24

Social and Semantic Web Technologies for the Text-To-Knowledge Translation Process in Biomedicine

Carlos Cano[1], Alberto Labarga[1], Armando Blanco[1] and Leonid Peshkin[2]
[1]*Dept. Computer Science and Artificial Intelligence. University of Granada. c/. Daniel Saucedo Aranda, s/n 18071, Granada,*
[2]*Dept. of Systems Biology, Harvard Medical School. 200 Longwood Ave, Boston, MA 02115*
[1]*Spain*
[2]*USA*

1. Introduction

Currently, biomedical research critically depends on knowledge availability for flexible re-analysis and integrative post-processing. The voluminous biological data already stored in databases, put together with the abundant molecular data resulting from the rapid adoption of high-throughput techniques, have shown the potential to generate new biomedical discovery through integration with knowledge from the scientific literature.

Reliable information extraction applications have been a long-sought goal of the biomedical text mining community. Both named entity recognition and conceptual analysis are needed in order to map the objects and concepts represented by natural language texts into a rigorous encoding, with direct links to online resources that explicitly expose those concepts semantics (see Figure 1).

Naturally, automated methods work at a fraction of human accuracy, while expert curation has a small fraction of computer coverage. Hence, mining the wealth of knowledge in the published literature requires a hybrid approach which combines efficient automated methods with highly-accurate expert curation. This work reviews several efforts in both directions and contributes to advance the hybrid approach.

Since Life Sciences have turned into a very data-intensive domain, various sources of biological data must often be combined in order to build new knowledge. The Semantic Web offers a social and technological basis for assembling, integrating and making biomedical knowledge available at Web scale.

In this chapter we present an open-source, modular friendly system called BioNotate-2.0, which combines automated text annotation with distributed expert curation, and serves the resulting knowledge in a Semantic-Web-accessible format to be integrated into a wider bio-medical inference pipeline. While this has been an active area of research and development for a few years, we believe that this is an unique contribution which will be widely adopted to enable the community effort both in the area of further systems development and knowledge sharing.

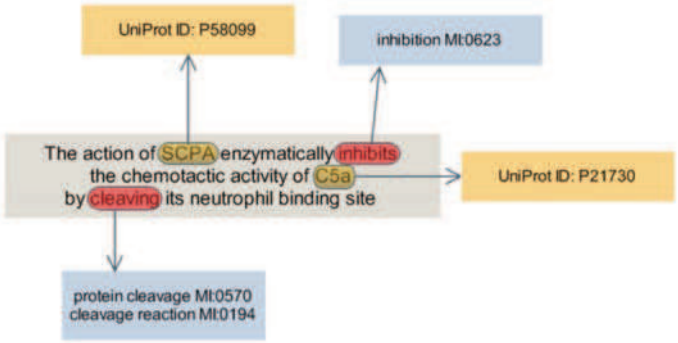

Fig. 1. Some annotations on a piece of biomedical text. Entities of interest and evidences of interaction are marked up in the text and mapped to external resources. In this case, genes and proteins are mapped to UniProt entries, and interaction keywords are linked to terms from the ontology PSI-Molecular Interactions (PSI-MI). Annotated snippets constitute a *corpus*. Large *corpora* are required to train Machine Learning systems.

Particularly, this chapter describes the design and implementation of BioNotate-2.0 for: 1) the creation and automatic annotation of biomedical corpora; 2) the distributed manual curation, annotation and normalization of extracts of text with biological facts of interest; 3) the publication of curated facts in semantically enriched formats, and their connection to other datasets and resources on the Semantic Web; 4) the access to curated facts with a Linked Data Browser.

Our aim is to provide the community with a modular and open-source annotation platform to harness the great collaborative power of the biomedical community over the Internet and allow the dissemination and sharing of semantically-enriched biomedical facts of interest. Specifically, we illustrate several cases of use of BioNotate 2.0 for the annotation of biomedical facts involving the identification of genes, protein-protein, and gene-disease relationships. By design, the provided tools are flexible and can implement a wide variety of annotation schemas.

1.1 Information extraction systems in biology

Efficient access to information contained in on-line scientific literature collections is essential for life science research, playing a crucial role from the initial stage of experiment planning to the final interpretation and communication of the results. The biological literature also constitutes the main information source for manual curation of biological databases and the development of ontologies (Krallinger, Valencia & Hirschman, 2008). However, biological databases alone cannot capture the richness of scientific information and argumentation contained in the literature (Krallinger, Valencia & Hirschman, 2008; Baumgartner Jr et al., 2007).

The biomedical literature can be seen as a large integrated, but unstructured data repository. It contains high quality and high-confidence information on genes that have been studied for decades, including the gene's relevance to a disease, its reaction mechanisms, structural information and well characterized interactions. However, an accurate and normalized representation of facts and the mapping of the information contained within papers onto existing databases, ontologies and online resources has traditionally been almost negligible.

Extracting facts from literature and making them accessible is approached from two

directions: first, manual curation efforts develop ontologies and vocabularies to annotate gene products based on statements in papers; second, text mining aims to automatically identify entities and concepts in text using those controlled vocabularies and ontologies and employing information retrieval and natural language processing techniques (Winnenburg et al., 2008).

The best known community-wide effort for the evaluation of text-mining and information extraction systems in the biological domain is the BioCreative (Critical Assessment of Information Extraction systems in Biology) challenge (Krallinger, Morgan, Smith, Leitner, Tanabe, Wilbur, Hirschman & Valencia, 2008). The goal of BioCreative has been to present algorithmic challenges that would result in scalable systems targeted for the general community of biology researchers as well as for the more specialized end-users, such as annotation database curators. One special case is the Gene Mention recognition, which evaluates systems that find mentions of genes and proteins in sentences from PubMed abstracts. Another case is the Gene Normalization, focused on providing direct links between texts and actual gene and protein-focused records in biological databases. In contrast to these indexing challenges, the Interaction Article Subtask (IAS) addressed the first step in many biological literature review tasks, namely the retrieval/classification and ranking of relevant articles according to a given topic of interest. Particularly, in the second edition of the challenge, the goal was to classify a collection of PubMed titles and abstracts based on their relevance for the derivation of protein-protein interaction annotations. As a direct result of these competitive evaluations, a platform which integrates the participant's servers for detecting named entities and relationships has been released (Leitner et al., 2008). This platform, called BioCreative Meta-Server (`http://bcms.bioinfo.cnio.es/`), allows to simultaneously annotate a piece of biomedical text using different NLP tools and systems and visualize and compare their results. U-compare (`http://u-compare.org/`) is a similar platform which allows the user to design a customized text-mining annotation pipeline using different available tools and corpora (Kano et al., 2009).

