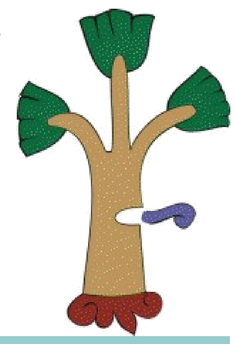
Digital assisted curation to the rescue of traditional literature curation

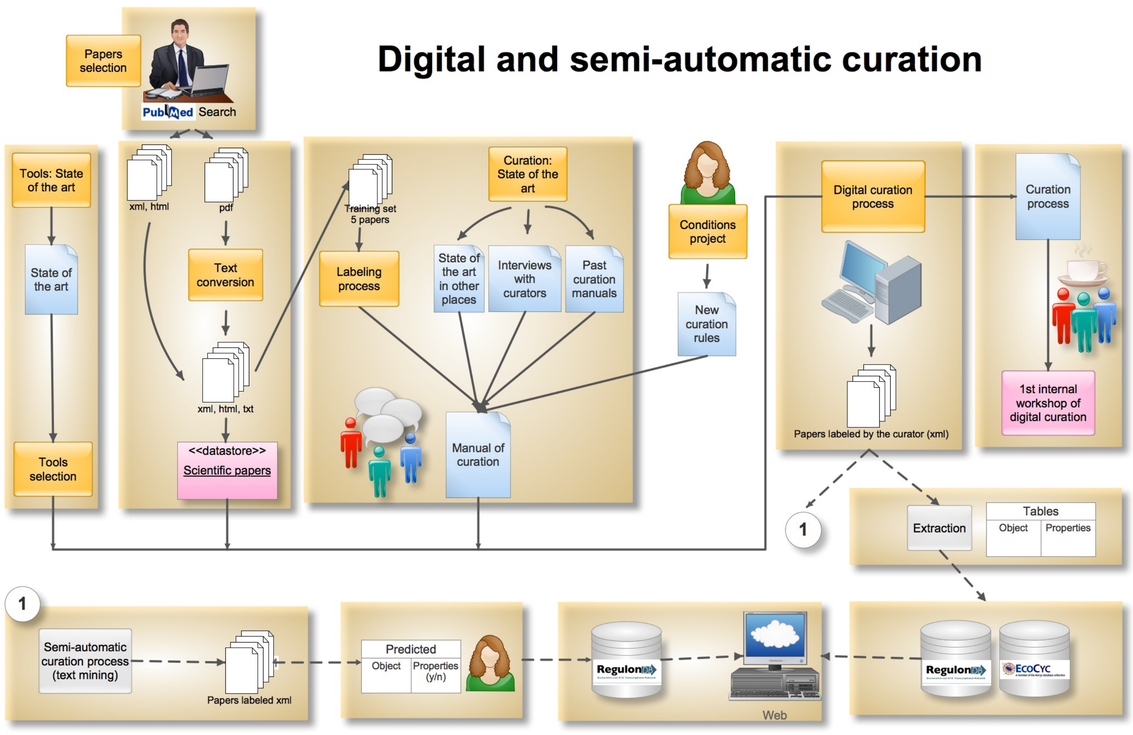
Fabio Rinaldi[1,2], Socorro Gama-Castro[1], Yalbi Itzel Balderas-Martínez[1], Oscar Lithgow[1], Hilda Solano[1], Mishael Sánchez-Pérez[1], Alejandra Lopez-Fuentes[1], Luis José Muñiz Rascado[1], Cecilia Ishida-Gutiérrez[1], Carlos-Francisco Méndez-Cruz[1], Alberto Santos-Zavaleta[1], Julio Collado-Vides[1]

1 Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Cuernavaca, Morelos 62100
2 Swiss Institute of Bioinformatics and Institute of Computational Linguistics, University of Zurich, Switzerland

- Traditional curation is not cost-effective and cannot keep up with the constantly increasing rate of publication of novel results
- Assisted curation is a promising avenue to increase the throughput of the curation process without compromising the quality of the results
- RegulonDB is testing and gradually introducing assisted curation techniques in its curation pipeline
- Digital curation: manual curation using a customized curation environment
- Semi-automated curation: a text mining system provides candidate annotations which partially guide the curation process
- Automated curation: higher throughput but lower quality, the user must be informed how the data has been generated

