ISSN 2395-8618

An Approach towards Semi-automated Biomedical Literature Curation and Enrichment for a Major Biological Database

Fabio Rinaldi, Oscar Lithgow-Serrano, Alejandra López-Fuentes, Socorro Gama-Castro, Yalbi I. Balderas-Martínez, Hilda Solano-Lira, and Julio Collado-Vides

***Abstract*—As part of a large-scale biocuration project, we are developing innovative techniques to process the biomedical literature and extract information relevant to specific biological investigations. Biological experts routinely extract core information from the scientific literature using a manual process known as scientific curation. The aim of our activity is to improve the efficiency of this process by leveraging upon natural language processing technologies in a text mining system. There are two lines of investigation that we pursue: (1) finding information relevant for curation and present it in an adaptive interface, and (2) use sentence-similarity techniques to create interlinks across articles in order to allow a process of knowledge discovery.**

***Index Terms*—Text mining, natural language processing, biocuration.**

I. Introduction

THANKS to novel technological developments in genomics and the emergence of multiple high-throughput (H-T) strategies, we live in a time when studies are producing a tsunami of data. With the next generation sequencing technology, the amount of genomic data is now growing faster than the computational power [1]. In spite of the large number of databases and bioinformatics resources, a critical barrier in the field is how to accelerate access to and processing of such large amounts of information and knowledge. H-T-generated data produces large collections of individual, disconnected elements. On the other hand, pregenomic scientific papers tend to discover several interrelated elements with experiments that support more integrated perspectives, but limited to specific biological systems. The gathering in an organized and accessible database of detailed, manually curated collections of such well-studied biological systems provides the framework for an integrated understanding that is fundamental in genomics research.

RegulonDB [2] is a database, with manually curated knowledge, extracted from the literature, describing information related to transcriptional regulation in *Escherichia coli* K-12.[1] It contains biological objects such as genes, promoters, transcription factors (TFs), transcription factor binding sites (TFBSs), terminators and operons. It also contains relations of regulation among TFs and genes, promoters and operons. RegulonDB, first published in 1998, marked an effort that continues to this day for continuous and expanded curation [2], [3]. Briefly, RegulonDB facilitates access to organized information on the mechanisms of transcription initiation and it has been successful in this work; however, currently it does not facilitate access to fundamental concepts, generalizations, and knowledge of regulation of transcription initiation in *E. coli* (frequently found in reviews).

Several technical limitations have restricted the work to do so, and as a consequence, RegulonDB has only captured the knowledge contained in an estimated 10 to 15% of all sentences available in the literature of 5,244 original scientific papers supporting this database (version 8.6). This estimate is based on the number of sentences behind knowledge about TFs, TFBSs and their functions affecting promoters, the regulated TUs, and operons encoded in the database. Based on this diagnosis, we decided to improve the efficiency of biocuration process by leveraging upon natural language processing technologies in text mining systems.

Biomedical text mining can be used to partially automate the process of biomedical literature curation by using sophisticated algorithms for discovering biomedical entities together with interaction and events in which they participate. A successful biomedical text mining system is typically based on a pipeline which first discovers entities of interest in the text of a scientific article and subsequently looks for interactions between them. As described above, finding the unique database identifiers of the entities in focus is an important step in this process. Which database identifiers are used in this process depends largely on the application for which a text mining system is built, or in other words, the database for which the system is designed to extract information.

In order to accomplish the goal of digitally-assisted curation, we are working simultaneously on two main lines of research: (1) finding information relevant for curation and present

Manuscript received on June 18, 2015, accepted for publication on August 12, 2015, published on October 15, 2015.

Fabio Rinaldi (corresponding author) is with the Institute of Computational Linguistics, University of Zurich, Switzerland (e-mail: fabio.rinaldi@uzh.ch).

Oscar Lithgow, Alejandra López-Fuentes, Socorro Gama-Castro, Yalbi I. Balderas-Martínez, Hilda Solano-Lira, and Julio Collado-Vides are with the Computational Genomics Program, Center for Genomic Sciences, Universidad Nacional Autónoma de México Cuernavaca, Morelos, Mexico.

[1] http://regulondb.ccg.unam.mx/

ISSN 2395-8618

it in an adaptive interface, and (2) use sentence-similarity techniques to create interlinks across articles in order to allow a process of knowledge discovery. These steps are described in detail in this paper, together with the preliminary design of the integrated system.