Recently, the efforts have shifted from the localization and annotation of the character strings towards the identification of concepts (Hunter et al., 2008). Concepts differ from character strings in that they are grounded in well defined knowledge resources. Thus, concept recognition provides an unambiguous semantic representation of what the text denotes. A related initiative is the Collaborative Annotation of a Large Biomedical Corpus (CALBC, `http://www.calbc.eu`). CALBC is an European support action addressing the automatic generation of a very large, community-wide shared text corpus annotated with biomedical entities. Their aim is to create a broadly scoped and diversely annotated corpus (150,000 Medline immunology-related abstracts annotated with approximately a dozen semantic types) by automatically integrating the annotations from different named entity recognition systems. The CALBC challenge involves both Name Entity Recognition and Concept recognition tasks.

In theory, text mining is the perfect solution to transforming factual knowledge from publications into database entries. However, the field of computational linguistics have not yet developed tools that can accurately parse and analyse more than 30% of English sentences in order to transform them into a structured formal representation (Rebholz-Schuhmann et al., 2005).

On the other hand, manually curated data is precise, because a curator, trained to consult the literature and databases, is able to select only high-quality data, and reformat the facts according to the schema of the database. In addition, curators select quotes from the

text as evidence supporting the identified fact, and those citations are also added to the database. Curators know how to define standards for data consistency, in particular, the most relevant terminology, which has led to the design of standardized ontologies and controlled vocabularies. The issue with curation of data is that it is time consuming and costly, and therefore has to focus on the most relevant facts. This undermines the completeness of the curated data, and curation teams are destined to stay behind the latest publications. Therefore, an environment where manual curation and text mining can effectively and efficiently work together is highly desirable (Rebholz-Schuhmann et al., 2005).

1.2 Social annotation and tagging in life sciences

Web resources such as Delicious (`http://delicious.com`), or Connotea (`http://connotea.org`) facilitate the tagging of online resources and bibliographic references. These online tools harness the collective knowledge that is modeled by the collective tagging. Collaboration is thus based on similarities in tags and tagged objects. The more annotations the system gets, the better the chances are for users to interact with researchers who share similar interests, such as elucidating the same pathway, methodology or gene function.

General-purpose annotation tools, such as Knowtator *Knowtator* (n.d.), WordFreak *WordFreak* (n.d.), SAFE-GATE (Cunningham et al., 2002) and iAnnotate *iAnnotate* (n.d.), can be adapted to the annotation of biomedical entities and relationships in scientific texts. Some BioNLP groups have also created customized annotation tools implementing their specific annotation schemas, such as the Xconc Suite's implementation for annotating events in the GENIA corpus (Kim et al., 2008). While these tools allow a restricted group of well-trained annotators to curate corpora, they are not intended for massive annotation efforts by the broad research community.

In contrast, our work is largely inspired by the recent distributed and collaborative annotation efforts that have emerged, such as those in the image analysis domain (*Google Image Labeler*, n.d.; Russell et al., 2008) or related to the Amazon Mechanical Turk (AMT) annotation web services (*Amazon's Mechanical Turk*, n.d.; Callison-Burch, 2009). These efforts have shown a great potential since they allow any interested user world-wide to contribute in the annotation task.

In a recent work, Snow *et al*. (Snow et al., 2008) show the effectiveness of collaborative non-expert annotation on some traditional NLP problems such as emotion recognition from text, word synonymy, hypothesis inference, chronological ordering of facts and ambiguity resolution. Particularly, this work demonstrates that the accuracy achieved by a Machine Learning system trained with annotations by a few non-expert curators equals the accuracy achieved by the same system trained with annotations made by experts. For example, for the emotion recognition from text task, 4 non-expert annotations (in average) per item are enough to emulate the results of one expert annotation, with significantly reduced costs (Snow et al., 2008). After this pioneer work, others have proposed and evaluated the effectiveness of using AMT for massive collaborative annotation of corpora to train machine learning systems (Raykar et al., 2009; Donmez et al., 2009; Callison-Burch, 2009; Carlson et al., 2010).

Within the biomedical field, the notion of community annotation has also recently started to be adopted. For instance, WikiProteins (Mons et al., 2008) or WikiGene (Maier et al., 2005) deliver appropriate environments in which it is possible to address the annotation of genes and proteins. Since 2007, GoPubMed also includes a collaborative curation tool for the annotation of concepts and Pubmed authors profiles. While these efforts allow the wider research community to directly benefit from the generation and peer-review of knowledge at

minimal cost, they are not intended for the creation of corpora for training NLP tools. Such capabilities allow a feedback from the curation effort back to the automated processing in order to improve its accuracy, in turn enabling human curation to focus on more sophisticated instances.

Baral et al. (Baral et al., 2007), proposed a methodology where the community collaboratively contributes to the curation process. They used automatic information extraction methods as a starting point, and promote mass collaboration with the premise that if there are a lot of articles, then there must be a lot of readers and authors of these articles. Our approach is similar to that implemented by their system, called CBioC. This system allows the user to annotate relationships between biomedical concepts while browsing PubMed records. The user is presented with potential relationships from the current record extracted by automated tools or suggested by other users. Registered users can add new relationships and vote for suggested relationships. For a given PubMed record, a relationship is defined by providing the literals of the two interacting entities and the keywords of the interaction. However, CBioC does not allow to highlight the exact mentions of these words in the text. Furthermore, the users can only access to the annotated facts from within CBioC. The whole corpus of annotations is not directly available until it is distributed by the CBioC team.

Within the publishing industry, there has also been a series of efforts in promoting community interaction by Social Networks. BioMedExperts (BME, `http://www.biomedexperts.com`) is a professional network in which literature references are used to support interaction. Although this system does not support tagging by users, it does support automatic tagging based on a reference terminology; thus allowing the identification of researchers with similar interests. Nature Network (`http://network.nature.com/`) works in a similar way; however it does not facilitate any controlled vocabulary for annotating the literature references.

