Multi-Domain Collaboration for Web-Based Literature Browsing and Curation

David H. Mason, Marie-Jean Meurs, Erin McDonnell, Ingo Morgenstern, Carol Nyaga, Vahé Chahinian, Greg Butler, Adrian Tsang

Centre for Structural and Functional Genomics
Concordia University, Montréal, Canada

Abstract. We present Proxiris, a web-based tool developed at the Centre for Structural and Functional Genomics, Concordia University. Proxiris is an Open Source, easily extensible annotation system that supports teams of researchers in literature curation on the Web via a browser proxy. The most important Proxiris features are iterative annotation refinement using stored documents, Web scraping, strong search capabilities, and a team approach that supports specialized software agents. Proxiris is designed in a modular way, using a collection of Free and Open Source components best suited to each task.

1 Introduction

In the past five decades, an overwhelming volume of electronic publications has become available in many repositories, a quantity which continues to increase steadily. The English edition of Wikipedia currently contains more than four million articles, and at least 800 new pages are created each day. Within biomedical and life science literature, PubMed [27], the largest knowledge source available to biological researchers, reports more than 23 million articles indexed as of October 2013. Accessing this information is crucial for conducting research, designing experiments, and developing cutting-edge technologies.

Therefore, funding agencies and company R&D departments are supporting projects to mine this increasing volume of data. Both disciplinary and interdisciplinary research and development projects are promoted, creating new needs for knowledge workers. While disciplinary projects require browsing content within a single domain, interdisciplinary projects involve literature and data from multiple domains, and members with various backgrounds.

The development of human-machine collaboration that makes use of this huge amount of data, supporting users with various profiles and backgrounds interested in different domains, must also cope with the multiple displays of documents. Even if the way a given website displays its content is not supposed to impact the provided information, the very large number of design choices, along with the lack of universal technical standards, prevents users from benefiting from well-adapted, cross-domain mining tools that supports their research.

In the field of life science literature, a lot of research efforts are put into information extraction [12], and several tools have been developed to help researchers

and literature curators. Among them are Reflect, PubMed-EX and PubTator. Reflect [9, 23] is a free service that can be installed as a plug-in to web-browsers. Reflect tags gene, protein and small molecule names in web pages. The Reflect user can interact with the plug-in by reporting false positives or false negatives to the developer team. Reflect does not allow users to create their own categories of interest, or to collaborate as a team.

PubMed-EX [8, 28] is a browser extension that marks up PubMed bibliographic database [27] search results with additional text-mining information. PubMed-EX page mark-up includes section categorization, gene/disease name, and relation. The mark-ups of PubMed-EX highlight key terms in PubMed abstracts, and can provide additional information on these terms. PubMed-EX does not allow interaction between the user and the provided service.

Some recently developed web-based tools allow curators to create, save, and export annotations. For instance, PubTator [29] makes use of the Entrez API to search PubMed, then highlights genes, chemicals, diseases, and species in retrieved papers by running the SR4GN[30] text mining system for species recognition and gene normalization. DOMEO [16] proposes comparable functions and enables users to share their annotation sets with colleagues, groups or the community. DOMEO supports provenance records, and is compliant with text mining systems based on the UIMA framework [17]. Another tool that enables curators to interact with automatically annotated pubmed abstracts is ODIN [25]. Based on a client-server architecture, ODIN allows curators to check the annotations provided by the OntoGene pipeline [5] in abstracts, to compare them to the gold standard ones, and to modify them if needed.

Specially designed for pdf documents, Utopia Documents [24] is optimized for scientific literature. It connects the article content with online relevant data sources (references, citations, supplementary data, etc.). Three annotators for molecular information can also be invoked in Utopia: GPCRDB [2], NuclearDB [4] and Reflect.

In broader fields, multilingual annotation systems have also been actively developed in the past few years. DBpedia Spotlight [20] automatically annotates mentions of DBpedia resources in text, linking it to the Linked Open Data cloud through DBpedia. AlchemyAPI [1] and Wikimeta [11, 15] systems annotate textual content of web documents with named entities (persons, locations, organizations). AlchemyAPI provides users with important keywords, concepts and sentiments extracted from the document, while Wikimeta adds semantic links to tagged named entities, and extracts concepts from the documents.

