Assisted Mining and Curation of Genomic Variants using Egas

Sérgio Matos[1,*], David Campos[2], Renato Pinho[1], Raquel Silva[1], Matthew Mort[3], David N. Cooper[3], José Luís Oliveira[1]

[1]IEETA/DETI, University of Aveiro, 3810-193 Aveiro, Portugal
{aleixomatos,renato.pinho,raquelsilva,jlo}@ua.pt
[2]BMD Software, Aveiro, Portugal
david.campos@bmd-software.com
[3]HGMD, Institute of Medical Genetics, Cardiff University, Heath Park, Cardiff, UK
{mortm,CooperDN}@cardiff.ac.uk

Abstract. Egas is a web-based platform for text-mining assisted literature curation, supporting the annotation and normalization of concept mentions and relations between concepts. Egas allows the definition of different curation projects with specific configuration in terms of the concepts and relations of interest for a given annotation task, as well as the ontologies used for normalizing each concept type. Annotations may be performed over raw documents or over the results of automated concept identification and relation extraction tools. Users can inspect, correct or remove automatic text mining results, manually add new annotations, and export the results to standard formats. In this paper, we describe Egas and demonstrate its use for the curation of inherited gene mutation data and associated clinical attributes, such as zygosity, penetrance or inheritance mode. The Egas tool is compatible with most recent versions of Google Chrome, Mozilla Firefox, Internet Explorer and Safari and is available at https://demo.bmd-software.com/egas/.

1 Introduction

Egas is a web-based platform for biomedical literature annotation, providing text-mining services and allowing both blind and collaborative curation work. Egas may be described as an "annotation-as-a-service" platform. Document collections, users, configurations, annotations and back-end data storage, are all managed centrally, as are the tools for document processing and text mining. This way, a curation team can use the service, configured according to the annotation guidelines, taking advantage of a centrally managed pipeline.

Egas allows the creation and configuration of different projects. Each project has its own workspace, and comprises 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 defined by the curation guidelines. Each project has a project administrator, who can add and manage users (curators) associated with the project, import documents, and define project characteristics, such as target concepts and relations, markup colors, and the normalization ontology to use for each concept type.

Projects may be defined as collaborative, in which case all users can work on the same documents and see annotations from other users, or blind, in which case curators can only see their own annotations. In the case of blind projects, the project administrator may assign different portions of the corpus to different users, with or without overlap, and visually compare their annotations. Calculation of the inter-annotator agreement is also available for blind annotation projects. Features for real-time collaboration include instant feedback of the changes introduced by other users as well as of their cursors' position, and a project chat to allow remote users to discuss details of the annotation task.

Curators may start from the raw text and add the concept and relation annotations as they review the documents, or they may start from pre-processed texts, containing automatically identified concepts and relations that they will revise. This can be achieved by importing a previously processed document collection in one of the supported formats or by using the integrated concept and/or relation extraction services to pre-process a set of documents in the collection.

2 System Description

Project management

Project management allows project administrators to specify configurations of the annotation task, such as defining project settings and annotation guidelines, managing curators, and defining the concepts and relations to annotate. Projects may be public or private, that is they are only accessible by users that have been added by the project manager, and may be defined as collaborative or blind.

The project administrator can freely define the relevant concept and relation types, according to the requirements of the task. Furthermore, to facilitate the annotation work, each different concept and relation type may be associated with a markup color. Once concept types are defined, a relation type can be defined by specifying the types of the intervening concepts. For example, to associate a mutation with a gene, a relation between concept types "Mutation" and "Gene" could be defined. Fig. 1 illustrates the definition of a concept type and respective normalization ontology through the project administration panel.