1.3 The emerging role of the semantic web technologies in life sciences

Current research in biology heavily depends on the availability and efficient use of information. Life sciences have turned into a very data-intensive domain and, in order to build new knowledge, various sources of biological data must often be combined. Therefore, scientists in this domain are facing the same challenges as in many other disciplines dealing with highly distributed, heterogeneous and voluminous data sources.

The Semantic Web offers a social and technological basis for assembling, integrating and making biomedical knowledge available at Web scale. Its emphasis is on combining information using standard representation languages and allowing access to that information via standard web protocols and technologies to leverage computation, such as in the form of inference and distributable query.

As the Semantic Web is being introduced into the Life Sciences, the basis for a distributed knowledge-base that can foster biological data analysis is laid. Biomedical ontologies provide essential domain knowledge to drive data integration, information retrieval, data annotation, natural-language processing and decision support, and so, new ontologies are being developed to formalize knowledge (Shah et al., 2009). Such major bioinformatics centers as the European Bioinformatics Institute or the National Center for Biotechnology Information provide access to over two hundred biological resources. Links between different databases are an important basis for data integration, but the lack of a common standard to represent and link information makes data integration an expensive business.

Recently, such key databases as Uniprot (Bairoch et al., 2005) began providing data

access in RDF format. Resource Description Framework (http://www.w3.org/RDF/) is a core technology for the World Wide Web Consortium's Semantic Web activities (http://www.w3.org/2001/sw/) and is therefore well suited to work in a distributed and decentralized environment. The RDF data model represents arbitrary information as a set of simple statements of the form subject-predicate-object. To enable the linking of data on the Web, RDF requires that each resource must have a (globally) unique identifier. These identifiers allow everybody to make statements about a given resource and, together with the simple structure of the RDF data model, make it easy to combine the statements made by different people (or databases) to allow queries across different datasets. RDF is thus an industry standard that can make a major contribution to solve two important problems of bioinformatics: distributed annotation and data integration.

The Bio2RDF project has successfully applied these semantic web technologies to publicly available databases by creating a knowledge space of RDF documents linked together with normalized URIs and sharing a common ontology (Belleau et al., 2008). The benefits promised by the Semantic Web include aggregation of heterogeneous data using explicit semantics, simplified annotation and sharing of findings, the expression of rich and well-defined models for data aggregation and search, easier reuse of data in unanticipated ways, and the application of logic to infer additional insights. The Linking Open Drug Data (LODD) (Jentzsch et al., n.d.) task within the W3C's Semantic Web for Health Care and Life Sciences Interest Group is another related initiative that gathered a list of data sets that include information about drugs, and then determined how the publicly available data sets could be linked together. The project has recently won the first prize of the Linking Open Data Triplification Challenge, showing the importance of Linked Data to the health care and life sciences.

In addition, the concept of nanopublication (Mons & Velterop, 2009) has recently emerged to contribute to model and share Life Sciences discoveries using Semantic Web technologies. A nanopublication is defined as a core scientific statement (e.g. "malaria is transmitted by mosquitos") with associated annotations (i.e. evidence supporting this biological fact, references to the authors of this assertion, etc.) which can be represented as a Named Graph /RDF model. Such representation makes for efficient vehicle of knowledge dissemination and large-scale aggregation, due to its machine-readable characteristics.

2. Proposed approach

In this work we present an integrated approach to concept recognition in biomedical texts, which builds upon both the Semantic Web, which values the integration of well-structured data, and the Social Web, which aims to facilitate interaction amongst users by means of user-generated content. Our approach combines automated named entity recognition tools with manual collaborative curation and normalization of the entities and their relations for a more effective identification of biological facts of interest. Identified facts are converted to a standardized representation for making connections to other datasets and resources on the Semantic Web.

The system is composed of five basic modules which cover the different stages of the annotation pipeline: administration, search, automatic annotation, manual curation and publication. Figure 2 shows how these modules are interconnected.

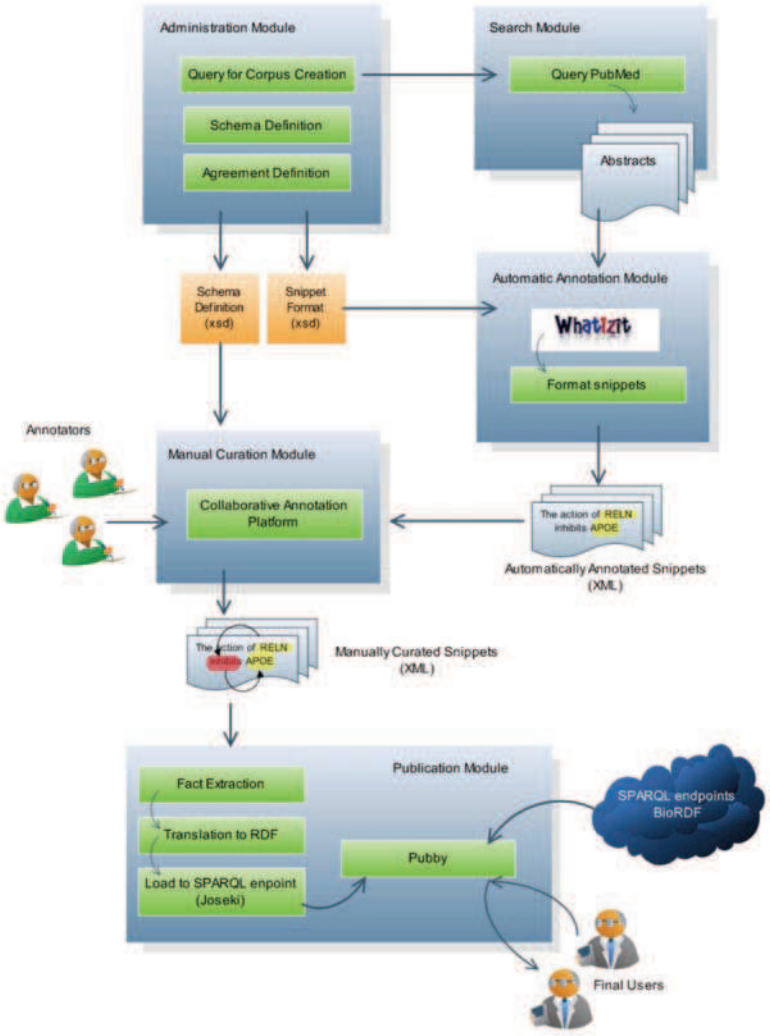

Fig. 2. System architecture of BioNotate-2.0 represents distinct modules and their interconnections.

