search [42-44]. The NCBO search capability allows specifying some text and finding terms across multiple ontologies that contain it. The names, synonyms, and properties for a term are searched for matches to the entered text.

It is important to note that Domeo allows users to connect any search service and to customize the list of available vocabularies. By simply changing the set of vocabularies used for performing the annotation, it is possible to tackle any desired domain. It is even possible to re-use the BioPortal infrastructure for uploading and managing the desired ontologies and providing some of the web services necessary to Domeo to operate.

Domeo currently presents the search results in a linear list (Figure 1). However, motivated by users' feedback, we are exploring alternative visualization that would allow users to browse the desired terminology or ontology.

Once the annotation – in this case a qualifier – is created, the annotated span of text of the document is highlighted. Clicking on the span of text allows inspection of the annotation items associated with it, through a popup (Figure 2). Full annotation provenance - who created the annotation, when, with what tool... - is recorded, however only a summary is displayed by the popup.

The above annotation types are part of the standard distribution of Domeo. Additional types can be added by developing new software components or plugins. These



Text: *APP*

Qualifiers:

APP — *162 terms in 2 vocabularies* — Display: All

□ Vocabulary	Term & Description
□ PRotein Ontology	amyloid beta A4 protein
□ NIFSTD	Amyloid Precursor Protein
□ PRotein Ontology	amyloid beta A4 protein sequence variant APPA692G (human)
□ PRotein Ontology	amyloid beta A4 protein sequence variant APPD678N (human)
□ PRotein Ontology	amyloid beta A4 protein sequence variant APPD694N (human)
□ PRotein Ontology	amyloid beta A4 protein sequence variant APPE682K (human)
□ PRotein Ontology	amyloid beta A4 protein sequence variant APPE693del (human)
□ PRotein Ontology	amyloid beta A4 protein sequence variant APPE693G (human)
□ PRotein Ontology	amyloid beta A4 protein sequence variant APPE693K (human)
□ PRotein Ontology	amyloid beta A4 protein sequence variant APPE693Q (human)
□ PRotein Ontology	amyloid beta A4 protein sequence variant APPI716V (human)
□ PRotein Ontology	amyloid beta A4 protein sequence variant APPK724N (human)
□ PRotein Ontology	amyloid beta A4 protein sequence variant APPKM670/671NL (human)
□ PRotein Ontology	amyloid beta A4 protein sequence variant APPL723P (human)
□ PRotein Ontology	amyloid beta A4 protein sequence variant APPT714A (human)
□ PRotein Ontology	amyloid beta A4 protein sequence variant APPT714I (human)
□ PRotein Ontology	amyloid beta A4 protein sequence variant APPV715A (human)

Vocabularies included in the terms search widget:
- ☑ PRotein Ontology (PRO)
- ☑ Chemical entities of biological interest (CHEBI)
- ☑ Gene Ontology (GO)
- ☑ Disease Ontology
- ☑ Cell Ontology
- ☑ Plant Ontology
- ☑ NIFSTD Ontology

of Aβ. We transfected APP or APP RNA interference into primary neurons and

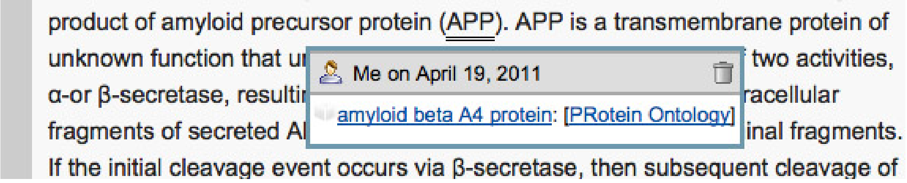
Figure 1 NCBO BioPortal webservice search results. Results of a terms search against the NCBO BioPortal through the BioPortal webservice. The results are currently displayed as a list and they include the terms belonging to the terminologies/ontologies enabled by the user's preferences.

define user interface components, semantic aspects of the new annotation topics, and connectors to external services when needed. Already developed plugins include features for modeling scientific discourse according to the model provided by the SWAN [45,46] ontology and features for modeling antibody usage. The latter has been developed in collaboration with the NIF (Neuroscience Information Framework) project and consists of annotating text with one of the antibody entries of http://antibodyregistry.org and, optionally, with the methods and species involved in the particular study reported in the document content.

The SWAN ontology plugin

The SWAN plugin allows key assertions or claims in any paper to be recorded, along with their primary evidence in the literature, comments by the reader / reviewer / curator, and mappings to biomedical terminologies. Terminology mappings can be generated automatically – with user override – for any assertion. Bibliographic references are treated as structured semantic annotation, and are normalized against the National Library of Medicine's article metadata using NCBI PubMed web services.

Annotation created by this plugin, as with all Domeo annotation, is independently retrievable, selectively shareable (for example with collaborators), and retains location-based contextualization in the original literature [47].



product of amyloid precursor protein (APP). APP is a transmembrane protein of
unknown function that u … two activities,
α-or β-secretase, resulti … racellular
fragments of secreted A … inal fragments.
If the initial cleavage event occurs via β-secretase, then subsequent cleavage of

Me on April 19, 2011

amyloid beta A4 protein: [PRotein Ontology]

Figure 2 Clicking on annotated text to inspect the associated annotation. Users click on the annotated text to inspect the associated annotation items, in this example a semantic entity represented by a term from the PRotein Ontology. Some of the available provenance data are also displayed.

This plugin was initially designed to allow users to create SWAN content through Domeo. However we believe it will also be useful in the broader context of general scientific note-taking.