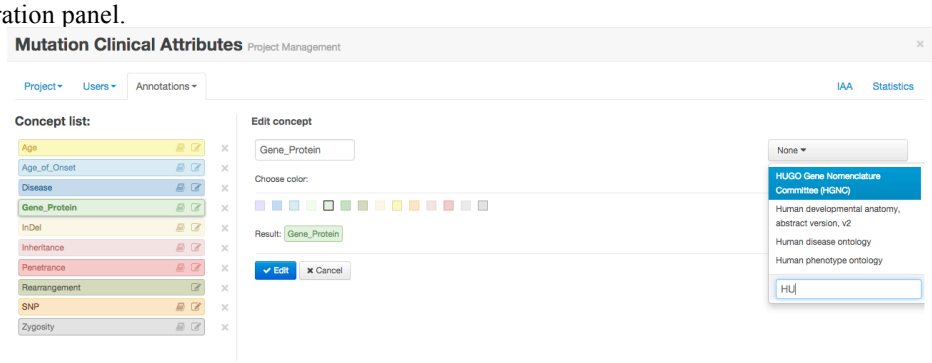

Fig. 1. Administration interface showing the definition of concept type "Gene_Protein" and link to the HUGO Gene Nomenclature Committee (HGNC) for normalization.

Blind annotation

Blind annotation projects allow project administrators to assign different portions of the corpus to different users, with a configurable overlap. Administrators may evaluate the annotation consistency by visually comparing the annotations or by calculating the inter-annotator agreement (see Results section). Fig. 2 shows the 'Users' tab in the user management panel, where the project administrator can partition the corpus and assign partitions to curators. Partitions may be defined manually or by selecting the percentage of overlap between curators, with the remaining documents being equally distributed. In the example shown, an overlap of 33% is defined and each of the three curators has been assigned 55% of the corpus. Additionally, to facilitate assigning (or removing) a partition to a curator, the colored partition boxes may be dragged over to (or from) a curator on the right-hand side of the panel.

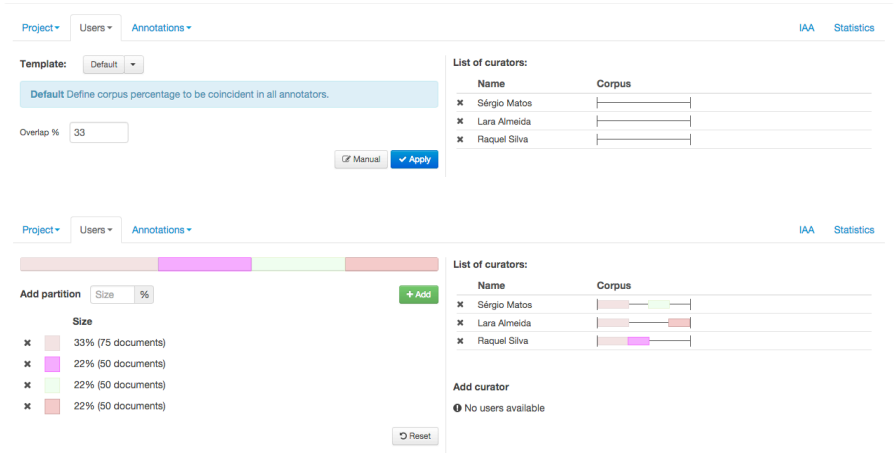

Fig. 2. Administration interface showing the definition and assignment of corpus partition to curators.

Importing and exporting documents

Documents can be imported from the client machine in RAW, A1 [1] and BioC [2] formats. The A1 and BioC formats allow importing of both the text of the documents and any concepts and relations that have been pre-annotated by external concept and relation extraction tools. Documents may also be remotely imported from both PubMed and PubMed Central, either through a list of identifiers or by submitting a search query to one of these resources. In either case, the returned documents are displayed to the users, so they can select which ones to import.

Documents may be exported in A1 or BioC formats. This feature allows users to store the generated information locally in order, for instance, to add it to a local knowledge base or for use in text mining pipelines.

Automatic concept and relation mining

In addition to importing pre-annotated documents, Egas provides an interface that allows the use of external automatic annotation tools that are available as web-services. For example, users can automatically annotate a document with specific concepts and respective relations, and then correct the provided annotations. It currently integrates an automatic service for protein-protein interactions (PPIs) annotation, providing the following annotations: a) protein concepts; b) relations between proteins (PPIs); c) relations marking equivalent protein mentions (e.g. acronyms and long forms); and d) active words that may indicate the presence of PPIs. The service is implemented on top of Neji [3], using Gimli [4] to perform machine learning-based protein name recognition. BioThesaurus [5] is used to normalize recognized names, through the application of prioritized dictionary matching [3]. Equivalent protein relations are added using a simple abbreviation resolution technique, and PPIs are recognized through a rule-based approach using dependency-parsing trees.