2.1 Administration module

The administration module allows users to generate the problem definition, the annotation schema and the format for the snippets that will be employed in the annotation and curation tasks. It consists of an intuitive user interface in which administrators can define entities and relationships of interest for the annotation task and provide a function for determining whether two annotations made by different users significantly agree. As part of the problem definition, administrators can also provide the references to the bio-ontologies

or terminological resources which will be used to normalize the entities of interest in the annotation task. Finally, they are also allowed to upload their own corpus or create it by providing query terms and making use of the automatic retrieval module.

2.2 Automatic retrieval and annotation module

To generate the base collection, users can start by sending a query to the system. This query is forwarded to Pubmed or Citexplore. Returned abstracts (and full texts, if available) are presented to the user, who can refine the query, remove non relevant articles and save the query for later updates.

This module also includes resources for carrying out an initial automatic annotation of the retrieved publications using Named Entity Recognition (NER) and automatic text-mining systems. This first annotation eases latter manual curation efforts by providing fast and moderately accurate results, and enables textual semantic markup to be undertaken efficiently over big collections. Our system uses Whatizit web services (Mcwilliam et al., 2009; Rebholz-Schuhmann et al., 2008) to annotate entities of interest and their relationships as defined by the administrator. Whatizit is a Java-powered NER system that searches for terms in the text that match those included in vast terminological resources, allowing morphological variations (Kirsch et al., 2006). Whatizit also considers syntactic features and POS tags obtained by TreeTagger (Schmid, 1994). Whatizit implements different modules depending on the NEs or relations to be identified. Our system includes the following:

- whatizitSwissprot: focuses on the identification and normalization of names of genes and proteins.
- whatizitChemical: focuses on the identification of chemical compounds based on the ChEBI terminology (Degtyarenko et al., 2008) and the OSCAR3 NER system (Corbett & Murray-Rust, 2006).
- whatizitDisease: focuses on the extraction of names of diseases based on MEDLINE terminology.
- whatizitDrugs: identifies drugs using DrugBank terminology (`http://redpoll.pharmacy.ualberta.ca/drugbank/`).
- whatizitGO: identifies GO terms.
- whatizitOrganism: identifies species and organisms based on NCBI taxonomy.
- whatizitProteinInteraction: identifies protein-protein (gene-gene) interactions using Protein Corral (`http://www.ebi.ac.uk/Rebholz-srv/pcorral`).
- whatizitSwissprotGo: detects protein-GO term relationships using UniProtKb/Swiss-Prot terminological resources.

For a complete list of available Whatizit modules, refer to `http://www.ebi.ac.uk/webservices/whatizit/info.jsf`.

2.3 Collaborative annotation tool

The central idea of our approach is to leverage annotations contributed by the users and utilize it as feedback to improve automatic classification. Annotation is generally a simple task for the user and may amount to a ”yes/no” vote on whether the current annotation is correct when examining an individual entry. More sophisticated schemas may require specialized domain expertise on the problem addressed. The collaborative manual annotation tool is

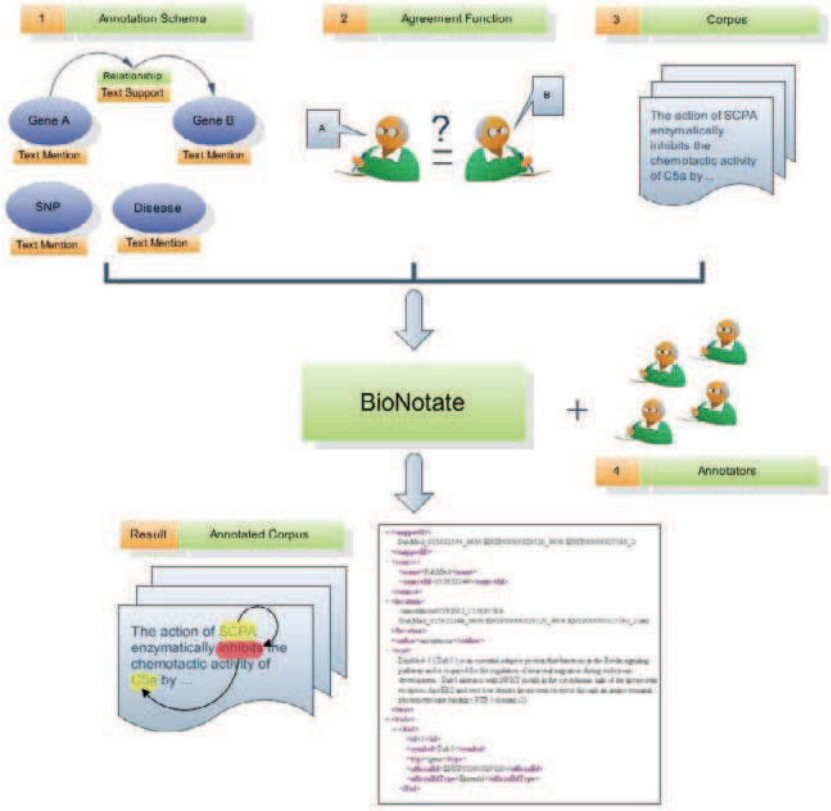

Fig. 3. Information workflow *in* and *out* of the collaborative annotation module. The annotation schema, function for agreement among annotators, and un-annotated corpora are provided as input. The collaborative annotation tool supports a distributed annotation effort with contributions from different users. Consensus annotations are provided as output.

based on BioNotate (Cano et al., 2009)– an open source platform for supporting distributed collaborative annotation efforts via a simple interface over a standard Internet browser-based client/server system.

2.3.1 Architecture of the annotation tool

Although BioNotate was originally designed for annotating gene-gene and gene-disease relationships, current releases of the tool allow users to define arbitrary annotation schemas. Figure 3 shows the information workflow *in* and *out* of the collaborative annotation module. The annotation schema, agreement function and corpora are provided by the administrator. BioNotate supports a distributed annotation effort with contributions from different users. Consensus annotations following the proposed schema are provided as output.