All the aforementioned tools improve the user experience when browsing the web. However, none of them provides users with integrated, multidisciplinary, and flexible natural language processing services. Our motivations for building a new system are based on the need for user friendly systems for knowledge workers such as researchers and biocurators [18, 19], which (i) provide additional information while preserving the original content and format of browsed web pages, (ii) allow user interaction while browsing and in a collaborative space, and (iii) can be easily adapted to various research fields. Our Proxiris system is a

web-based tool developed to support users who need to mine huge volumes of web publications. Proxiris is an Open Source project designed to meet all three requirements. The next section describes services provided by Proxiris to its users. Section 3 proposes three use cases while Section 4 presents Proxiris architecture and implementation. Section 5 discusses results and future improvements.

2 Description

Proxiris is a web-based tool developed at the Centre for Structural and Functional Genomics, Concordia University. Proxiris is primarily intended to support researchers, curators and experimenters working on the Genozymes project. The main objectives of this project are discovery and development of effective fungal enzyme cocktails which can convert lignocellulose into fermentable sugars. While our current target is mainly academic literature, Proxiris works with most common types of text-focused digital content including Web pages. The flexible design of Proxiris will also support journalist team members in charge of the environmental, economic, ethical, legal, and societal dimensions of genome science in the Genozymes context. Proxiris has also been successfully trialed in consumer health and grey literature research in the PatientSense Observastory project [26, 7].

When browsing through literature, researchers mainly read fragments of interesting papers (for instance, sections describing methods or results). Biocurators study relevant papers with an exhaustive approach since their goal is translation and integration of relevant information into a database. Mining literature related to environmental impacts and public engagement, researchers need multi-domain annotation including biological entities, classical named entities, sentiments, and topics.

Proxiris eases literature mining for both categories of readers by highlighting entities of interest that are listed in an interactive sidebar. For these entities, it gives access to added content from external databases in the form of identifiers or direct links to web pages. Proxiris augmented browsing combines the annotations of a number of text mining systems, managed in hierarchies shaped by Linked Data. Currently, biological annotations are provided by our on-site developed mycoMINE [21] system as shown in Figure 2. Named entity annotations rely on DBpedia Spotlight [20] and Wikimeta [15]. A simple sentiment analysis and opinion mining service is provided by the Sentimental module [10], based on the AFINN-111 wordlist [22].

The design basis of Proxiris simply accommodates additional annotators with compatible or transformable output. Proxiris uses a team metaphor in its workflow, allowing teams to be composed of both text mining systems and human annotators. Generally, large scale text processing will be done by software programs. For example, an individual web page may have person or location annotations added by a named entity recognizer system. Then human curators can override the generated annotations as appropriate, for example, remove an annotation if it is not correct in the page context. As well, new quote-based and document level