The SWAN project (Semantic Web Applications in Neuromedicine) developed a practical, common, semantically-structured, framework for scientific discourse, which can be applied to significant problems in Alzheimer Disease (AD) research and many other biomedical disorders. In the initial workflow, curators of AlzSWAN, the SWAN Alzheimer Database [48], used the SWAN Workbench to curate content which was then published via the SWAN Browser web application (http://hypothesis.alzforum.org). Output of SWAN's curation process represented the article's scientific discourse by means of formal discourse elements from the SWAN ontology: questions, hypotheses, claims,and evidence. For each discourse element the curator selected related publications (motivating or evidence), associated proteins and genes, plus a few other properties. The curator also connected certain discourse element for individual publications, with others already in the knowledge base.

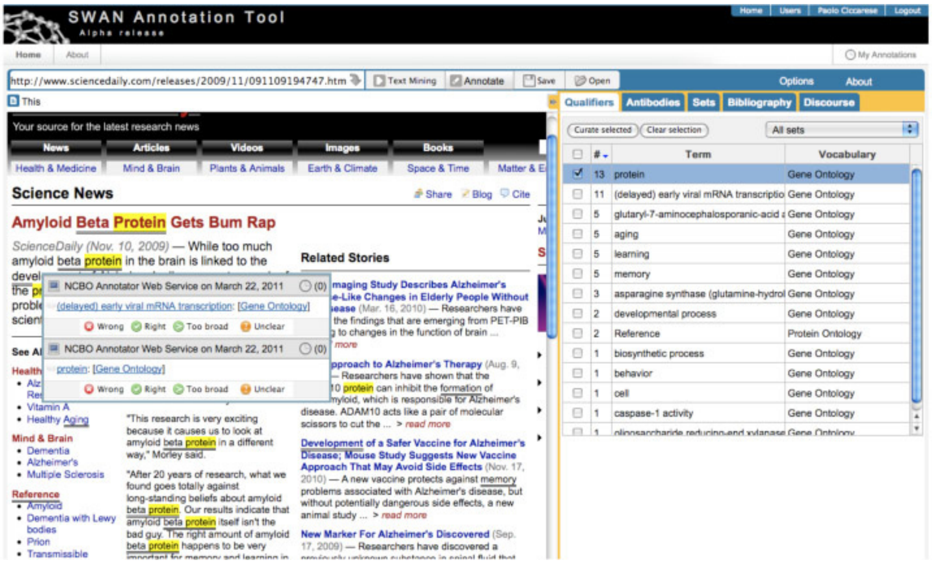
It eventually became clear that because the curation software was removing the annotated discourse elements from their context in the original publications, they were also extracted from the normal scientific activity of reading, which focuses on articles, not databases. We wanted to restore this connection.

This limitation can be overcome with Domeo (Figure 4) using the SWAN ontology plugin. This module allows highlighting a section of a document and annotating it with a SWAN discourse element: Claim, Hypothesis, Question or Definition. It is possible to attach terms relevant to the discourse elements through a search against the NCBO BioPortal web service. Domeo also allows specification of evidence for claims, as represented by cited publications. Directly citing datasets in an archive such as Dataverse [49,50] is a new feature currently in development.

Evidence, as modeled by the SWAN ontology, can be classified as supportive, inconsistent or relevant. In the current implementation, publications metadata are retrieved through a search in PubMed via NCBI web services. It is also possible to relate the new research statement with others already in the knowledge base once again through the relationships provided by the SWAN ontology: consistent, inconsistent and alternativeTo. The process is currently manual but we are planning to experiment with methods for retrieving automatically related statements in the knowledge base. Once the desired discourse elements have been created it is possible to see them summarized in the Domeo 'discourse' perspective (Figure 5).

Semi-automatic annotation

In many cases, the efficiency of mass-scale manual annotation can be significantly augmented by annotation algorithms. DOMEO allows implementing the RECS (Run, Encode, Curate, Share) process. Using this process, it is possible to select and run external text mining services, encode the results in the AO format, display the results in the context of the annotated document (Figure 3) to enable the curation process. In the current version of the tool we integrated the NCBO Annotator web service. The NCBO Annotator is an ontology-based Web service that annotates public datasets with biomedical ontology concepts based on their textual metadata. It is possible,



http://www.sciencedaily.com/releases/2009/11/091109194747.htm

Science News

Amyloid Beta Protein Gets Bum Rap

ScienceDaily (Nov. 10, 2009) — While too much amyloid beta protein in the brain is linked to the devel
the p
proble
scien

NCBO Annotator Web Service on March 22, 2011 ○ (0)
(delayed) early viral mRNA transcription: [Gene Ontology]
Wrong Right Too broad Unclear

NCBO Annotator Web Service on March 22, 2011 ○ (0)
protein: [Gene Ontology]
Wrong Right Too broad Unclear

See Al
Health
- Alz Res
- Vitamin A
- Healthy Aging

Mind & Brain
- Dementia
- Alzheimer's
- Multiple Sclerosis

Reference
- Amyloid
- Dementia with Lewy bodies
- Prion
- Transmissible

"This research is very exciting because it causes us to look at amyloid beta protein in a different way," Morley said.