2.3.2 Technical features of the annotation tool

Here we provide some technical features of the annotation tool:

- Support for parallel and simultaneous annotations by different users.
- Annotator management. The system tracks all the annotations being made by every user.
- Distribution of annotating tasks among annotators. The system implements a *k*-blind annotation process. When an annotator logs in the system and requests a snippet to annotate, he is assigned a new one from the pool of documents pending annotation. The assigned document is picked at random from the documents not previously annotated by this user. Each snippet is annotated by at least *k* different annotators. If the *k* annotations of a snippet do not meet a minimum degree of agreement, the snippet is presented to another annotator at random. The process continues until at least *k* annotations performed on the snippet meet a minimum degree of agreement (See Figure 4).
- Access to the annotation system is available from any computer with Internet access and a modern web browser. Annotators do not need to install any extra software. Our browser-based system allows the annotators to log in the system from any machine and add new annotations at any time. On the client side, the application consists of an intuitive user interface where snippets are displayed and the user can perform annotations on the snippets by highlighting arbitrary chunks of text and assigning any of the available labels. On the server side, several Perl scripts serve on request snippets not yet annotated by the user, save the annotations and check the agreement with previous users (see Figure 5).
- Custom annotation schema. The tool is able to deploy arbitrary annotation schemas provided by the administrator in Extensible Markup Language (XML) format. The current version of BioNotate supports the definition of *entities* and *questions* in the annotation schema. *Entities* refer to the different types of named entities of interest and the relationships between them (i.e., each label available in the annotation interface is defined by an *entity* in the annotation schema). *Questions* define the questions to be asked to the annotator for each snippet (if any), with the set of available answers. The schema definition is rendered into the annotation interface by Extensible Stylesheet Transformations (XSLT).

2.3.3 Annotation schema and function for agreement between annotators

For the demo installation of BioNotate-2.0, we have customized the annotation schema for annotating protein-protein (gene-gene) interactions. Furthermore, we require that the interactors are normalized against Uniprot and the interaction keywords are associated to selected terms from the ontology PSI-Molecular Interactions (PSI-MI) (see Figure 6).

Description of the annotation process

The annotator will be shown a snippet and a pair of proteins (genes) of interest. All the mentions of these two proteins of interest detected by automated annotation tools will be highlighted in the text of the snippet in advance. For each snippet the annotator is asked to:

1. Indicate (Yes/No) whether the text implies that there is an interaction between the provided proteins.
2. Highlight the minimal and most important phrase in the text (if any) that supports his answer. This text is labeled as ”interaction”.
3. Locate and highlight the one mention of each of the entities of interest which is essential to the relation of interest. These are the mentions which, if altered, would result in a phrase which no longer conveyed the same relation. For example, in the snippet: ”Protein A is

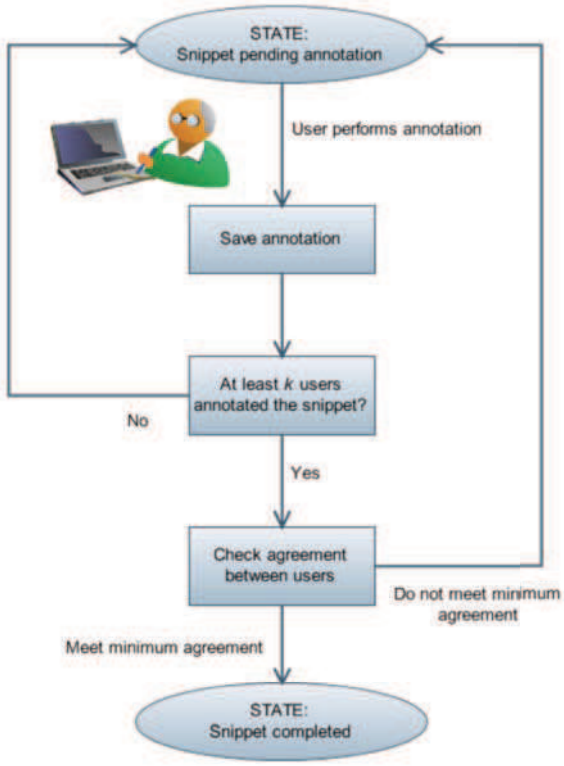

Fig. 4. State diagram representing the annotation of a snippet: from state *pending annotation* to state *annotation completed*. The annotation of one snippet is completed when it has been annotated by at least k different users and their annotations meet a minimum agreement. Once the annotation of a snippet is completed, the snippet is not served again for annotation.

found in tissue T. Protein A interacts with protein B", only the second mention of "Protein A" should be highlighted.

4. Provide a PSI-MI term for the keywords highlighted in 2, and Uniprot identifiers for the two proteins highlighted in 3. Only one mention of every gene/disease of interest should be highlighted in each snippet. The annotator should check whether the highlighted regions comply with these guidelines, and correct annotations that do not.

The resulting sets of available tags for the annotation are the following:

– PROTEIN: for protein and gene mentions (e.g. RELN, GRM8, WNT2);

– INTERACTION: minimal, most relevant phrase that supports the Yes/No decision (e.g. "binds to", "inhibits", "phosphorylates").

Detailed annotation guidelines together with annotation examples are provided at the project site. However, disagreement among experts and nomenclature specific to different scientific fields make it non-trivial to create annotation guidelines. *k* given annotations are said to meet the minimum agreement if they satisfy the following three conditions:

– The Yes/No answer is the same.
– The token sequences highlighted with labels PROTEIN completely overlap and their provided Uniprot identifiers are the same.
– The token sequences highlighted with label INTERACTION overlap (up to 1 different token with respect to the shortest highlighting is allowed between every pair of the k annotations) and their provided PSI-MI ontology matching terms are the same.

Similar agreement requirements have been previously proven to provide good quality results with acceptable agreement rates (Cano et al., 2009). Community voting is another paradigm that we are planning to include in the future.