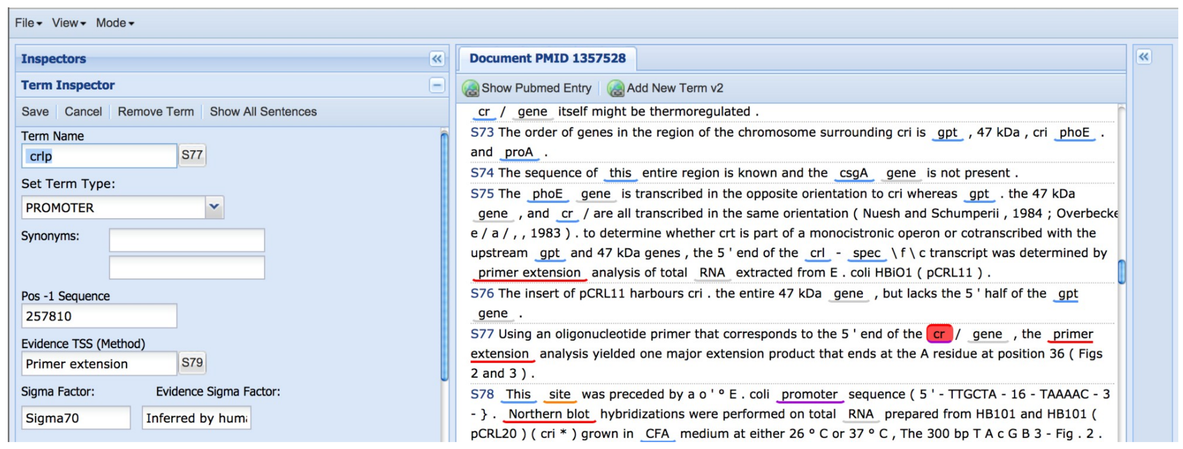


Assisted curation pipeline



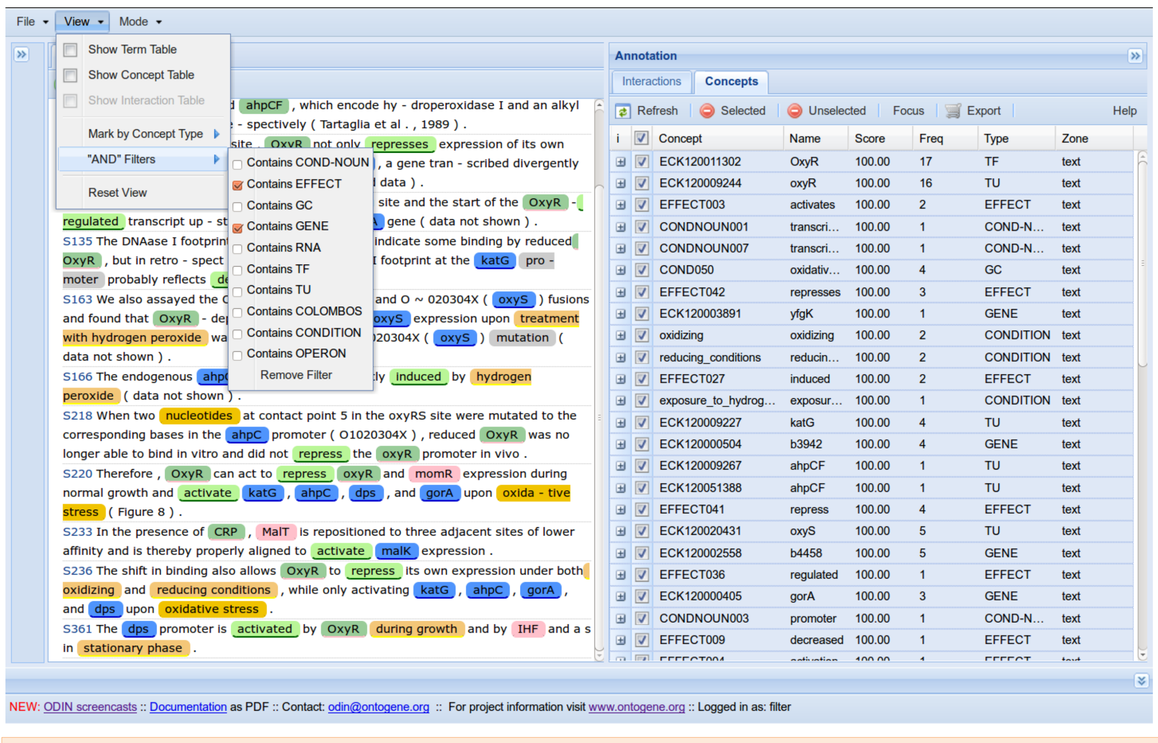
The introduction of text mining technologies in an assisted curation pipeline is expected to increase the throughput of the curation process without compromising the quality of the results.