Annotation interface

Fig. 3 illustrates an annotated document in Egas' main user interface. The central box displays the content of the text being curated, showing the concepts and relations that have been identified. During the curation task, concepts and relations may be added, edited or removed. Additionally, concept mentions can be normalized, that is, associated with an ontology term. To create a new concept annotation, the curator selects a text span and the corresponding concept type. Subsequently, a list of possible concept identifiers is suggested along with some information to assist the curator in selecting the correct one. If the correct match is not found, the curator can refine the term in the normalization search box. Annotations can also be left un-normalized by simply clicking the check button. Concepts are shown as colored boxes, using the colors defined in the project configuration. Hovering the mouse over an annotation shows the corresponding semantic type and normalization information. Right-clicking an annotation, a menu opens that allows the removal or editing of the annotation.

Relations are shown as lines, tagged with the relation type. Colored boxes connected by the relation markup are placed under the concepts that participate in the relation, making it easy to identify the entire relation. To add a relation, the user clicks the two concepts that participate in the relation while pressing the "Alt" key. Same as for concepts, right-clicking over an existing relation allows removing or editing the relation.

Implementation

Text-processing modules, such as the concept and relation annotation services, 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 rapid 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, allowing easy integration with mobile devices, such as smartphones and tablets.

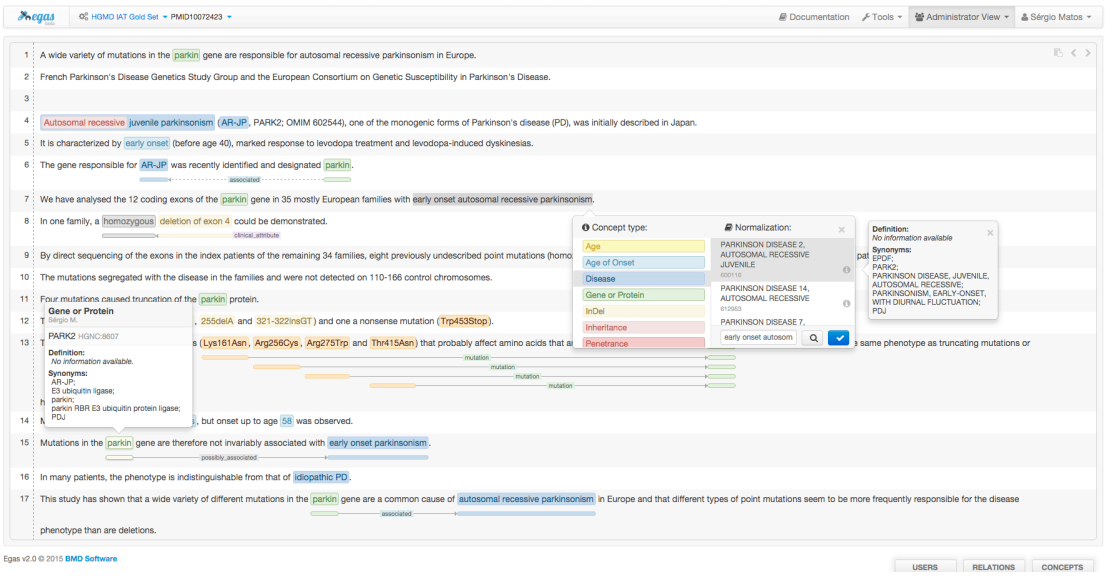

Fig. 3. Egas annotation interface.

3 Curation task

To demonstrate the use of Egas, we proposed a curation task involving the identification of human inherited gene mutations and associated clinical attributes, such as inheritance mode and penetrance, described in PubMed abstracts. The task was organized in concert with the Human Gene Mutation Database (HGMD®), a comprehensive collection of germline mutations in nuclear genes that underlie, or are associated with, human inherited disease [6]. By March 2015, this database contained over 170,000 different lesions detected in over 6,900 different genes, with new mutation entries currently accumulating at a rate exceeding 12,000 per annum. HGMD is used as a central unified disease-oriented mutation repository by human molecular geneticists, genome scientists, molecular biologists, clinicians and genetic counsellors as well as by those specializing in biopharmaceuticals, bioinformatics and personalized genomics. The public version of HGMD (http://www.hgmd.org) is freely available to registered users from academic institutions/non-profit organizations whilst the subscription version (HGMD Professional) is available to academic, clinical and commercial users under license via Qiagen Inc [http://www.biobase-international.com/product/hgmd].