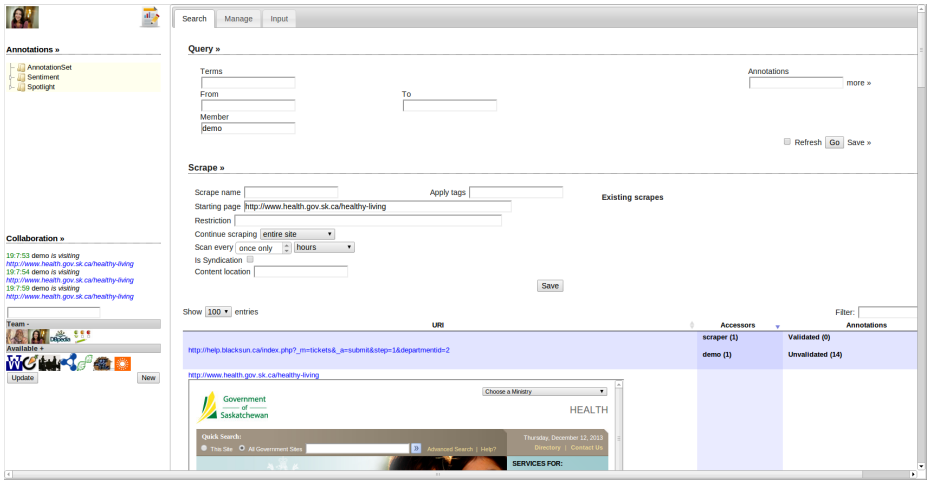

Fig. 1. Proxiris dashboard

annotations can be created. Annotations and document text are saved in a leading text database system for keyword or free text search supporting ranges and fuzzy search. Supporting queries of different conceptual levels of granularity, annotations are hierarchical. For example, the "EducationalOrganization" annotation will be considered as a general concept (or an upper ontology element) which covers "Concordia University", tagged as a named entity instance. As shown in Figure 1, the dashboard can also be used to query using free text, by contributor, or annotations. A facet search is also available, which enables quickly narrowing results to specific data dimensions, along with the ability to subscribe to saved searches, and export to reusable formats such as CSV. Export supporting Open Annotation context descriptions is planned. A scraping feature allows web scraping based on starting pages using parameters such as term incidence. Retrieved documents are automatically annotated, on an ad hoc or scheduled basis.

3 Use cases

Three examples of use cases are described in the following subsections.

3.1 Topic exploration

A team of researchers is exploring a new topic. They assemble a *search team* including the DBpedia Spotlight annotator, and human members. They also create some term sets (keywords organized as a topic) based on anticipated useful results. They upload a list of links in spreadsheet format which Proxiris automatically retrieves and annotates, and a zip file containing multiple Office and PDF documents. Then, they search and browse relevant web sites using

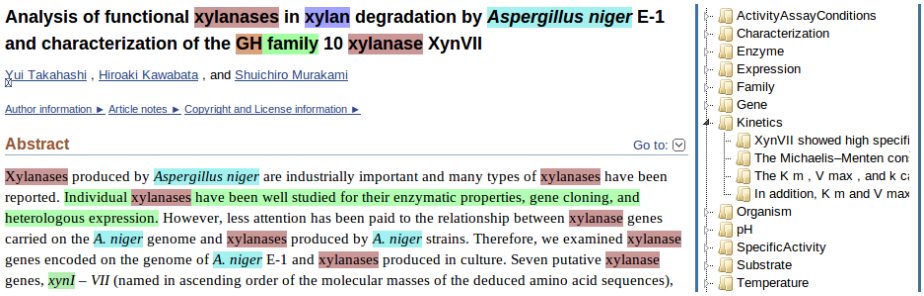

Fig. 2. Biological annotations provided through Proxiris

their *search team*. Contents are stored as they are being accessed, dynamically building a document database related to the explored topic. After the browsing sessions, the researchers use the facet browser to quickly drill down to find terms in common with relevant documents, refine their term sets and generate a new set of results against the stored content. The term sets can be edited according to observations during browsing. Once satisfied with their term sets, they launch a broader scrape that continues as long as terms are found within one link of a scraped document. The updated results appear in near real time in the search feature of the dashboard, and are immediately useful for assisted (bio)curation.

3.2 Assisted biocuration

The current version of the Proxiris prototype enables the Genozymes biocurators to curate papers which have been automatically pre-annotated by the mycoMINE text-mining system [21]. The annotations are displayed in the browsable side-bar as shown in Figure 2. With a quick look at these annotations, biocurators curating PubMed abstracts get an accurate idea of the content of the abstracts. Triage decisions (i.e. whether biocurators will read the full paper or not) are then easier and faster to make. While curating selected full papers, biocurators can also make use of the provided annotations to go directly to a given section of the paper for immediate reading, according to the annotations it contains.

3.3 Annotation system development

A software developer team works with biologists to create a new Proxiris annotator, which discovers, for instance, various expressions of Biological Value mentions, and stores found values. The software developers can work in the programming language of their choice, connecting the new annotator to Proxiris via command line or simple Web service. Since Proxiris has few technological demands and supports loosely coupled components, the integration is quick, iterative, and highly flexible. Once the annotator is made available in Proxiris, the users can see it in the annotator list under the *Manage* tab (see Fig. 3), and also in the *Available* section (see Fig. 1) from where they can designate a team using drag