Assisted form filling



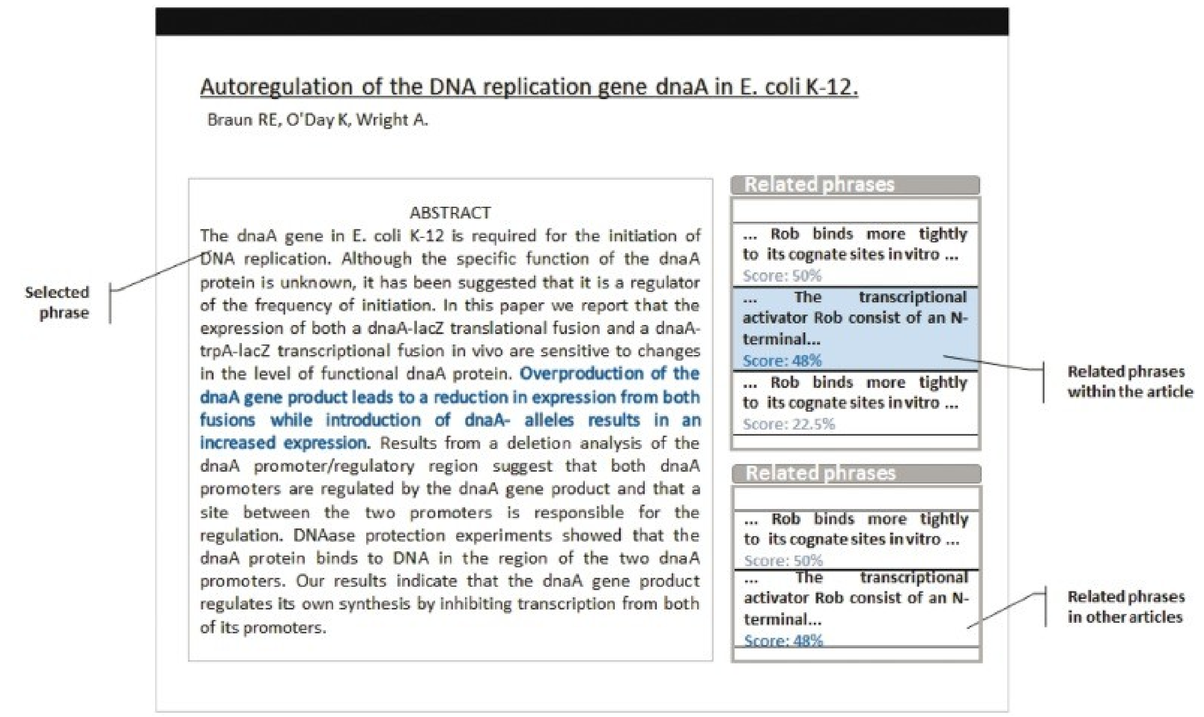
For each item to be curated (Transcription Factors, Promoters, etc.) a specific data input form has been defined. Where possible, the system tries to pre-fill some slots of the form, which the curator can accept or change if needed. The system also supports efficient two-clicks select-and-copy from the paper into the input form.

Identification of relevant domain entities (OntoGene)