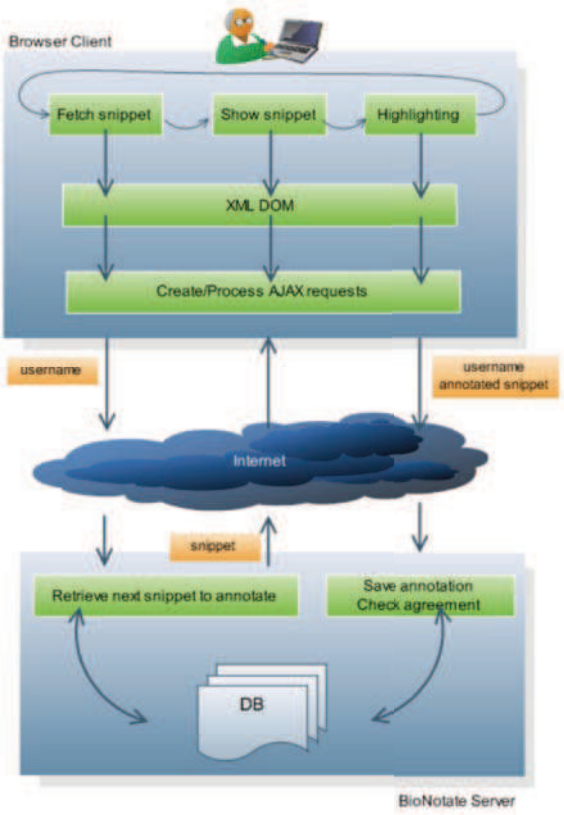

Fig. 5. Components and communications in the annotation tool. The client side iterates in the loop: Fetch the next snippet for the current annotator, present it to the user, allow him to add the annotations and save them. The process continues until the annotator closes the browser window. Each of these modules in the client side communicates with the server side. On the server side, one CGI Perl script serves snippets to the client browsers and another script attends requests for saving annotations.

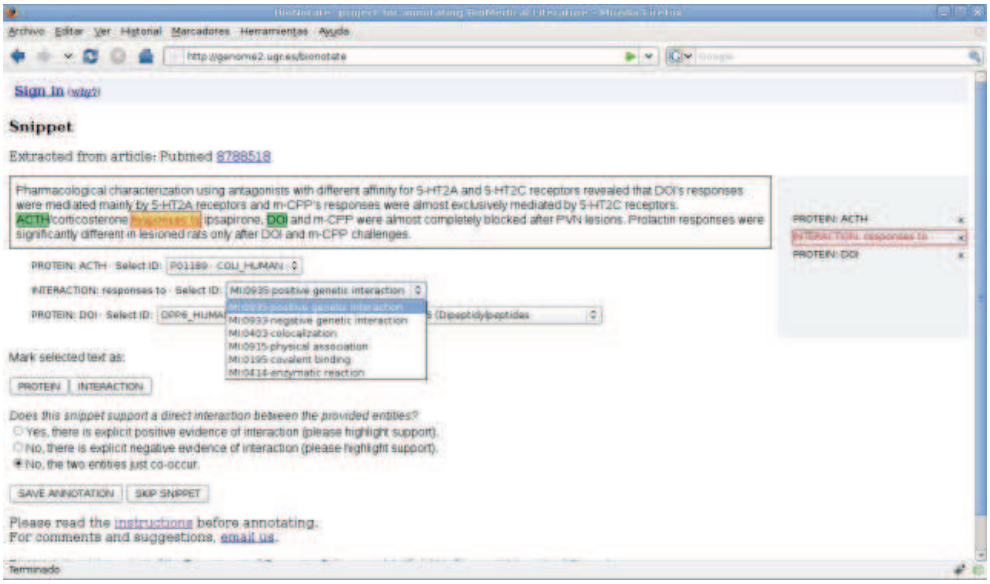

Fig. 6. Demo installation of BioNotate's annotation interface showing two proteins and an interaction keyword in a snippet. Entities may be normalized using bio-ontologies and other terminological resources. In the Figure, proteins are normalized against Uniprot, and interaction keywords are normalized against the ontology PSI-Molecular Interactions (PSI-MI).

2.4 Publication module

The overall goal of the Bionotate framework is to improve the quality and accessibility of life science research data on the Web by bringing together automatic annotation and manual curation. From the perspective of the life scientist, discovery is the final objective for the generation and use of the information generated by the system.

The dawn of the Semantic Web Era has brought a first wave of reduction of ambiguity in the Web structure, as terms and other tokens are increasingly mapped to shared identifiers for the concepts they denote. Linked Data is a style of publishing data on the Semantic Web that makes it easy to interlink, discover and consume data on the semantic web by making the connections between the datasets explicit in the form of data links. This connection should be done at the concept level rather than at the term level.

The connection at the concept level can be accomplished using RDF as a standardized data representation format, HTTP as a standardized access mechanism, and through the development of algorithms for discovering the links between data sets. Such explicit links allow scientists to navigate between data sets and discover connections they might not have been aware of previously. The standardized representation and access mechanisms allow generic tools, such as Semantic Web browsers and search engines, to be employed to access and process the data (Jentzsch et al., 2009).

But beyond the term to concept we need to go further, from concepts to statements and then to annotate these statements with context and provenance, treating richly annotated statements as nanopublications (Mons & Velterop, 2009).

In the context of Bionotate, we have made the resulting knowledge extracted after the

annotation process available as open linked data connected to the existing resources on the web like Bio2RDF or LODD. Particularly, the system allows to export the gathered information as RDF (actually using Notation3 language) or to publish it as a SPARQL endpoint. SPARQL is a query language able to retrieve and manipulate data stored in RDF.
The publication module allows to publish the RDF data to any available RDF server, such as Sesame or Joseki. The reference implementation uses Joseki (`http://www.joseki.org`), an HTTP engine that supports the SPARQL Protocol and the SPARQL RDF Query language. It is part of Jena, a Java framework for building Semantic Web applications developed at HP labs. In order to expose the data as Linked data, we use Pubby `http://www4.wiwiss.fu-berlin.de/pubby/`), a system developed by the Free University of Berlin that makes it easy to turn a SPARQL endpoint into a Linked Data server.