II. The OntoGene Text Mining System

We use an existing biomedical text mining system (OntoGene) in order to process a collection of documents relevant to the curation purposes of RegulonDB. Ontogene offers a powerful and flexible entity recognition module based on a dictionary lookup approach allowing for some variants and a post-annotation filtering module based on maximum entropy techniques. The aim of this section is to provide a brief description of the OntoGene Text Mining pipeline, which in the RegulonDB application is used to provide the basic preprocessing capabilities as well as for the identification and normalization of domain entities. For additional details the reader is invited to consult some of the related publications [4], [5], [6]. The publications used in the experiments described in this paper were either downloaded from PubMed Central (an open access repository of biomedical literatures) in XML format, when available, or converted from PDF using the PDFlib Text Extraction Toolkit[2].

The full sequence of processing steps offered by the OntoGene pipeline is the following, however the last three steps where not used for the applications described in this paper.

- Transformation of input format into the OntoGene-specific XML format
- Zoning: partition of the document in sections such as title, abstract, references, etc.
- Sentence splitting
- Tokenization
- Part of Speech Tagging
- Lemmatization
- Stemming
- Named Entity Recognition
- Chunking
- Dependency Parsing
- Detection of Interactions

The named entity recognition step is based on a large internal database of domain terms, sourced from life science databases, and customizable by the end user of the application. Several life science databases can be considered a rich terminological resources, since they provide not only concept descriptions, but also the terms that are actually used by researchers to refer to a particular concept. The OntoGene database can be automatically generated from a subset of such resources, taking from them the preferred names and synomyms of user-selected term categories. As term names are stored together with their original database identifiers, it is always possible to retrieve all information associated with a given concept. The OntoGene system takes automatically into account a number of possible minor variants of the terms (e.g. hyphen replaced by space), thus increasing the flexibility of term recognition. The annotation step automatically adds to the XML representation of the document a list of possible database identifiers for each term where a match was found.

The OntoGene pipeline is also optionally capable of generating candidate interactions among the detected domain entities. The approach is based on a preliminary generation of potential interactions by combinatorial pairwise combinations of entities in a given text span (typically a paragraph). In order to balance the low precision of such an approach, a machine-learning based reranking is performed after the initial combination. The reranking takes into account lexical features, such as word stems and PoS tags, syntactic features, such as dependency parses, and global distribution features, such as relative frequencies of terms in the specific paper compared to the average distribution in the whole collection.

The system is trained using a distant learning approach taking a reference database as provider of the "ground truth". For example, in a recent industrial application aimed at large-scale detection of protein-protein interactions from the literature, the BioGrid database was used as a reference. BioGrid is a very large scale manually curated resource about protein interactions and genetic interactions. In the specific application described in this paper, the RegulonDB database itself is taken as the ground truth reference.

In practice the OntoGene system uses a supervised machine learning method (based on a maximum entropy classifier) in order to compute a probability of a term/concept pair to be part of a relationship in the reference database. This probability score can be used to weed out false positive entities erroneously provided by the high-recall dictionary-based annotation system. Additionally, given that each annotated term can be associated to several identifiers in the reference database, the most likely association can be selected, thus leading to the disambiguation of the possible meanings of the term.

As a second step, the probabilities of the two concepts that participate in a candidate interactions are combined using their harmonic mean, producing a score for the relationship. These scores allow a ranking of the candidate relationships, and therefore either an automated selection based on a threeshold, or a manual selection based on the inspection of the most likely candidates by domain experts.

The end result of the processing steps described above is an XML version of the original document enriched with information coming from the various modules. In particular, the information of relevance for the end users is the annotations of the domain entities, and (optionally) a set of candidate interactions. This rich XML format can be browsed through a specifically designed interface called OntoGene Document Inspector (ODIN) [7], [5], [8].

For the particular application discussed in this paper, the first step of processing consists in annotation of the

