ODIN: a customizable literature curation tool

Fabio Rinaldi[1], Allan Peter Davis[2], Christopher Southan[3], Simon Clematide[1], Tilia Renate Ellendorff[1], Gerold Schneider[1]

[1] Institute of Computational Linguistics, University of Zurich
[2] Department of Biology, North Carolina State University, Raleigh, NC 27695-7617, USA
[3] IUPHAR Database and Guide to PHARMACOLOGY web portal. The University British Heart Foundation Centre for Cardiovascular Science. The Queen's Medical Research Institute. University of Edinburgh, Edinburgh EH16 4TJ, United Kingdom

Introduction

ODIN is a lightweight graphical interface for literature curation that can be run within a web browser. ODIN has been developed by the OntoGene group (http://www.ontogene.org/) at the University of Zurich, which specializes in biomedical text mining, in particular extraction of domain entities and their relationships from the scientific literature. The quality of their text-mining technologies has been evaluated several times through participation in community-organized competitive evaluation challenges, where OntoGene frequently obtained top-ranked results [2].

Currently ODIN is coupled with the OntoGene pipeline, which provides its text mining capabilities; however, nothing prevents ODIN from being interfaced with other text-mining services, as long as they support the same data exchange format. In order to achieve optimal performance and user satisfaction, the OntoGene team typically customizes the OntoGene pipeline and ODIN for the specific curation task. OntoGene and ODIN have already been customized for some experiments in assisted curation in collaboration with well-known databases, in particular PharmGKB, CTD and RegulonDB, which have been described in a number of journal publications [3].

As part of their participation in the triage task (task 1) of BioCreative 2012 [4], the OntoGene team produced a version of OntoGene/ODIN for the Comparative Toxicogenomics Database (CTD; http://ctdbase.org/) [6]. Customization involves using the entities of interest for the specific database, and minor adaptations of the interface to better suit the specific needs of the curators (*e.g.*, providing links from annotated entities to the web pages of the reference database).

While the OntoGene team obtained the best overall results in the official evaluation, this was concerned only with the capability of the underlying text mining system to deliver entities of

interest for CTD (genes, chemicals, and diseases) and articles ranked according to their relevance for CTD curation, so ODIN was not part of the evaluation.

In February 2013, the OntoGene team initiated a collaboration with the RegulonDB group (http://regulondb.ccg.unam.mx/) [5], which aims at improving their curation efforts through adoption of state-of-the-art text-mining techniques and advanced curation interfaces. As part of this collaboration the two groups decided on a joint participation in the interactive task (task 3) of BioCreative 2013. ODIN was, therefore, gradually modified according to suggestions provided by RegulonDB.

In August 2013, the organizers of the shared task required access to the curation interface in order to allow external curators to experiment with it. Since at this point the customization for RegulonDB was not yet completed, the OntoGene team decided to make available for this purpose the CTD version of ODIN, which in the meantime had been extended and already included several of the new features developed for RegulonDB.

As a result of these circumstances, two slightly different versions of ODIN have been evaluated in BioCreative 2013: ODIN-RegulonDB (described in a separate paper [1]) and ODIN-CTD. In the rest of this paper we will briefly describe the latter and the results of the independent evaluation.

Methods and Results

ODIN-CTD (like every version of ODIN) is available as a web application (HTML + ExtJS) and can be used from any browser. However due to incomplete support of web standards by some browser vendors, the OntoGene team recommends to use Firefox, Safari or Chrome (in this order).

Three curators invited by the task organizers were given access to ODIN-CTD (one of them chose to remain anonymous). All of them used firefox. At the first access the user is prompted to enter a login identifier that he/she can freely chose as long as it used consistently at later access of the system (the anonymous user pointed out that it was not clear that the identifier was not pre-assigned). The login identifier is stored as a cookie by the browser and therefore the user will not be prompted for it again as long as the same browser and machine are used. There is, however, an option to change the identifier if needed.

While we provide an extensive user manual, the experiments showed that this might not be a very effective way to explain how to use the tool given the limited time that curators have to perform the assigned tasks. Therefore, at a later stage in the evaluation, we added a series of screencasts that describe in a simple fashion the main functionalities of the system.

After login, the users have the option to either inspect one of the sample files provided by the system, or process an arbitrary PubMed abstract by entering the corresponding PubMed ID. The abstract will then be downloaded by the OntoGene server, processed (in this case using the CTD entity vocabulary) and delivered to the user's browser. The user can then inspect all entity annotations and candidate interactions created by OntoGene.

The annotations are visible in two formats: either as highlighted text spans in the document panel or as a table in the annotation panel (the two panels are shown side by side). A customizable color-coding shows different entity types in the documents. Hovering the mouse over an annotated span will show the type and identifier values (IDs) of the annotation (IDs depend on specific database: CTD in this case). The concept panel shows all entity IDs that have been assigned to annotations in the document. The two panels are linked, so that when a user selects an item in the concept panel, the corresponding span(s) are highlighted in the document. We summarize the experience of the users in the rest of this section.