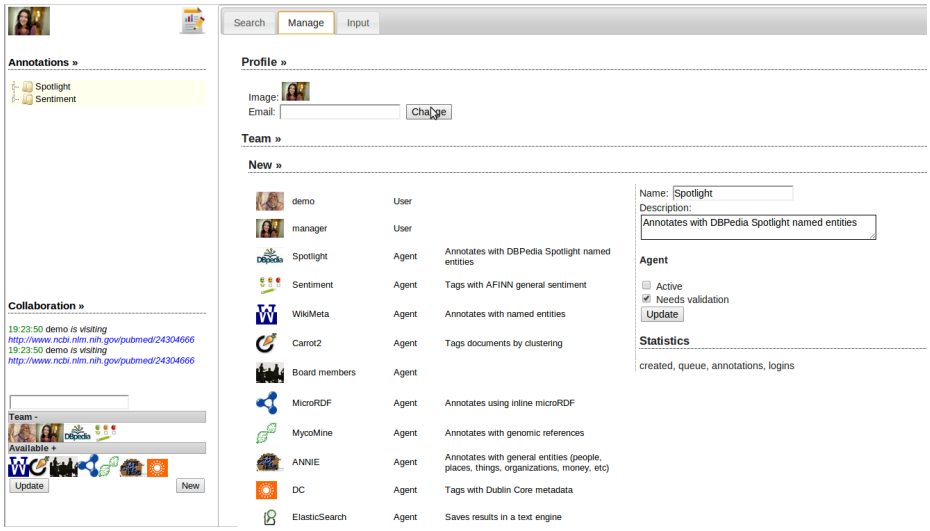

Fig. 3. Annotator list in the *Manage* tab

and drop. Results can be queried using ranges and in combination with other field values. The users can easily evaluate the annotator, and improve accuracy through feedback to the developers.

4 Implementation

Proxiris implementation is based on a proxy server approach depicted in Figure 4. Proxy servers act as intermediaries for web access. Configuring proxies is a standard setting in desktop browsers. The design of Proxiris relies on a publish-subscribe model. Rather than a standard-request response approach, actions are broadcast in connected components both for server processing and server-browser interaction. This results in a server architecture which can accommodate a large number of distributed software programs, and a client design which is simple in concept and supports highly distributed and collaborative systems with real time updates.

Proxiris is designed in a modular way, using a collection of Free and Open Source components best suited to each task. The server itself is implemented in node.js, a high-performance asynchronous Javascript server environment. Its asynchronous nature is important to support many clients and many active processing tasks without tying up system resources. Annotator components can be accessed via simple web requests (typically REST), or via command line execution via a thin wrapper. Different software programs can be designated for different domains and page sections, and requests can select which team members (software programs and humans) should be included. Since Javascript is used on the server and browser, re-use of code is enabled between server and browser in

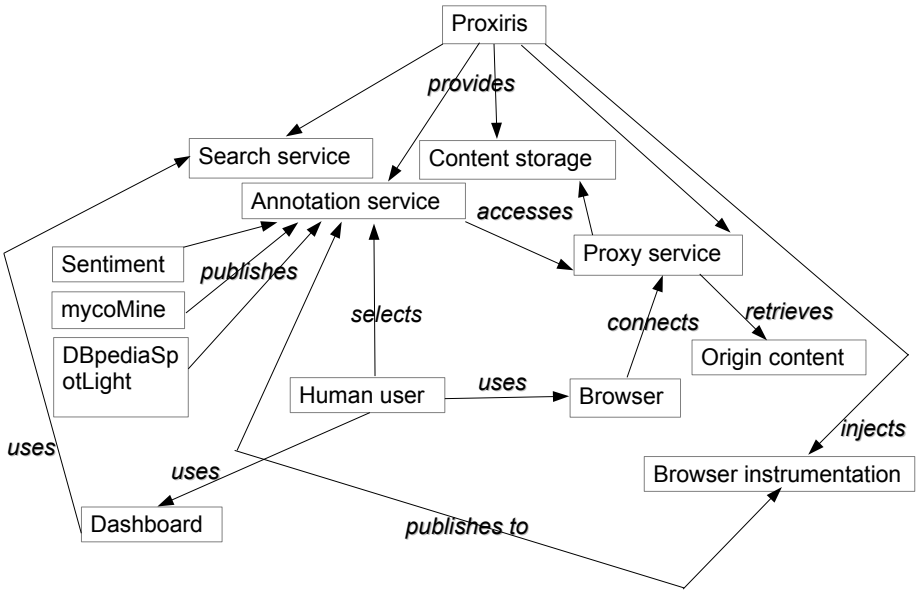

Fig. 4. Proxiris workflow

a widely distributed system. For example, the same code can be used for parsing JSON encapsulated annotations.