[2]http://www.pdflib.com/?id=12

ISSN 2395-8618

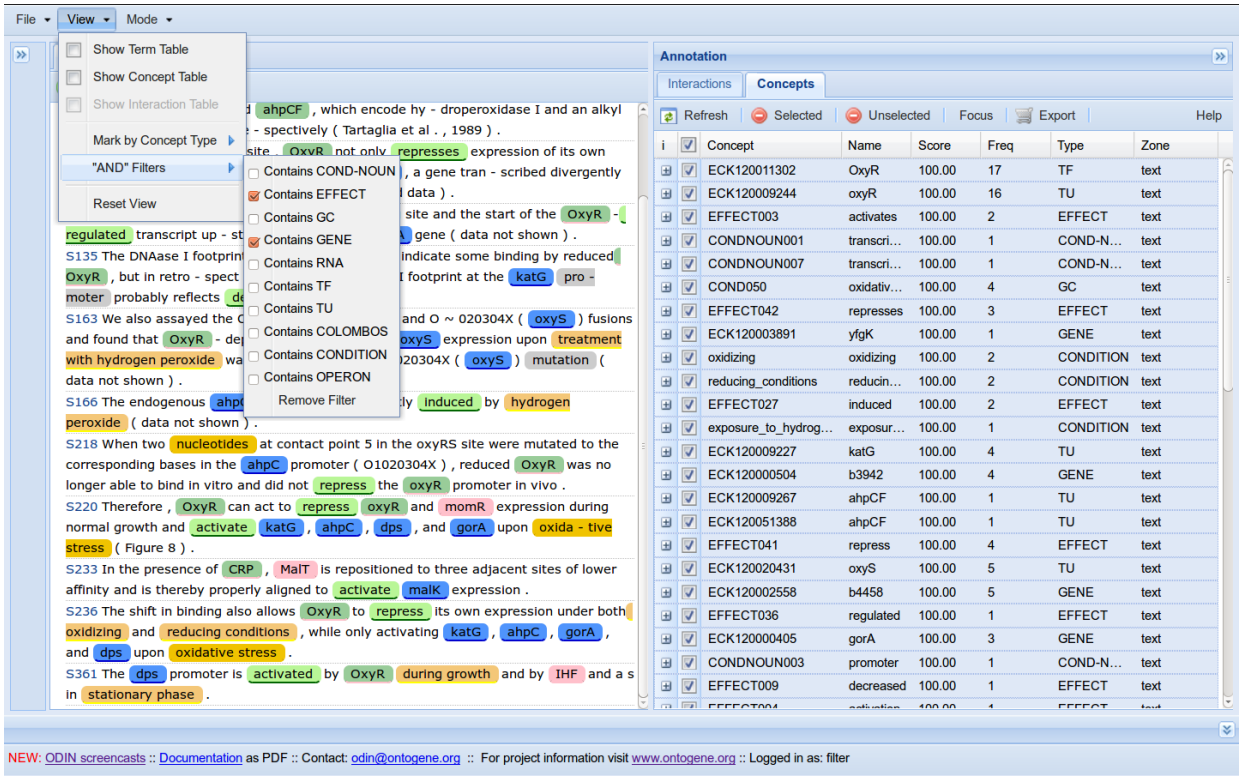

Fig. 1. ODIN customized for RegulonDB

domain entities which express the core knowledge of interest for RegulonDB curators: genes, transcription factors, growth conditions, etc. The list of genes of *E. coli* is derived from GenBank. In a recent experiment [9], we analysed a small set of articles relevant for the topic of genetic response to oxydative stress. All articles were annotated by the OntoGene pipeline and inspected by RegulonDB curators through the ODIN interface. In particular, ODIN offers a functionality called "sentence filteres", which allow the curators to select sentences which satisfy a simple logical condition defined by the user. Typically such a filter is defined by the presence in the same sentence of entities of two predefined types (e.g. "gene" and "effect"). Such condition is defined in order to locate sentences which are likely to contain the information that the curators need to extract from the documents. The experiment mentioned above showed that the curators could identify the desired items by reading fragments of the papers equivalent to only 11% of the total material that they would have had to read if they worked without the support of the assisted curation tool.

