COLLABORATIVE BIOMEDICAL ANNOTATION AS A SERVICE

David Campos¹, Joni Lourenço¹, Tiago Nunes¹, Rui Vitorino², Pedro Domingues², Sérgio Matos¹, José Luís Oliveira¹

¹DETI/IEETA, University of Aveiro, Aveiro, Portugal
²Chemistry Department, University of Aveiro, Aveiro, Portugal

1 INTRODUCTION
We describe a web-based platform for biomedical text mining and collaborative curation. The web tool allows users to annotate texts with concept occurrences as well as with relations between concepts. Annotations can be performed manually or based on the results of automated concept identification and relation extraction tools. These automatic annotations may have been previously added to the documents, using one of the accepted input formats, or may be added during the annotation process, by calling a document annotation service. Users can inspect, correct or remove automatic text mining results, manually add new annotations, and export the results to standard formats.

Currently, a concept annotation service is available through the tool to automatically identify and annotate concepts in texts, but the user interface was designed to also display and allow annotation and edition of relations between entities. In the future, we expect to include a processing module for automatic extraction of this information from the raw documents.

The tool is available at http://bioinformatics.ua.pt/software/egas/ and is compatible with the most recent versions of Google Chrome, Mozilla Firefox and Safari.

2 SYSTEM DESCRIPTION
The annotation tool follows what we termed an “annotation-as-a-service” paradigm. Management of document collections, users, configurations, annotations, back-end data storage, is all done centrally, as well as the tools for document processing and text mining. In this way, a curation team can use the service, configured according to their requisites, taking advantage of a centrally managed pipeline.

The tool is based on the idea of Projects. A Project consists of a curation or document annotation task, performed on a collection of documents, by a team of (one or more) curators, and considering a pre-defined set of concept and relation types.

Three different user profiles are supported, in order to properly delegate responsibilities and streamline project management:
• Super administrator: responsible for creating new projects and associate them with project administrators;
• Project administrator: responsible for managing project characteristics, such as curators, annotation guidelines and target concepts and relations;
• Curator: responsible for manually annotating documents.

The creation of new users is provided through an e-mail invitation system. Moreover, since each project may be public or private, users only have access to public projects and to private projects to which they are associated.

To create the document collection for a project, the tool offers three import options:

• Local - Import from a local collection of documents, using one of the accepted input formats: raw text, A1 and BioC. The documents may include concept and relation annotations resulting from the application of external annotation tools;
• Remote - Import from PubMed or PubMed Central using a list of identifiers;
• Search - Import from the results of a search in PubMed or PubMed Central.

Curators can start from the raw text and add the concept and relation annotations as they review the documents, or they can start from preprocessed texts, containing automatically identified concepts and relations that they will revise. This can be achieved by importing a previously processed document collection (in A1 format, for example) or by using the integrated concept recognition tool to pre-process a document in the collection. A mapping between the annotations provided by the tool and the required semantic types can be configured. For example, the semantic types ‘Molecular Function’ and ‘Biological Process’ could be combined into a single semantic type.

For each Project, the project administrator can freely define the relevant concept and relation types, according to the requisites of the task. Each different concept and relation type is associated with a markup color, to facilitate the annotation work. A specific relation type is defined by selecting the types of the intervening concepts and assigning a name for the relation. For example, for PPIs, after defining a concept type Protein, an “interact” relation can be defined between two concepts of type Protein.

3 CURATION TASK

The proposed curation task consists in the identification and extraction of biomolecular events described over PubMed abstracts related to neuropathological disorders, including protein-protein interactions, protein expression and post-translational modifications. Curators will have access to a collection of documents that have been automatically processed by a concept recognition tool. Curators will then be asked to revise the concept recognition results, correcting any errors and adding missed concepts, and annotate relations and events mentioned in the texts, using the graphical interface. The time taken to curate the collection and the number of annotated concepts, relations and events will be recorded.

For this curation task, a corpus consisting of more than 135 thousand PubMed abstracts was created with the PubMed search:

The following terminologies will be considered for annotating this corpus:

- GO: Cell Component
- GO: Biological Process
- GO: Molecular Function
- ChEBI
- Uniprot
- KEGG Pathways
- ENZYME
- FMA
- OMIM
- Protein class
- PSI-MOD
- PRotein Ontology

## 4 Implementation

Text-processing modules, such as the concept annotation tool, were implemented in Java, the article fetching modules were also built in Java, and the web interface was developed using HTML5, CSS3, and JavaScript, in order to allow fast processing of large documents and support mobile devices. The resulting information, such as annotations and relations, is stored in a relational database. Finally, all database operations are performed using secured RESTful web-services, which allow easy integration with mobile devices, such as smartphones and tablets.

### 4.1 Automatic Concept Recognition

The concept identification modules for recognizing species, anatomical concepts, miRNAs, enzymes, chemicals, drugs, diseases, metabolic pathways, cellular components, biological processes and molecular functions are based on dictionary-matching, using highly efficient deterministic finite automaton. A database of concepts was compiled from multiple meta-sources, including UMLS, LexEBI, Jochem, and NCBI BioSystems. Genes and proteins are identified using a Conditional Random Fields tagger with entity normalization, built over Gimli (http://bioinformatics.ua.pt/software/gimli/).