"After 20 years of research, what we found goes totally against long-standing beliefs about amyloid beta protein. Our results indicate that amyloid beta protein itself isn't the bad guy. The right amount of amyloid beta protein happens to be very

Related Stories

maging Study Describes Alzheimer's e-Like Changes in Elderly People Without ease (Mar. 16, 2010) — Researchers have the findings that are emerging from PET-PIB to changes in the function of brain ... more

pproach to Alzheimer's Therapy (Aug. 9, — Researchers have shown that the 10 protein can inhibit the formation of myloid, which is responsible for Alzheimer's disease. ADAM10 acts like a pair of molecular scissors to cut the ... > read more

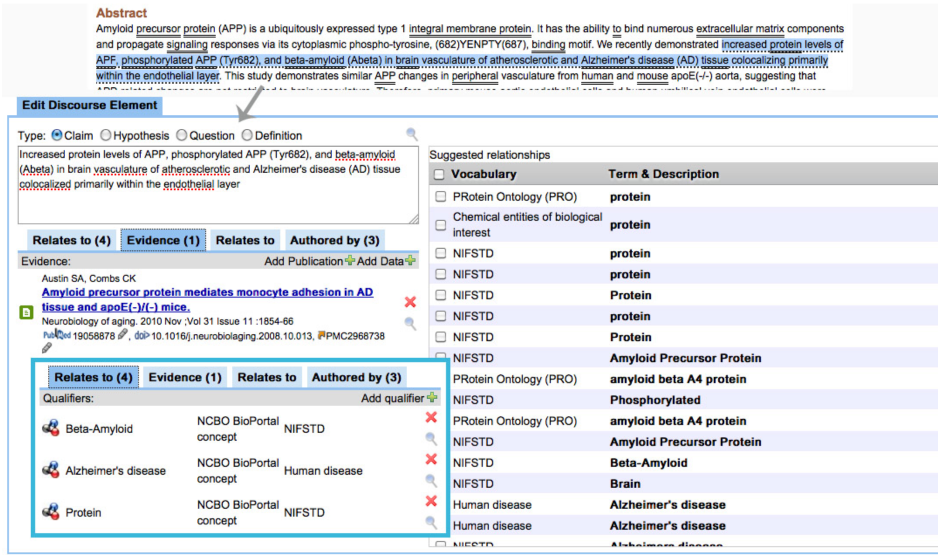
Development of a Safer Vaccine for Alzheimer's Disease; Mouse Study Suggests New Vaccine Approach That May Avoid Side Effects (Nov. 17, 2010) — A new vaccine protects against memory problems associated with Alzheimer's disease, but without potentially dangerous side effects, a new animal study ... > read more

New Marker For Alzheimer's Discovered (Sep. 17, 2009) — Researchers have discovered a

□	#	Term	Vocabulary
☑	13	protein	Gene Ontology
□	11	(delayed) early viral mRNA transcriptio	Gene Ontology
□	5	glutaryl-7-aminocephalosporanic-acid a	Gene Ontology
□	5	aging	Gene Ontology
□	5	learning	Gene Ontology
□	5	memory	Gene Ontology
□	3	asparagine synthase (glutamine-hydrol	Gene Ontology
□	2	developmental process	Gene Ontology
□	2	Reference	Protein Ontology
□	1	biosynthetic process	Gene Ontology
□	1	behavior	Gene Ontology
□	1	cell	Gene Ontology
□	1	caspase-1 activity	Gene Ontology
□	1	oligosaccharide reducing-end xylanase	Gene Ontology

Figure 3 Curation popup for user review and response to automatically generated annotation items. The text-mining results are displayed on the document and the curation popup lets the user review and respond to automatically generated annotation items.

through the Domeo preferences panel, to specify which ontologies to consider when running the service. The current list of allowed ontologies for running the NCBO Annotator is the same list of ontologies allowed in the terms search through the NCBO bioportal mentioned in the previous sections.



Abstract
Amyloid precursor protein (APP) is a ubiquitously expressed type 1 integral membrane protein. It has the ability to bind numerous extracellular matrix components and propagate signaling responses via its cytoplasmic phospho-tyrosine, (682)YENPTY(687), binding motif. We recently demonstrated increased protein levels of APP, phosphorylated APP (Tyr682), and beta-amyloid (Abeta) in brain vasculature of atherosclerotic and Alzheimer's disease (AD) tissue colocalizing primarily within the endothelial layer. This study demonstrates similar APP changes in peripheral vasculature from human and mouse apoE(-/-) aorta, suggesting that

Type: ● Claim ○ Hypothesis ○ Question ○ Definition

Increased protein levels of APP, phosphorylated APP (Tyr682), and beta-amyloid (Abeta) in brain vasculature of atherosclerotic and Alzheimer's disease (AD) tissue colocalized primarily within the endothelial layer

Evidence:
Austin SA, Combs CK
Amyloid precursor protein mediates monocyte adhesion in AD tissue and apoE(-)/(-) mice.
Neurobiology of aging. 2010 Nov ;Vol 31 Issue 11 :1854-66
PubMed 19058878, doi 10.1016/j.neurobiolaging.2008.10.013, PMC2968738

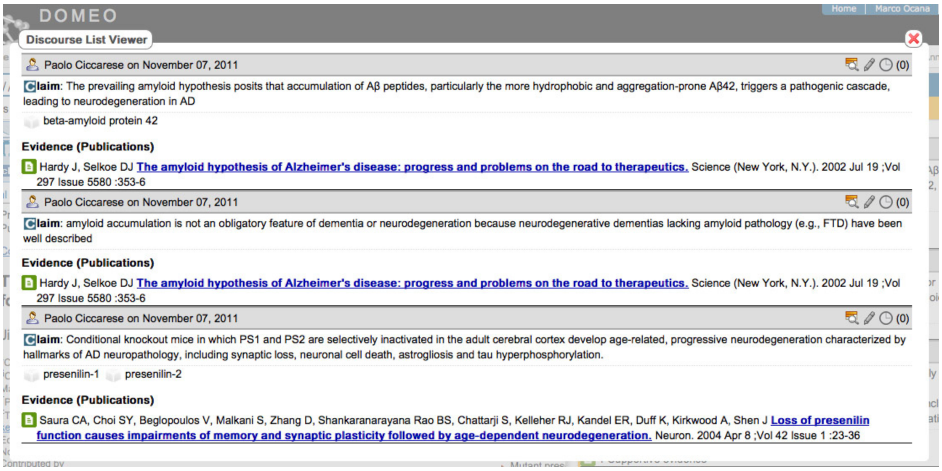
Qualifiers:

Beta-Amyloid	NCBO BioPortal concept	NIFSTD
Alzheimer's disease	NCBO BioPortal concept	Human disease
Protein	NCBO BioPortal concept	NIFSTD

Suggested relationships

□ Vocabulary	Term & Description
□ PRotein Ontology (PRO)	protein
□ Chemical entities of biological interest	protein
□ NIFSTD	protein
□ NIFSTD	protein
□ NIFSTD	Protein
□ NIFSTD	protein
□ NIFSTD	Protein
□ NIFSTD	Amyloid Precursor Protein
PRotein Ontology (PRO)	amyloid beta A4 protein
NIFSTD	Phosphorylated
PRotein Ontology (PRO)	amyloid beta A4 protein
NIFSTD	Amyloid Precursor Protein
NIFSTD	Beta-Amyloid
NIFSTD	Brain
Human disease	Alzheimer's disease
Human disease	Alzheimer's disease
□ NIFSTD	Alzheimers disease

Figure 4 Highlighted text fragment with associated discourse element creation panel. Above, the highlight of the text fragment from the PubMed abstract at the URL http://www.ncbi.nlm.nih.gov/pubmed/19923279. Below, the discourse element creation panel. The left side of the interface is dedicated to the type and description of the statement together with all the related terms, publications, other statements in the knowledge base and authors. In this particular example, it is possible to see a supportive publication and the list of related ontological terms. The right section of the interface is used for suggestions and for searching for terms, evidence - publication and data, and other statements in the knowledge base.



Paolo Ciccarese on November 07, 2011 (0)

Claim: The prevailing amyloid hypothesis posits that accumulation of Aβ peptides, particularly the more hydrophobic and aggregation-prone Aβ42, triggers a pathogenic cascade, leading to neurodegeneration in AD

beta-amyloid protein 42

Evidence (Publications)

Hardy J, Selkoe DJ The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics. Science (New York, N.Y.). 2002 Jul 19 ;Vol 297 Issue 5580 :353-6

Paolo Ciccarese on November 07, 2011 (0)

Claim: amyloid accumulation is not an obligatory feature of dementia or neurodegeneration because neurodegenerative dementias lacking amyloid pathology (e.g., FTD) have been well described

Evidence (Publications)

Hardy J, Selkoe DJ The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics. Science (New York, N.Y.). 2002 Jul 19 ;Vol 297 Issue 5580 :353-6

Paolo Ciccarese on November 07, 2011 (0)

Claim: Conditional knockout mice in which PS1 and PS2 are selectively inactivated in the adult cerebral cortex develop age-related, progressive neurodegeneration characterized by hallmarks of AD neuropathology, including synaptic loss, neuronal cell death, astrogliosis and tau hyperphosphorylation.

presenilin-1 presenilin-2

Evidence (Publications)

Saura CA, Choi SY, Beglopoulos V, Malkani S, Zhang D, Shankaranarayana Rao BS, Chattarji S, Kelleher RJ, Kandel ER, Duff K, Kirkwood A, Shen J Loss of presenilin function causes impairments of memory and synaptic plasticity followed by age-dependent neurodegeneration. Neuron. 2004 Apr 8 ;Vol 42 Issue 1 :23-36

Figure 5 View of the scientific discourse related to a publication. Example of view of the scientific discourse related to a publication. Three claims have been encoded with their supportive evidence and related ontological terms.

Curation is a crucial aspect of scientific publication and therefore an important aspect for both our annotation ontology and our annotation tool. We enable curation for annotation generated by both humans and text mining services. In the case of automatic generated annotation, the tool allows curators to judge each annotation item (or set of annotation items) according to a configurable set of judgment categories. By default avalable categories are: “wrong”, “right”, “too broad”, “unclear” – where “unclear” means the curator is unable to understand or judge the result.

Every time a curator judges and responds to a result, s/he can also provide motivation that can be used later on for further evaluation. The users can also provide, through manual annotation, the list of entities that have been overlooked by the text mining algorithm. Eventually, the curated results and the terms suggestions can be exported and sent back to the text mining providers for measuring the performance of their tools or even for implementing incremental learning of their algorithms.

As several users may produce annotation on the same document, several users or curators may therefore curate the same results. The annotation tool enables both concurrent and collaborative annotation and curation processes.