3. Discussion and conclusions

We have developed a framework where users can manage biological literature and related information, share their knowledge with their peers and discover hidden associations within the shared knowledge. Basically, it provides web based tools for: 1) the creation and automatic annotation of biomedical corpora 2) the distributed manual curation, annotation and normalization of extracts of text with biological facts of interest 3) the publication of curated facts in semantically enriched formats 4) the access to curated facts with a Linked Data Browser.
Our implementation is based on several open-source projects and allows disparate research groups to perform literature annotation to suit their individual research needs, while at the same time contributing to a large-scale annotation effort. There are multiple levels of integration built into the system. At one level, several annotators could collaborate on processing statements from a single corpus on their own server. At another level, multiple corpora could be created on different servers, and the resulting corpora could be integrated into a single overarching resource.
The system we propose uses BioNotate-1.0 as the collaborative annotation platform (Cano et al., 2009). BioNotate-1.0 was shown to be effective for the annotation of a small corpus on interacting genes related to autism, with averaged inter-annotator agreement rates over 75% (Cano et al., 2009). In this case, the annotation task involved providing spans of text for the interacting entities and the interaction keywords. Previous annotation efforts on gene identification and normalization reported agreement rates ranging from 91% to as low as 69% for certain contexts (Colosimo et al., 2005). BioNotate's agreement rates are thus similar to other annotation tasks, showing that the approach we propose is effective for collaborative annotation.
Our aim is to provide the community with a modular and open-source annotation platform to harness the great collaborative power of biomedical community over the Internet and create a substantially sized semantically-enriched corpus of biomedical facts of interest. Specifically, we focus on the annotation of genes and gene-disease and protein-protein relationships. However, the provided tools are flexible and can implement a variety of annotation schemas. Such versatility has already led to a wide interest in diverse user communities. One visible success story for BioNotate is the integration with the Autism Consortium research efforts (see Figure 7 or `http://bionotate.hms.harvard.edu/autism`). In that application, BioNotate is integrated with a Java-based tool kit for text processing using computational linguistics called LingPipe (Carpenter & Baldwin, 2008). Particularly, the gene-spotter class of LingPipe is used for pre-processing of autism-related abstracts. This exemplifies the flexibility

Fig. 7. Annotation interface for AutismNotate, a project for identifying autism-related genes based on BioNotate platform. `http://bionotate.hms.harvard.edu/autism`.

and modular structure of BioNotate-2.0.

In this chapter we have presented BioNotate-2.0, an open-source system which takes advantage of both manual and automated curation methods to generate reliable biological facts on protein relationships that are highly interconnected with other databases and resources. There are many directions in which this resource can be developed in both automated and manual aspect. We plan to add new Named Entity Recognition tools and incorporate new types of entities and relationships in the annotation process: diseases, drugs and SNPs. This platform is available to the community to download and integrate into diverse computational environments. In many cases, anyone with a web browser can contribute to the curation effort within a few minutes of studying the instructions.

BioNotate-2.0 extends our previous system, BioNotate-1.0, by providing several new modules for populating and automatically annotating a corpus before the current manual curation process is carried out. By assisting curators with automated annotations we expect their work to be considerably reduced in terms of time and complexity, since they have to correct previous annotations rather than create them from scratch. BioNotate-2.0 also includes a new publication module which allows curated facts to be published and shared with the community in RDF format, and conveniently accessed and browsed with the provided Linked Data front-end. All in all, we hope that the flexibility and functionality of BioNotate-2.0 will be well received by the community of bio-informatics developers and adopted for a multitude of versatile applications.

3.1 Availability

BioNotate-2.0 is available at `http://genome2.ugr.es/bionotate2/`. BioNotate-1.0 is available at `http://bionotate.sourceforge.net/`. The demo installation of BioNotate on protein-protein interactions is available at `http://genome2.ugr.es/bionotate/`.

The installation of BioNotate for assisting the Autism Consortium research efforts is available at `http://bionotate.hms.harvard.edu/autism/`.

4. Acknowledgements

C. Cano and A. Blanco are supported by the projects P08-TIC-4299 of J. A., Sevilla and TIN2009-13489 of DGICT, Madrid

5. References

Amazon's Mechanical Turk (n.d.). http://aws.amazon.com/.

Bairoch, A., Apweiler, R., Wu, C., Barker, W., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M. et al. (2005). The universal protein resource (UniProt), *Nucleic Acids Research* 33(Database Issue): D154.

Baral, C., Gonzalez, G., Gitter, A., Teegarden, C. & Zeigler, A. (2007). CBioC: beyond a prototype for collaborative annotation of molecular interactions from the literature, *Computational Systems Bioinformatics: CSB2007 Conference Proceedings, Volume 6: University of California, San Diego, USA, 13-17 August 2007*, Imperial College Pr, p. 381.

Baumgartner Jr, W., Cohen, K., Fox, L., Acquaah-Mensah, G. & Hunter, L. (2007). Manual curation is not sufficient for annotation of genomic databases, *Bioinformatics* 23(13): i41.

Belleau, F., Nolin, M., Tourigny, N., Rigault, P. & Morissette, J. (2008). Bio2RDF: Towards a mashup to build bioinformatics knowledge systems, *Journal of biomedical informatics* 41(5): 706–716.

Callison-Burch, C. (2009). Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk, *Proceedings of EMNLP 2009* .

Cano, C., Monaghan, T., Blanco, A., Wall, D. & Peshkin, L. (2009). Collaborative text-annotation resource for disease-centered relation extraction from biomedical text, *Journal of Biomedical Informatics* 42(5): 967–977.

Carlson, A., Betteridge, J., Wang, R., Hruschka Jr, E. & Mitchell, T. (2010). Coupled Semi-Supervised Learning for Information Extraction, *Proceedings of the Third ACM International Conference on Web Search and Data Mining (WSDM)*, p. 110.

Carpenter, B. & Baldwin, B. (2008). Lingpipe. http://alias-i.com/lingpipe/.

Colosimo, M., Morgan, A., Yeh, A., Colombe, J. & Hirschman, L. (2005). Data preparation and interannotator agreement: BioCreAtIvE task 1B, *BMC bioinformatics* 6(Suppl 1): S12.

Corbett, P. & Murray-Rust, P. (2006). High-throughput identification of chemistry in life science texts, *Lecture Notes in Computer Science* 4216: 107.

Cunningham, D., Maynard, D., Bontcheva, D. & Tablan, M. (2002). GATE: A framework and graphical development environment for robust NLP tools and applications, *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*. www.gate.ack.uk .

Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcantara, R., Darsow, M., Guedj, M. & Ashburner, M. (2008). ChEBI: a database and ontology for chemical entities of biological interest, *Nucleic acids research* 36(Database issue): D344.

Donmez, P., Carbonell, J. & Schneider, J. (2009). Efficiently learning the accuracy of labeling sources for selective sampling, *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM New York, NY, USA,

pp. 259–268.

Google Image Labeler (n.d.). http://imagesgoogle.com/imagelabeler/.