When accessing augmented browsing using Proxiris, the user's browser is configured to use the Proxiris proxy server. Browser web requests go to the proxy server, which retrieves and stores the requested page in ElasticSearch, a highly scalable indexing server. The web page is then returned to the user, injected with lightweight Proxiris instrumentation, composed of display code and interactive Javascript to support Proxiris operations.

The Javascript then publishes a request for annotations for that page. Proxiris retrieves any stored annotations, and executes additional relevant annotators as required, saving results to ElasticSearch. The browser code then displays these annotations in an editable format. Saved annotations are published back to the server and other browsers, where displays are immediately updated. Additional annotators can be requested without re-requesting the page using drag and drop controls.

Because pages and annotations are stored in an indexing server, results can also be accessed and shared in reformulated versions on the dashboard.

5 Conclusion

Two curators evaluated a previous prototype of Proxiris, on the triage of 114 PubMed abstracts for full paper curation. Using the tool, the time needed for triage was reduced by 21%, showing the relevance of the approach [14].

The proxy approach employed in Proxiris is flexible and easily accessible. It permits shared access and easy re-processing of content. A proxy also obviates the need to limit browser selection or to write custom browser code for all popular browsers. Not only does this approach preserve the format of the original document (including pictures, tables and embedded services), it also works around the same origin policy without the need for a browser-specific plugin, enabling greater control for instrumentation. Using a proxy enables caching of publications from selected websites which can be quickly retrieved by the user.

Due to its scalable and distributed design, Proxiris can support large teams of software agents and humans with collaborative features. Proxiris modules are under development for connecting Proxiris to Linked Data resources for the life sciences. These modules will mainly rely on the Bio2RDF [13] open source project, which currently provides the largest network of Linked Data for the Life Sciences. We also anticipate creating more complex workflows to refine content between software and human teams. Our current work is focused on user interaction with the application to document triage and curation.

Our goal is to align Proxiris with existing and emerging technologies and ontologies, in particular Open Annotation [6] and the ontologies supported by annotating software. In this version our annotations are designed to be adaptable to existing manual annotation services such as Hypothes.is [3]. We can consider mixing components of other annotating system to better support particular features, such as PDF annotation, or exchanging Open Annotation in a network of specialized systems. Proxiris components nature makes this particularly easy since JSON is a Javascript-based format, and ElasticSearch uses JSON documents to store data.

The Proxiris modules are available on GitHub at `https://github.com/TsangLab/Proxiris`. Considering a number of other open projects are focused on open annotation to better link information, we hope to encourage conversations and collaboration to contribute to this space.

Acknowledgments.
Funding for this work was provided by Genome Canada and Génome Québec.