Provenance, access control and RDF sharing

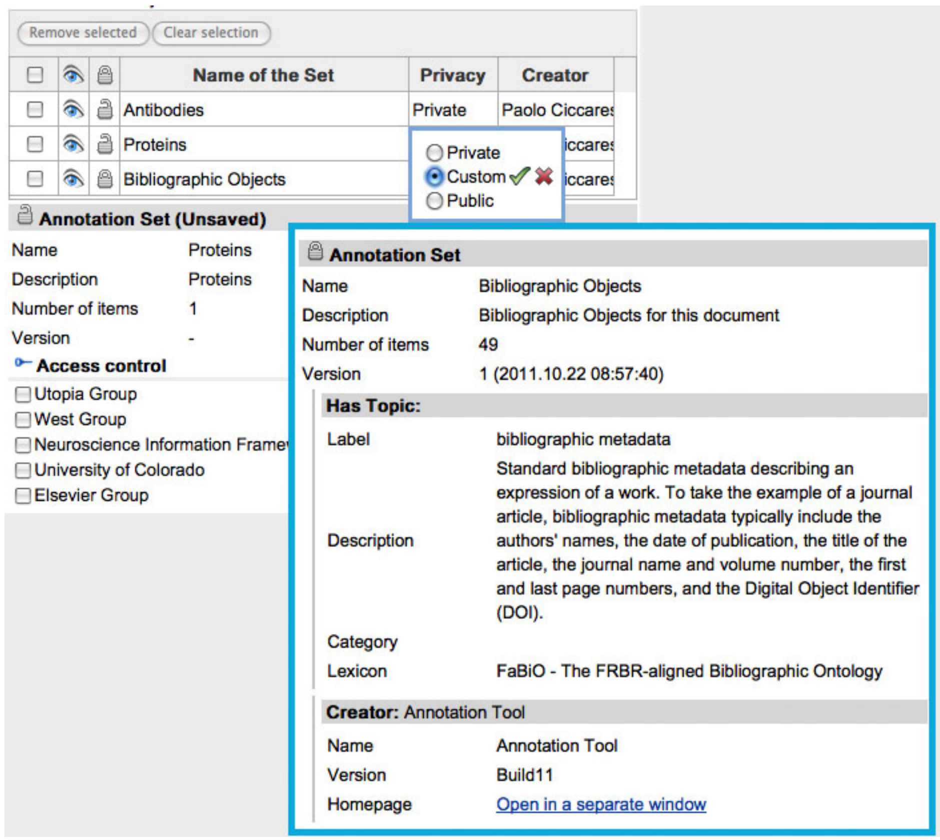
In working with online scientific communities, we are particularly aware of the importance of provenance tracking for establishing trust and properly documenting evolution of the science. AO offers a rich set of properties for modeling provenance based on the Provenance Authoring and Versioning (PAV) ontology originally developed for the SWAN project [22,51]. Our annotation tool tracks all the provenance aspects transparently while the user performs the annotation process. For every piece of annotation and annotation curation, the tool records the originating user, date, and the specific version of any software or web service involved. DOMEO also implements another feature of AO: the Annotation Set, a mechanism for grouping annotation items. The notion of an Annotation Set was included in AO to assist in annotation organization. Sets can be used, for instance, to collect items of the same type – i.e. proteins or genes

–, to show/hide multiple items, and to define the corresponding access policy. Using the annotation tool it is also possible to define, for each set, which users will be able to access the annotation items (Figure 6): only the creator (personal annotation), selected groups, or everybody (public annotation).

The annotation and curation items, together with all the provenance data, can be then serialized in RDF format according to the AO model. Serialization includes RDF representing aspects of the domain ontologies used in any annotation, as well as the AO RDF itself.

Conclusions

The Domeo annotation tool is a web component developed using a combination of the Google Web Toolkit (GWT) and pure JavaScript. As the GWT code is eventually compiled into JavaScript, the final product consists in a JavaScript file to be imported by the hosting page. In order to work properly, Domeo needs a series of services that can be developed potentially in any language or platform able to eventually publish a page that accommodates the Domeo JavaScript component. After the Grails implementation, we are now in the process of starting the integration of Domeo with the Drupal open source Content Management System - using the PHP programming language. We are confident the same process could be replicated, in the near future, with other CMSs and programming languages.



□			Name of the Set	Privacy	Creator
□			Antibodies	Private	Paolo Ciccares
□			Proteins		iccares
□			Bibliographic Objects		iccares

○ Private
● Custom ✓ ✗
○ Public

Annotation Set (Unsaved)

Name	Proteins
Description	Proteins
Number of items	1
Version	-

Access control

□ Utopia Group
□ West Group
□ Neuroscience Information Frame
□ University of Colorado
□ Elsevier Group

Annotation Set

Name	Bibliographic Objects
Description	Bibliographic Objects for this document
Number of items	49
Version	1 (2011.10.22 08:57:40)

Has Topic:

Label	bibliographic metadata
Description	Standard bibliographic metadata describing an expression of a work. To take the example of a journal article, bibliographic metadata typically include the authors' names, the date of publication, the title of the article, the journal name and volume number, the first and last page numbers, and the Digital Object Identifier (DOI).
Category	
Lexicon	FaBiO - The FRBR-aligned Bibliographic Ontology

Creator: Annotation Tool

Name	Annotation Tool
Version	Build11
Homepage	Open in a separate window

Figure 6 Annotation Sets access control. Annotation Sets access control.

Fifty alpha testers have been using Domeo in the past year. Users' feedback has been important to fix bugs and drive user interface improvments. As today, the beta release for the NIF (Neuroscience Information Framework) is deployed within NIF and it is planned to be brought into production within one of the next releases of the NIF portal (probably version 4.2).

A detailed roadmap has been defined to further improve the features most important for text miners. Significant work has also been carried out to integrate Domeo services with the Apache UIMA framework, so that textminers using that architecture will be able to display and curate the results their text mining with our tool.

With the Domeo tool, the Domeo Annotation Framework in general, and the collaborations currently in place, we expect to be able to publish large quantities of high quality annotations on scientific documents in RDF AO format. The published annotation will include the content of the AlzSWAN knowledge base (http://hypothesis.alzforum.org) with the discourse elements – claims, hypotheses, and questions – linked to the correspondent text in original papers. We also note that annotation produced with our tool can be displayed on the corresponding PDF documents in the Utopia application [2,52] as Utopia can now consume AO RDF. We are currently working to with the Utopia group to enable the opposite workflow: producing annotation on a PDF of a scientific paper, and displaying it on the HTML version.