Hunter, L., Lu, Z., Firby, J., Baumgartner, W., Johnson, H., Ogren, P. & Cohen, K. (2008). OpenDMAP: An open source, ontology-driven concept analysis engine, with applications to capturing knowledge regarding protein transport, protein interactions and cell-type-specific gene expression, *BMC bioinformatics* 9(1): 78.

iAnnotate (n.d.)Protege Plug-in. Available at:
http://www.dbmi.columbia.edu/ cop7001/iAnnotateTab/iannotate.htm.

Jentzsch, A., Hassanzadeh, O., Bizer, C., Andersson, B. & Stephens, S. (2009). Enabling Tailored Therapeutics with Linked Data, *Proceedings of the 2nd Workshop about Linked Data on the Web*.

Jentzsch, A., Zhao, J., Hassanzadeh, O., Cheung, K., Samwald, M. & Andersson, B. (n.d.). Linking Open Drug Data.
http://triplify.org/files/challenge_2009/LODD.pdf.

Kano, Y., Baumgartner, W., McCrohon, L., Ananiadou, S., Cohen, K., Hunter, L. & Tsujii, J. (2009). U-Compare: share and compare text mining tools with UIMA, *Bioinformatics* 25(15): 1997.

Kim, J., Ohta, T., Tsujii, J. et al. (2008). Corpus annotation for mining biomedical events from literature, *BMC bioinformatics* 9(1): 10. http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/home/wiki.cgi.

Kirsch, H., Gaudan, S. & Rebholz-Schuhmann, D. (2006). Distributed modules for text annotation and IE applied to the biomedical domain, *International journal of medical informatics* 75(6): 496–500.

Knowtator (n.d.). http://knowtator.sourceforge.net.

Krallinger, M., Morgan, A., Smith, L., Leitner, F., Tanabe, L., Wilbur, J., Hirschman, L. & Valencia, A. (2008). Evaluation of text-mining systems for biology: overview of the Second BioCreative community challenge, *Genome biology* 9(Suppl 2): S1.

Krallinger, M., Valencia, A. & Hirschman, L. (2008). Linking genes to literature: text mining, information extraction, and retrieval applications for biology, *Genome biology* 9(Suppl 2): S8.

Leitner, F., Krallinger, M., Rodriguez-Penagos, C., Hakenberg, J., Plake, C., Kuo, C., Hsu, C., Tsai, R., Hung, H., Lau, W. et al. (2008). Introducing meta-services for biomedical information extraction, *Genome Biology* 9(Suppl 2): S6.

Maier, H., Dohr, S., Grote, K., O'Keeffe, S., Werner, T., de Angelis, M. & Schneider, R. (2005). LitMiner and WikiGene: identifying problem-related key players of gene regulation using publication abstracts, *Nucleic acids research* 33(Web Server Issue): W779.

Mcwilliam, H., Valentin, F., Goujon, M., Li, W., Narayanasamy, M., Martin, J., Miyar, T. & Lopez, R. (2009). Web services at the European Bioinformatics Institute-2009, *Nucleic Acids Research* 37(Web Server issue): W6.

Mons, B., Ashburner, M., Chichester, C., van Mulligen, E., Weeber, M., den Dunnen, J., van Ommen, G., Musen, M., Cockerill, M., Hermjakob, H. et al. (2008). Calling on a million minds for community annotation in WikiProteins, *Genome biology* 9(5): R89.

Mons, B. & Velterop, J. (2009). Nano-Publication in the e-science era, *Workshop on Semantic Web Applications in Scientific Discourse (SWASD 2009)*.

Raykar, V., Yu, S., Zhao, L., Jerebko, A., Florin, C., Valadez, G., Bogoni, L. & Moy, L. (2009). Supervised Learning from Multiple Experts: Whom to trust when everyone lies a bit, *Proceedings of the 26th Annual International Conference on Machine Learning*, ACM New

York, NY, USA.
Rebholz-Schuhmann, D., Arregui, M., Gaudan, S., Kirsch, H. & Jimeno, A. (2008). Text processing through Web services: calling Whatizit, *Bioinformatics* 24(2): 296.
Rebholz-Schuhmann, D., Kirsch, H. & Couto, F. (2005). Facts from text-is text mining ready to deliver?, *PLoS Biol* 3(2).
Russell, B., Torralba, A., Murphy, K. & Freeman, W. (2008). LabelMe: a database and web-based tool for image annotation, *International Journal of Computer Vision* 77(1): 157–173.
Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees, *Proceedings of International Conference on New Methods in Language Processing*, Vol. 12, Manchester, UK.
Shah, N., Jonquet, C., Chiang, A., Butte, A., Chen, R. & Musen, M. (2009). Ontology-driven indexing of public datasets for translational bioinformatics, *BMC bioinformatics* 10(Suppl 2): S1.
Snow, R., O'Connor, B., Jurafsky, D. & Ng, A. (2008). Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, pp. 254–263.
Winnenburg, R., Wachter, T., Plake, C., Doms, A. & Schroeder, M. (2008). Facts from text: can text mining help to scale-up high-quality manual curation of gene products with ontologies?, *Briefings in Bioinformatics* 9(6): 466.
WordFreak (n.d.). http://wordfreak.sourceforge.net.

Biomedical Engineering, Trends, Research and Technologies

Edited by Dr. Sylwia Olsztynska

ISBN 978-953-307-514-3
Hard cover, 644 pages
Publisher InTech
Published online 08, January, 2011
Published in print edition January, 2011

This book is addressed to scientists and professionals working in the wide area of biomedical engineering, from biochemistry and pharmacy to medicine and clinical engineering. The panorama of problems presented in this volume may be of special interest for young scientists, looking for innovative technologies and new trends in biomedical engineering.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Carlos Cano, Alberto Labarga, Armando Blanco and Leonid Peshkin (2011). Social and Semantic Web Technologies for the Text-to-Knowledge Translation Process in Biomedicine, Biomedical Engineering, Trends, Research and Technologies, Dr. Sylwia Olsztynska (Ed.), ISBN: 978-953-307-514-3, InTech, Available from: http://www.intechopen.com/books/biomedical-engineering-trends-research-and-technologies/social-and-semantic-web-technologies-for-the-text-to-knowledge-translation-process-in-biomedicine

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

© 2011 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike-3.0 License, which permits use, distribution and reproduction for non-commercial purposes, provided the original is properly cited and derivative works building on this content are distributed under the same license.