References

1. AlchemyAPI. available at http://www.alchemyapi.com (accessed on 06 Feb. 2014)
2. GPCRSB. available at http://www.gpcr.org/7tm/ (accessed on 06 Feb. 2014)
3. Hypothes.is. available at http://hypothes.is/ (accessed on 06 Feb. 2014)
4. NucleaRDB. available at http://www.receptors.org/nucleardb/ (accessed on 06 Feb. 2014)
5. OntoGene PharmGKB. available at http://www.ontogene.org/pharmgkb/ (accessed on 06 Feb. 2014)
6. Open Annotation Data Model. available at http://www.openannotation.org/spec/core/ (accessed on 06 Feb. 2014)
7. PatientSense. available at http://patientsense.net/ (accessed on 06 Feb. 2014)
8. Pubmed-EX. available at https://sites.google.com/site/hongjiedai/projects/pubmed-ex/ (accessed on 06 Feb. 2014)
9. Reflect. available at http://reflect.ws (accessed on 06 Feb. 2014)
10. Sentimental. available at https://npmjs.org/package/Sentimental (accessed on 06 Feb. 2014)
11. Wikimeta. available at http://www.wikimeta.com (accessed on 06 Feb. 2014)
12. BioCreative IV. In: Fourth BioCreative Challenge Evaluation Workshop. Bethesda, Maryland, USA (2013)
13. Callahan, A., Cruz-Toledo, J., Ansell, P., Dumontier, M.: Bio2RDF Release 2: Improved Coverage, Interoperability and Provenance of Life Science Linked Data. In: The Semantic Web: Semantics and Big Data, pp. 200–212. Springer (2013)
14. Chahinian, V., Meurs, M.J., Mason, D.H., McDonnell, E., Morgenstern, I., Butler, G., Tsang, A.: Proxiris, an augmented browsing tool for literature curation. In: 9th International Conference on Data Integration in the Life Sciences. Montreal, QC, Canada (07/2013 2013)
15. Charton, E., Gagnon, M.: A disambiguation resource extracted from wikipedia for semantic annotation. In: LREC. pp. 3665–3671 (2012)
16. Ciccarese, P., Ocana, M., Clark, T.: DOMEO: a web-based tool for semantic annotation of online documents. Bio-Ontologies 2011 (2012)
17. Ferrucci, D., Lally, A.: UIMA: an architectural approach to unstructured information processing in the corporate research environment. Natural Language Engineering 10(3-4), 327–348 (2004)
18. Hirschman, L., Burns, G.A.C., Krallinger, M., Arighi, C., Cohen, K.B., Valencia, A., Wu, C.H., Chatr-Aryamontri, A., Dowell, K.G., Huala, E., et al.: Text mining for the biocuration workflow. Database: the journal of biological databases and curation 2012 (2012)
19. Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., Hill, D.P., Kania, R., Schaeffer, M., St Pierre, S., et al.: Big data: The future of biocuration. Nature 455(7209), 47–50 (2008)
20. Mendes, P.N., Jakob, M., Garcia-Silva, A., Bizer, C.: DBpedia Spotlight: Shedding Light on the Web of Documents. In: Proceedings of the 7th International Conference on Semantic Systems (I-Semantics) (2011)
21. Meurs, M.J., Murphy, C., Morgenstern, I., Butler, G., Powlowski, J., Tsang, A., Witte, R.: Semantic text mining support for lignocellulose research. BMC Medical Informatics and Decision Making, Vol 12 Suppl 1 (2012)
22. Nielsen, F.Å.: A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. arXiv preprint arXiv:1103.2903 (2011)

23. Pafilis, E., O'Donoghue, S.I., Jensen, L.J., Horn, H., Kuhn, M., Brown, N.P., Schneider, R.: Reflect: augmented browsing for the life scientist. Nature Biotechnology 27, 508–510 (2009)
24. Pettifer, S., McDermott, P., Marsh, J., Thorne, D., Villeger, A., Attwood, T.K.: Ceci n'est pas un hamburger: modelling and representing the scholarly article. Learned Publishing 24(3), 207–220 (2011)
25. Rinaldi, F., Clematide, S., Garten, Y., Whirl-Carrillo, M., Gong, L., Hebert, J.M., Sangkuhl, K., Thorn, C.F., Klein, T.E., Altman, R.B.: Using ODIN for a PharmGKB revalidation experiment. Database 2012 (2012)
26. Rizo, C., Deshpande, A., Seeman, N., et al.: A rapid, Web-based method for obtaining patient views on effects and side-effects of antidepressants. Journal of Affective Disorders 130(1), 290–293 (2011)
27. Sayers et al.: Database resources of the National Center for Biotechnology Information. Nucleic Acids Research 38(suppl 1), D5–D16 (2009)
28. Tsai, R., Dai, H., Lai, P., Huang, C.: PubMed-EX: a web browser extension to enhance PubMed search with text mining features. Bioinformatics 25(22), 3031–3032 (2009)
29. Wei, C.H., Harris, B.R., Li, D., Berardini, T.Z., Huala, E., Kao, H.Y., Lu, Z.: Accelerating literature curation with text-mining tools: a case study of using PubTator to curate genes in PubMed abstracts. Database 2012 (2012)
30. Wei, C.H., Kao, H.Y., Lu, Z.: SR4GN: a species recognition software tool for gene normalization. PloS one 7(6) (2012)