Methods

Domeo was developed upon an initial set of requirements accumulated in developing curation-intensive biomedical knowledge bases and scientific online communities, including

- the AlzSWAN knowledge base (http://hypothesis.alzforum.org) [53], a customization of the Semantic Web Applications in Neuromedicine (SWAN) platform for hypothesis management in Alzheimer disease research;
- the Science Commons Antibodies Resource [54], an OWL model for formally representing antibodies and an associated collection of formally defined commercial antibodies;
- StemBook [55] (http://www.stembook.org), a web portal for the Stem Cell community collecting a comprehensive set of original review articles indexed by NLM; and
- PDOnline [56] (http://pdonlineresearch.org), a web portal for the Parkinson Disease researcher community, collecting several relevant resources including extensive online discussions by scientists.

Our approach was iterative and the application code was developed in tandem with the OWL representation of the annotation metadata in Annotation Ontology (AO). After completing an initial pilot and deploying some research code, we began taking on additional incremental use cases. These included:

- Annotation and curation of hypotheses in pharmaceutical drug discovery, based on requirements of a major international drug company;
- Annotation and curation of antibodies in literature, linked to the Neuroscience Information Framework (NIF) antibodies registry;
- Layering of multi-community annotation sets on documents;
- Curation, comparison and correction of annotations developed by automated text-mining algorithms; and

- Creating scientific claims linked directly to evidence in the form of archived datasets.

For each of these use cases a corresponding community of users and user representative was identified. These were consulted extensively about detailed incremental requirements, user interface and fitness for purpose of the developed software. The NIF antibodies project now uses this toolkit on an ongoing basis. We have received several requests from biomedical textmining groups to use this software as well and are currently working to organize textmining users in a way that will make ongoing support as straightforward as possible.

Availability and requirements

- Project name: Domeo
- Project home page: http://annotationframework.org
- Operating system(s):
  - ○ Client: browser-based, platform independent
  - ○ Server: Linux
- Programming language:
  - ○ Client: Google Web Toolkit, Javascript
  - ○ Server: JAVA and Groovy, using the Grails framework
- License: release under Apache 2.0 planned for January 2012
- Any restrictions to use by non-academics: compliance with Apache 2.0 license
- Current deployment status: beta release.
- Website: http://annotationframework.org

List of abbreviations used

AF: Domeo Annotation Framework; AlzSWAN: Alzheimer Disease Hypothesis Knowledge Base; AO: Annotation Ontology; CMS: Content Mangement System; GWT: Google Web Toolkit; HTML: HyperText Markup Language; NCBO: National Center for Biomedical Ontology; NIF: Neuroscience Information Framework'; OWL: Web Ontology Language; PAV: Provenance, Authoring and Versioning Ontology; PDF: Portable Document Format; PHP: PHP: Hypertext Preprocessor; RDF: Resource Description Framework; RECS: (Run, Encode, Curate, Share); REST: Representational State Transfer; SKOS: Simple Knowledge Organization System; SWAN: Semantic Web Applications in Neuromedicine; URI: Uniform Resource Identifier; URL: Uniform Resource Locator; XHTML: eXtensible HyperText Markup Language; XML: eXtensible Markup Language.

Acknowledgements

Development of Domeo was funded by a grant from the National Institute on Drug Abuse, National Institutes of Health, as part of the Neuroscience Information Framework; by a grant from EMD Serono, Inc. as part of the MS Discovery Forum project; by a grant from Elsevier Laboratories; and by a grant from Eli Lilly and Company. We are most grateful for the financial support of these organizations.

We thank Professor Maryann Martone and Anita Bandrowski of the University of California at San Diego; Anita deWaard, Bradley Allen and Anthony Scerri of Elsevier B.V., Adam West and Ernest Dow of Eli Lilly and Company, and Carole Goble of the University of Manchester, for their continuing support and for many fruitful discussions and much joint work. We also thank Steve Pettifer of University of Manchester for the work toward integration of our tool with the Utopia PDF annotator.

This article has been published as part of *Journal of Biomedical Semantics* Volume 3 Supplement 1, 2012: Proceedings of the Bio-ontologies Special Interest Group 2011. The full contents of the supplement are available online at http://www.jbiomedsem.com/supplements/3/S1

Author details

[1]Harvard Medical School and Massachusetts General Hospital, Boston MA, USA. [2]Balboa Systems, Newton MA, USA. [3]University of Manchester, School of Computer Science, Manchester, UK.

Authors' contributions

PC is the principal author of the Annotation Ontology (AO), and the software architect and principal developer of the Domeo Annotation Toolkit. Dr. Ciccarese is the main contributor to this paper and produced the figures.
MO co-developed the server-side software and the first pilot of the Domeo annotation tool.
TC conceived of and provided overall guidance for development of the Domeo Annotation Toolkit and Annotation Ontology projects, wrote the background section of this paper, co-wrote other sections, did the overall editing, and is principal investigator of the project.

Competing interests

The authors of this paper declare that they have no competing interests.