The various concept recognition modules were tested on the CRAFT, AnEM and NCBI Diseases corpora, achieving f-measure results for overlap matching of 76% for genes and proteins, 95% for species, 65% for chemicals, 83% for cellular components, 92% for cells, 63% for molecular functions and biological processes, 83% for anatomical entities, and 85% for diseases.

## 5 User Interaction

### 5.1 Interface

Figure 1 shows the tool’s main user interface. The central box displays the content of the text being curated, showing the concepts and relations that have been identified. Concepts are shown as colored boxes, using the colors defined in the project configuration. Relations are shown as lines, tagged with the relation type. The colored boxes connected by the relation markup are placed under the concepts that participate in the
relation, and are colored with the same color as the respective concept, making it easy to identify the entire relation. Moreover, hovering the cursor over the relation markup also highlights the involved concepts.

The boxes on the lower right corner allow curators to select the concept and relation types they want to appear highlighted in the text.

Figure 1: Annotation interface.

Figure 2: Project and document selection.
5.2 REVIEWING ANNOTATIONS

During the curation task, concepts and relations can be added, edited or removed. Figure 3 illustrates the interactions for editing concept annotations. Hovering the mouse over an annotation shows the corresponding semantic type and, by right-clicking, a menu opens that allows removing the annotation. A new concept annotation is added by selecting a text span in the annotation window. This opens a concept type selection box for choosing the concept type for the new annotation.

Figure 3: Interface options related with concept annotations: 1) get feedback of the type of concept annotation; 2) add a new concept annotation; and 3) remove an existing concept annotation.

Figure 4 shows the user interactions for adding and removing relations between concepts annotated in the text. To add a relation, the user clicks the first concept in the relation while pressing the Alt key, and then clicks the second concept also while pressing the Alt key. Relations are considered directional, so the order in which the concepts are added to a relation is important. For example, if a relation “promotes” is defined, the order needs to be considered. As for concepts, right-clicking over an existing relation allows removing it.

Figure 4: Interface options related with relation annotations: 1) add a new relation between two concepts; and 2) remove an existing relation annotation.

5.3 PROJECT MANAGEMENT

Project management allows the project administrator to specify configurations of the annotation task, such as:

- Project: manage project information;
- Users: manage curators;
- Concepts: manage concepts to annotate;
- Relations: manage relations to annotate;
Through the project panel, the administrator can provide a brief description of the annotation task, and supply documents describing the guidelines of the annotation task. Moreover, the project administrator can also indicate which curators are allowed to annotate the documents associated with the project.

Concept management, which allows adding or removing target concept types, is presented in Figure 5. To add a new concept, users have to set the name and choose the annotation color of preference. Removing existing concepts is also simple, since the user just has to click on the cross to the right of the concept name. If there are annotations of that concept in the project, users are not allowed to remove it.

Concepts relation management, which allows adding or removing target concept relations, is presented in Figure 6. To add a new relation, the user must specify the two concept types, which are collected from the concept list, and set the type of relation. Removing relations is also performed by clicking the cross to the right of the relation. Similarly to concepts, removing relations that are annotated in the project is not possible.
5.4 IMPORTING AND EXPORTING

Figure 7 presents the process of importing documents from the client machine. The formats A1 and BioC are supported. Users can drag-and-drop or browse for files to import. Afterwards, they can select which files to import by using the check boxes. Regarding the A1 format, if no corresponding annotation file is provided, only the text of the document is imported. Otherwise, if the corresponding annotation files contain concept and relation annotations, these are imported to the database.

![Import Articles](image)

Figure 7: Import documents from the client machine, supporting both A1 and BioC formats.

Importing documents from remote resources, which supports both PubMed and PubMed Central, is presented in Figure 8. First, users must provide a list of PubMed or PubMed Central identifiers. Afterwards, the system loads the documents from the remote resource and displays them to the users, so they can choose the documents to import.
Figure 8: Import documents from remote resources, such as PubMed and PubMed Central.

Importing documents by searching remote resources, which also supports both PubMed and PubMed Central, is presented in Figure 9. First, the user must submit a query to search the remote resource, using their search infrastructures through web-services. Afterwards, the system loads the documents from the remote resource and displays them to the users, so they can choose the documents to import.
Figure 9: Import documents by searching on remote resources, such as PubMed and PubMed Central.

Figure 10 presents the interface that allows exporting the project documents to an external resource. This operation supports both A1 and BioC formats. Such feature allows users to store the generated information locally, in order to add it to a local knowledge base or for using in text mining pipelines, for instance.

Finally, Figure 11 presents the interface that allows using external automatic annotation tools that are available as web-services. For instance, users can automatically annotate a document with specific concepts, and then correct the provided annotations.
6 CONCLUSION

A tool for collaborative document annotation and curation is proposed. The tool allows teams of curators to work on a shared curation project, following a set of configurable concept and relation types. The curation task can be performed over a collection of raw text document or by reviewing automatic concept and relation annotations, obtained either with the included concept recognition tool or through external annotation tools.

Documents can be imported in raw text, A1 and BioC formats, and the final annotations may be exported to A1 and BioC formats. Apart from the local import option, it is also possible to create a document collection by importing from PubMed and PubMed Central either through a list of identifiers or by submitting a search to these services.

The tool is in active development by the Bioinformatics group at the University of Aveiro, Portugal (http://bioinformatics.ua.pt/), aiming to provide an annotation-as-a-service solution through a flexible, configurable and user-friendly environment.