For the task, curators were asked to annotate mentions of genes, genetic diseases, mutations, and clinical attributes, as well as associations between these in a corpus comprising 100 abstracts. Abstracts were selected after prioritizing, using a classifier trained with information from the documents previously used to curate information in HGMD, the 28,000 results obtained from the PubMed search:

```
genetic disease, inborn[MeSH Terms] AND (polymorphism, genetic[MeSH
Terms] OR deletion[Title/Abstract] OR substitution[Title/Abstract] OR
insertion[Title/Abstract] OR duplication[Title/Abstract] OR in-
del[Title/Abstract] OR delin[Title/Abstract] OR conver-
sion[Title/Abstract] OR translocation[Title/Abstract] OR inver-
sion[Title/Abstract]) AND hasabstract[text] AND "humans"[MeSH Terms]
AND English[lang] NOT Review[ptyp] NOT genome wide
```

In order to ground the annotations, concepts were assigned a concept identifier from an established ontology. The following terminologies were used for annotating this corpus:

- Online Mendelian Inheritance in Man (OMIM)
- HUGO Gene Nomenclature Committee (HGNC)
- Human Phenotype Ontology (HPO)
- NCBI Metathesaurus
- NCI Thesaurus

Table 1 lists the set of clinical attributes considered for this task and the Unified Medical Language System (UMLS) concept unique identifiers (CUI) used to represent the categorical values of these attributes.

Table 1. List of clinical attributes for annotating human variants.

Category	Clinical metadata	UMLS CUI
Mode of inheritance	Autosomal dominant	**C0265385**
	Autosomal recessive	**C0441748**
Zygosity	Homozygous	**C0019904**
	Heterozygous	**C0019425**
	Hemizygous	**C1881036**
Penetrance	Complete penetrance	**C1840470**
	Reduced penetrance	**C1867989**
	Variable penetrance and expressivity	**C3276568**
	Incomplete penetrance of some features	**C2750454**
	Incomplete, age-associated penetrance	**C3280136**
Age	Age of patient (years)	**C0001779**
	Age of onset (years)	**C0206132**

Seven curators were selected and were asked to annotate documents that were pre-analyzed by an automatic concept recognition tool (half of the corpus), and raw documents (the remaining corpus), in order to evaluate the added benefit of text mining-assisted curation. For this, Neji [3] was used to identify and annotate concept mentions. In these documents, curators had to revise the automatically generated annotations, correct any erroneous concept annotation and add missing ones, normalize the concept mentions, and add associations between the identified concepts. For the raw documents, curators had to add all the concept, normalization and relation annotations. The tool recorded the time taken by each curator to curate each document, as well as the number of annotated concepts and relations.

4 Results

Three curators annotated the complete corpus while two other curators followed a time-limited work plan, that is, they worked on the curation task for four hours, half of the time curating pre-annotated documents and the other half working on raw (non-annotated) documents. The remaining two curators annotated a small portion of the corpus: 13 and 9 documents.

Fig. 4 illustrates the inter-annotator agreement (IAA) panel in Egas. The IAA is calculated as the average of f-scores between each pair of curators, taking into account the documents shared by both. For this task the complete corpus was assigned to all curators (100% overlap) in order to maximize the number of shared documents. An overall IAA of 0.74 was obtained, with paired agreements varying between 0.62 and 0.95.

Table 2 shows the IAA values for each different concept type. As can be observed, there is a large spread of values, with agreement results for the less frequent concepts varying between 0.05 and 0.15, and the remaining concept types with values from 0.72 to 0.94.

We compared text-mining assisted versus non-assisted curation by evaluating the time taken to perform the task on each set of documents. As can be seen from the results on Table 3, it took in general a shorter time to curate documents that had been previously annotated by the concept recognition tool, although the results are not conclusive. One of the seven curators was not considered for these results, as the curation time per article was considerably higher than the times from the remaining curators.