Published: 24 April 2012

References

1. Jacobs I, Walsh N: **Architecture of the World Wide Web, Volume One.** *W3C Recommendation* World Wide Web Consortium; 2004.
2. Attwood TK, Kell DB, McDermott P, Marsh J, Pettifer SR, Thorne D: **Utopia documents: linking scholarly literature with research data.** *Bioinformatics* 2010, **26(18)**:i568-i574.
3. Singh J: **Mendeley: A free research management tool for desktop and web.** 2010, **1.**
4. Altman RB, Bergman CM, Blake J, Blaschke C, Cohen A, Gannon F, Grivell L, Hahn U, Hersh W, Hirschman L, *et al*: **Text mining for biology–the way forward: opinions from leading scientists.** *Genome Biology* 2008, **9(Suppl 2)**:S7.
5. Arighi C, Lu Z, Krallinger M, Cohen K, Wilbur W, Valencia A, Hirschman L, Wu C: **Overview of the BioCreative III Workshop.** *BMC Bioinformatics* 2011, **12(Suppl 8)**:S1.
6. Arighi C, Roberts P, Agarwal S, Bhattacharya S, Cesareni G, Chatr-aryamontri A, Clematide S, Gaudet P, Giglio M, Harrow I: **BioCreative III Interactive Task: an Overview.** *BMC Bioinformatics* 2011.
7. Krallinger M, Vazquez M, Leitner F, Salgado D, Chatr-aryamontri A, Winter A, Perfetto L, Briganti L, Licata L, Iannuccelli M: **The Protein-Protein Interaction tasks of BioCreative III: classication/ranking of articles and linking bio-ontology concepts to full text.** *BMC Bioinformatics* 2011.
8. Lu Z, Kao H, Wei C, Huang M, Liu J, Kuo C, Hsu C, Tsai R, Dai H, Okazaki N: **The Gene Normalization Task in BioCreative III.** *BMC Bioinformatics* 2011.
9. Leitner F, Chatr-aryamontri A, Mardis S, Ceol A, Krallinger M, Licata L, Hirschman L, Cesareni G, Valencia A: **The FEBS Letters/BioCreative II.5 experiment: making biological information accessible.** *Nature biotechnology* 2009, **28**:897-899.
10. Kim J, Ohta T, Tsuruoka Y, Tateisi Y, Collier N: **Introduction to the Bio-Entity Task at JNLPBA.** *BioCreative Challenge Evaluation Workshop* 2004.
11. Hirschman L, Yeh A, Blaschke C, Valencia A: **Overview of BioCreAtIvE: critical assessment of information extraction for biology.** *BMC Bioinformatics* 2005, **6(Suppl 1)**:S1.
12. Stein L, Sternberg P, Durbin R, Thierry Mieg J, Spieth J: **WormBase: network access to the genome and biology of Caenorhabditis elegans.** *Nucleic Acids Res* 2001, **29**:82-86.
13. St Pierre S, McQuilton P: **Inside FlyBase: biocuration as a career.** *Fly* 2009, **3(1)**:112-114.
14. Tweedie S, Ashburner M, Falls K, Leyland P, McQuilton P, Marygold S, Millburn G, Osumi-Sutherland D, Schroeder A, Seal R, *et al*: **FlyBase: enhancing Drosophila Gene Ontology annotations.** *Nucleic Acids Res* 2009, **37(Database issue)**: D555-559.
15. Drysdale R: **FlyBase : a database for the Drosophila research community.** *Methods Mol Biol* 2008, **420**:45-59.
16. Ashburner M, Ball C, Blake J, Botstein D, Butler H, Cherry J, Davis A, Dolinski K, Dwight S, Eppig J, *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nature Genetics* 2000, **25(1)**:25-29.
17. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, *et al*: **The Gene Ontology (GO) database and informatics resource.** *Nucleic Acids Res* 2004, **32(Database issue)**:D258-261.
18. Hill DP, Smith B, McAndrews-Hill MS, Blake JA: **Gene Ontology annotations: what they mean and where they come from.** *BMC Bioinformatics* 2008, **9(Suppl 5)**:S2.
19. Schneider M, Lane L, Boutet E, Lieberherr D, Tognolli M, Bougueleret L, Bairoch A: **The UniProtKB/Swiss-Prot knowledgebase and its Plant Proteome Annotation Program.** *Journal of Proteomics* 2009, **72(3)**:567-573.
20. Shapin S: **Pump and Circumstance: Robert Boyle's Literary Technology.** In *The Scientific Revolution.* Oxford: Blackwell; Hellyer M 2003:.
21. **The Annotation Ontology on Google Code.** [http://code.google.com/p/annotation-ontology/].
22. Ciccarese P, Ocana M, Garcia Castro LJ, Das S, Clark T: **An open annotation ontology for science on web 3.0.** *J Biomed Semantics* 2011, **2(Suppl 2)**:S4.
23. McGuinness D, van Harmelen F: **OWL Web Ontology Language.** *W3C Recommendation* 2004.
24. Miles A, Bechhofer S: **SKOS Simple Knowledge Organization System Reference.** *W3C Recommendation* 2009.
25. Wang Z, Dong H, Kelly M, Macklin JA, Morris PJ, Morris RA: **Filtered-Push: A Map-Reduce Platform for Collaborative Taxonomic Data Management.** *Computer Science and Information Engineering, 2009 WRI World Congress on: March 31 2009-April 2 2009* 2009, 731-735.
26. **Tags4Labs.** [http://www.biotea.ws/node/3].
27. Garcia-Castro A, Labarga A, Garcia L, Giraldo O, Montaña C, Bateman JA: **Semantic Web and Social Web heading towards Living Documents in the Life Sciences.** *Web Semantics: Science, Services and Agents on the World Wide Web* 2010, **8(2-3)**:155-162.
28. Waldrop MM: **Big data: Wikiomics.** *Nature* 2008, **455(7209)**:22-25.
29. Waldrop MM: **Science 2.0.** *Scientific American* 2008, **298(5)**:68-73.
30. Bos N, Zimmerman A, Olson J, Yew J, Yerkie J, Dahl E, Olson G: **From shared databases to communities of practice: A taxonomy of collaboratories.** *Journal of Computer-Mediated Communication* 2007, **12(2)**:article 16.
31. Shotton D: **Semantic publishing: the coming revolution in scientific journal publishing.** *Learned Publishing* 2009, **22(2)**:85-94.