Since a given span could have multiple IDs (because of inherent ambiguity), and the same IDs could appear in several spans in the document, there is actually a many-to-many correspondence between items in the concept table and document spans. This aspect of the system was a bit confusing for some curators.

Items in the concept table can be easily sorted according to different criteria (*e.g.,* name, ID, frequency, type, etc). No problems were identified with this facility.

If the user enters an incorrect or non-existing PubMed ID, the system tries to download it from PubMed and appears to be processing it for a while, ending in a blank screen. The users correctly pointed out that an explanatory error message would be helpful.

All entity annotations can be edited: users can modify or remove existing terms and add new terms. Deletion of a new term is a trivial procedure. Modification of an existing term by type or ID is also relatively simple. Addition of a new term is simple as long as the new term does not overlap existing term annotations. In this case it is necessary to first delete the existing annotations in order to create the new one. This procedure was a source of some confusion. We intend to clarify it in future releases of ODIN.

Additionally ODIN provides a panel containing candidate interactions suggested by the system, and ranked according to an internal confidence score. While it is relatively easy for a user to inspect the interactions, and then confirm or reject them, the system still lacks a way to add completely new interactions. This is a planned extension in a forthcoming release of ODIN.

Currently it is possible to export selected entities or interactions as a plain text file, and the curators had no difficulty in performing this task. However it would of course be desirable to be able to export them also in other common formats (*e.g.*, Excel, BEL, etc.).

In general, all curators rated their experience with ODIN as either positive or very positive.

Conclusion and future work

The experiment briefly described in this paper shows that ODIN is a user-friendly, easy-to-use web interface that can address some of the problems that curators are confronted with during their daily activities.

However, it also pointed out to some problems and shortcomings that are not due to any intrinsic limitation of the system but rather insufficient field-testing. Problems such as missing or unclear error messages can be easily solved by OntoGene programmers. Additional help menus for specific panels and tasks are already available and need only to be verified and switched on. The curators also suggested enhancements that will be considered going forward. One of these was expanding the abstract to include any MeSH terms not in the the text. Another was the capability to past in text blocks from any source.

In fact some of the feedback provided by curators during the experiments was used already to improve ODIN before the official termination of the task. A revised version was released early in September which took into account much of the feedback received up to that point. Aspects of the system that were improved include a more consistent color highlighting scheme, removal of some discrepancies in the manual, novel filters to allow focuses inspection of selected sentences. We believe that the experience was extremely positive for all parties involved and we thank the BioCreative organizers for offering us this chance to partner developers and users of biomedical text mining technologies.

References

1. Socorro Gama-Castro, Fabio Rinaldi, Alejandra López-Fuentes, Yalbi Itzel Balderas-Martínez, Simon Clematide, Tilia Renate Ellendorff, Julio Collado-Vides. Assisted curation of growth conditions that affect gene expression in *E. coli* K-12. Proceedings of BioCreative 2013, Washington, October 2013.
2. Fabio Rinaldi, Thomas Kappeler, Kaarel Kaljurand, Gerold Schneider, Manfred Klenner, Simon Clematide, Michael Hess, Jean-Marc von Allmen, Pierre Parisot, Martin Romacker, Therese Vachon (2008). OntoGene in BioCreative II. Genome Biology, 2008, 9:S13, PMC2559984
3. Fabio Rinaldi, Simon Clematide, Yael Garten, Michelle Whirl-Carrillo, Li Gong, Joan M. Hebert, Katrin Sangkuhl, Caroline F. Thorn, Teri E. Klein, Russ B. Altman. *Using ODIN for a PharmGKB revalidation experiment. The Journal of Biological Databases and Curation*, Oxford Journals, 2012, bas021; doi:10.1093/database/bas021

4. Fabio Rinaldi and Simon Clematide and Simon Hafner and Gerold Schneider and Gintare Grigonyte and Martin Romacker and Therese Vachon. Using the OntoGene pipeline for the triage task of BioCreative 2012. The Journal of Biological Databases and Curation, Oxford Journals, 2013. doi:10.1093/database/bas053
5. Salgado H, Peralta-Gil M, Gama-Castro S, Santos-Zavaleta A, Muñiz-Rascado L, García-Sotelo JS, Weiss V, Solano-Lira H, Martínez-Flores I, Medina-Rivera A, Salgado-Osorio G, Alquicira-Hernández S, Alquicira-Hernández K, López-Fuentes A, Porrón-Sotelo L, Huerta AM, Bonavides-Martínez C, Balderas-Martínez YI, Pannier L, Olvera M, Labastida A, Jiménez-Jacinto V, Vega-Alvarado L, Del Moral-Chávez V, Hernández-Alvarez A, Morett E, Collado-Vides J. RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. Nucleic Acids Res. 2013 Jan;41(Database issue):D203-13. doi: 10.1093/nar/gks1201. Epub 2012 Nov 29.
6. Davis AP, Murphy CG, Johnson R, Lay JM, Lennon-Hopkins K, Saraceni-Richards C, Sciaky D, King BL, Rosenstein MC, Wiegers TC, Mattingly CJ. The Comparative Toxicogenomics Database: update 2013. Nucleic Acids Res. 2013 Jan 1;41(D1):D1104-14.