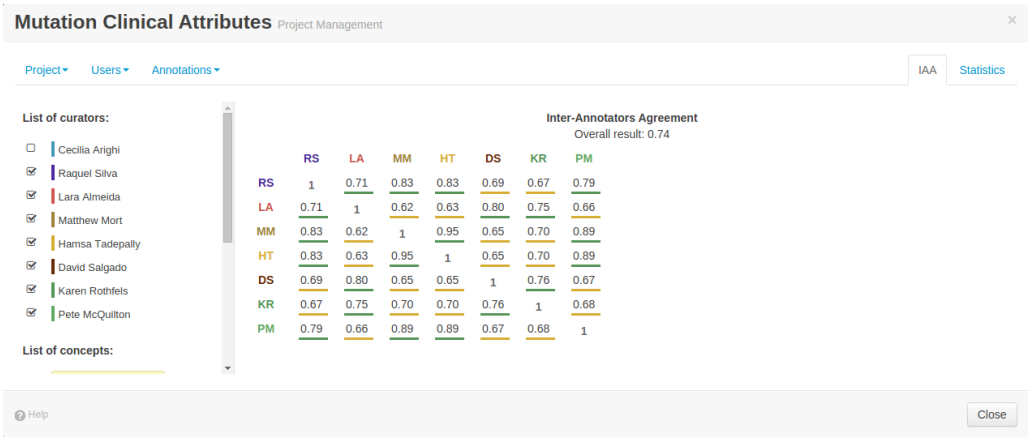

Fig. 4. Inter-annotator agreement (IAA) panel in Egas. The IAA is calculated as the f-score between each pair of curators, and the average of these values is taken as the overall result. Annotations for each concept type and each curator can be included or removed from the calculation by using the checkboxes on the left.

Table 2. IAA results for each concept type.

Concept	IAA	Concept	IAA
Age	0.05	Inheritance	0.73
Age of Onset	0.15	Penetrance	0.72
Disease	0.78	Rearrangement	0.09
Gene or Protein	0.73	SNP	0.80
InDel	0.09	Zygosity	0.94

Table 3. Comparison of text-mining assisted versus non-assisted curation. An automatic tool was used to annotate concepts but not relations. Values shown are averages for six curators. P-value obtained using a one-tailed paired t-test.

	# concepts	# relations	Time / article (s)	Time / concept (s)
TM assisted	495	138	198.6	9.6
Non-assisted	449	105	222.8	12.9
P-value	0.32	0.42	0.21	0.07

5 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 documents or by reviewing automatic concept and relation annotations, obtained either with the included concept and relation identification service 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.

Funding

This work was supported by the EU/EFPIA Innovative Medicines Initiative Joint Undertaking (EMIF grant no. 115372). SM is funded by Fundação para a Ciência e a Tecnologia – FCT under the FCT Investigator programme. DC has received support from the HemoSpec European project (EC contract number 611682). DNC and MM receive financial support from Qiagen Inc through a License Agreement with Cardiff University.

References

1. Standoff format - brat rapid annotation tool [http://brat.nlplab.org/standoff.html].
2. Comeau DC, Islamaj Doğan R, Ciccarese P, Cohen KB, Krallinger M, Leitner F, Lu Z, Peng Y, Rinaldi F, Torii M, Valencia A, Verspoor K, Wiegers TC, Wu CH, Wilbur WJ: BioC: a minimalist approach to interoperability for biomedical text processing. *Database (Oxford)* 2013, 2013:bat064.
3. Campos D, Matos S, Oliveira JL: A modular framework for biomedical concept recognition. *BMC Bioinformatics* 2013, 14.
4. Campos D, Matos S, Oliveira J: Gimli: open source and high-performance biomedical name recognition. *BMC Bioinformatics* 2013, 14:54.
5. Liu H, Hu ZZ, Zhang J, Wu C: BioThesaurus: a web-based thesaurus of protein and gene names. *Bioinformatics* 2006, 22:103–105.
6. Stenson PD, Mort M, Ball EV, Shaw K, Phillips A, Cooper DN: The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum. Genet*. 2014, 133:1-9.