32. Renear AH, Palmer CL: **Strategic Reading, Ontologies, and the Future of Scientific Publishing.** *Science* 2009, **325(5942)**:828-832.
33. Becket D: **RDF/XML Syntax Specification (Revised).** *W3C Recommnedation* 2004.
34. Rocher G, Brown J: **The Definitive Guide to GRAILS.** Berkeley CA: Apress; 2009.
35. Henry K: **A crash overview of groovy.** *Crossroads* 2006, **12(3)**.
36. **Survey of Text Mining: Clustering, Classification, and Retrieval.** Heidelberg Springer Verlag;Berry MW, Castellanos M , 2 2007:.
37. Krallinger M, Morgan A, Smith L, Leitner F, Tanabe L, Wilbur J, Hirschman L, Valencia A: **Evaluation of text-mining systems for biology: overview of the Second BioCreative community challenge.** *Genome Biol* 2008, **9(Suppl 2)**:S1.
38. Hirschman L, Yeh A, Blaschke C, Valencia A: **Overview of BioCreAtIvE: critical assessment of information extraction for biology.** *BMC Bioinformatics* 2005, **6(Suppl 1)**:S1.
39. Leitner F, Valencia A: **A text-mining perspective on the requirements for electronically annotated abstracts.** *FEBS letters* 2008, **582(8)**:1178-1181.
40. Leitner F, Chatr-aryamontri A, Mardis SA, Ceol A, Krallinger M, Licata L, Hirschman L, Cesareni G, Valencia A: **The FEBS Letters/BioCreative II.5 experiment: making biological information accessible.** *Nat Biotechnol* 2010, **28(9)**:897-899.
41. Kim J-D, Ohta T, Pyysalo S, Kano Y, Tsujii Ji: **Overview of BioNLP'09 shared task on event extraction.** *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task; Boulder, Colorado. 1572342* Association for Computational Linguistics; 2009, 1-9.
42. Jonquet C, Musen MA, Shah N: **A system for ontology-based annotation of biomedical data.** *International Workshop on Data Integration in the Life Sciences, DILS'08: 2008; Evry, France* 2008.
43. Jonquet C, Musen MA, Shah NH: **Help will be provided for this task: Ontology-Based Annotator Web Service.** Stanford, CA: Stanford Center for Biomedical Informatics Research, Stanford University School of Medicine; 2008, 16.
44. **Bioportal REST Services.** [http://www.bioontology.org/wiki/index.php/BioPortal_REST_services].
45. **Semantic Web Applications in Neuromedicine (SWAN) Ontology, W3C Interest Group Note 20 October 2009.** [http://www.w3.org/2001/sw/hcls/notes/swan/].
46. Ciccarese P, Wu E, Wong G, Ocana M, Kinoshita J, Ruttenberg A, Clark T: **The SWAN biomedical discourse ontology.** *J Biomed Inform* 2008, **41(5)**:739-751.
47. Clark T, Ciccarese P, Attwood T, Waard Ad, Pettifer S: **A Round-Trip to the Annotation Store: Open, Transferable Semantic Annotation of Biomedical Publications.** *Beyond the PDF: January 19-21, 2011* University of California at San Diego; 2011.
48. Gao Y, Kinoshita J, Wu E, Miller E, Lee R, Seaborne A, Cayzer S, Clark T: **SWAN: A distributed knowledge infrastructure for Alzheimer Disease research.** *Web Semantics: Science, Services and Agents on the World Wide Web* 2006, **4(3)**:222-228.
49. Altman M, Andreev L, Diggory M, King G, Sone A, Verba S, Kiskis DL: **A Digital Library for the Dissemination and Replication of Quantitative Social Science Research.** *Social Science Computer Review* 2001, **19(4)**:458-470.
50. Altman M, King G: **A Proposed Standard for the Scholarly Citation of Quantitative Data.** *DLib Magazine* 2006, **13(3/4)**.
51. **PAV Ontology on Google Code: PAV Ontology 2.0.** [http://code.google.com/p/pav-ontology/].
52. Attwood TK, Kell DB, McDermott P, Marsh J, Pettifer SR, Thorne D: **Calling International Rescue: knowledge lost in literature and data landslide!** *Biochemical Journal* 2009, **424(3)**:317-333.
53. Clark T, Kinoshita J: **Alzforum and SWAN: the present and future of scientific web communities.** *Brief Bioinform* 2007, **8(3)**:163-171.
54. **Science Commons Semantic Resources Project: Antibody Resource.** [http://neurocommons.org/page/Semantic_resources_project/Antibodies].
55. Das S, Girard L, Green T, Weitzman L, Lewis-Bowen A, Clark T: **Building biomedical web communities using a semantically aware content management system.** *Brief Bioinform* 2009, **10(2)**:129-138.
56. Das S, Rogan M, Kawadler H, Corlosquet S, Brin S, Clark T: **PD Online: a case study in scientific collaboration on the Web.** *Workshop on the Future of the Web for Collaborative Science, 19th International World Wide Web Conference: April 26-30, 2010; Raleigh, NC, USA* 2010.

doi:10.1186/2041-1480-3-S1-S1
Cite this article as: Ciccarese *et al.*: Open semantic annotation of scientific publications using DOMEO. *Journal of Biomedical Semantics* 2012 **3**(Suppl 1):S1.