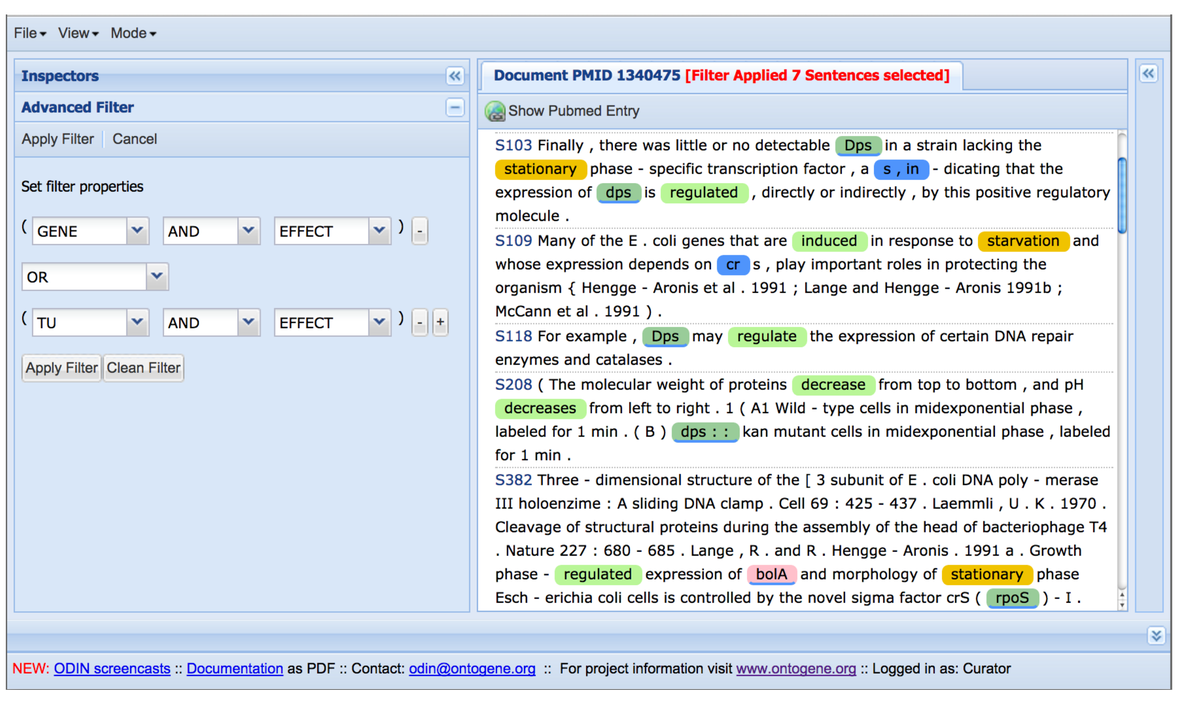
The text mining server (OntoGene) identifies all relevant entities in the document to be curated. The curator can inspect the results of the system through the ODIN interface, and, if necessary, simply remove some of them, add missing ones, or modify them.

Sentence Linking across collection



Reading whole articles sequentially is very time consuming. Based on the fact that the documents have several topics in common, we are designing a system that uses sentence similarity to link sentences about the same subject across all the articles in the set. This way, the "linear reading" is modified, allowing the expert to choose one sentence of interest and jump/navigate through other articles, guided by the current topic of interest

User-controlled selection of relevant sentences (ODIN)



The user can define filters that restrict the amount of text to be inspected, helping to identify the new items to be curated in a more efficient way.

References

F. Rinaldi, O. Lithgow, S. Gama-Castro, H. Solano, A. López-Fuentes, L. J. Muñiz-Rascado, C. Ishida-Gutiérrez, C. F. Méndez-Cruz, J. Collado-Vides; Strategies towards digital and semi-automated curation in RegulonDB. Database (Oxford), 2017 (1): bax012. doi:10.1093/database/bax012
http://www.ontogene.org/assisted-curation

Other OntoGene/BioMeXT resources

The Bio Term Hub (BHT) is a combined terminological resource created by dynamically sourcing entity names and their identifiers from reference databases. A web interface allows a user to access selected resources and download them in an uniform format.
http://www.ontogene.org/resources/termdb

The OntoGene's Biomedical Entity Recogniser (OGER) is a RESTful web service implemented on top of the BHT which allows a remote user to batch annotate a collection of documents. Recently, we have participated in a community-organized evaluation of Bio Text Mining services (BioCreative/TIPS), in which our system obtained the best results according to several of the evaluation metrics.
http://www.ontogene.org/resources/oger

Funding

Research reported in this poster is supported by the National Institutes of Health (NIH) under Award Number R01GM110597 (J.C. Vides). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.
The OntoGene/BioMeXT group at the SIB and University of Zurich is partially supported by the Swiss National Science Foundation (SNF), grants 100014-118396/1, 105315-130558/1, CR30I1_162758 (F. Rinaldi).