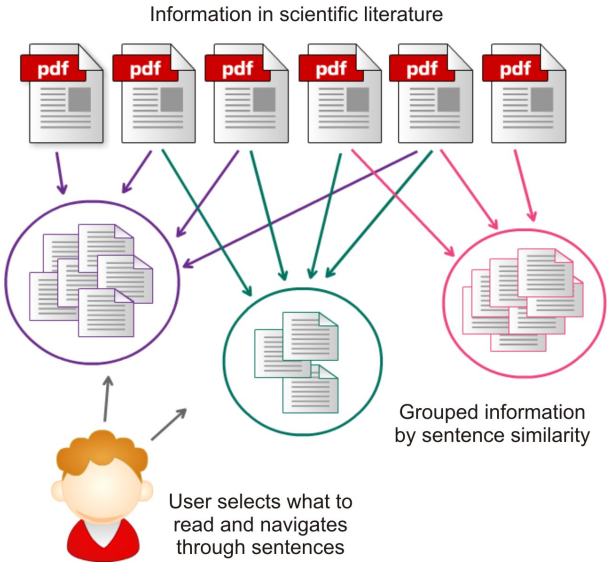

Fig. 2. Exemplification of the process of semantic linking

III. Linking Sentences Across Articles Based on their Similarity

In the context of Natural Language Processing (NLP), Semantic Similarity between two texts is the task of evaluate the likeness of their meaning. It is a recurrent and important approach to address the natural language understanding issue in tasks such as summarization, paraphrasing, QA-Systems, conversational agents, etc.

Common approaches to its computation are based on the combined use of syntactic and semantic features of the texts being compared, such as: the lexical class, ontological similarity, syntactic category, etc. These features correspond to

ISSN 2395-8618

knowledge of experts encoded in ontologies, lexical resources and syntactic parsers, among others.

Other kinds of features, also used, are those extracted from the automated analysis and statistical interpretation of large amount of texts (corpora) [10]. This distributional perspective relies on the hypothesis that the meaning of word and texts can be inferred by the context where it is used or, put in other words, that text occurring in similar context have similar meaning [11].

When doing biocuration, the experts read one by one a set of topic-related articles to annotate relevant information. This technique works well in the sense that relevant information is identified but having to read the whole articles sequentially is very time consuming. So based on the fact that the documents have several topics in common, we designed a system that uses sentence similarity to link sentences about the same subject across all the articles in the set. For instance, complex sentences (like examples a, b and c) will be related, since they are about the same topic:

a. *The oxidized form of Oxy is a transcriptional activator of a multitude of genes that assist in protecting the cell from oxidative damage* [12].

b. *Activated Oxy then induces transcription of a set of antioxidant genes, including katG (hydroperoxidase I), ahpCF (alkylhydroperoxidase), dps (a nonspecific DNA binding protein), gorA (glutathione reductase), grxA (glutaredoxin I), and oxyS (a regulatory RNA)* [13].

c. *A hallmark of the E. coli response to hydrogen peroxide is the rapid and strong induction of a set of Oxy -regulated genes, including dps, katG, grxA, ahpCF, and trxC* [14].

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. The system is formed by 4 components which are:

- A user-friendly web interface
- A web service layer
- A relational database
- A semantic similarity engine

The 3 former components are orchestrated by RestFul Web services which respond to the user's petition to search on the database, either key words or semantic similar sentences. On the other hand, the semantic similarity engine is an off-line process that is in charge of processing the articles and registers the results, sentences and relations, in the database.

The semantic similarity engine was built as a micro-framework where each involved step is an interface that can be re-implemented or extended in order to test different strategies. Besides, the implementation classes and external libraries are dynamically loaded from a configuration file; this alleviates the need of recompiling the framework in order to test different strategies.

In the current version, the processing sequence is the following:

- Apply Part-Of-Speech (POS) tagging using the Stanford POS tagger [15].
- Apply stemming using the snowball stemmer [16].
- Apply a rule based sentence selector, i.e. regular expressions that are based on the words' POS-tags. The motivation for this is to restrict the set of candidates and to keep those more informative. For example one of the rules is to select those sentences that contains the pattern "Noun-Verb-Noun with other optional tags like determinants, adjectives, etc.":
  `[IN] [DT] [JJ] NN [RB] VB [IN] [DT] [JJ] NN`
- Create a multidimensional vector representation of the sentence. Selected sentences are represented by vectors with as many dimensions as the length of the global vocabulary. Each vector dimension embodies a word of the vocabulary and the dimension's magnitude is the word frequency in the sentence. It is worth noting that counts are not normalized.
- Measure the similarity between each two sentences using the cosine similarity. This measure is particularly convenient because the length of the vectors is irrelevant in its computation. Moreover, it provides a confined similarity measure that ranges between -1 and 1, being 0 when the vectors are orthogonal (i.e. not related) and 1 when both vectors are identical. The similarity is computed using Efficient Java Matrix Library [17].

The novel web interface (see figure 3), currently in the implementation phase, provides to the user the means to search key words on one or several articles. Once the results are listed and the user decide to enter to an specific article, the article's sentences (content) are displayed and those which have semantic relations with other sentences, either in the same article or in others, are decorated with hyperlinks. When the user selects a hyperlink the related sentences that are located inside the current article are highlighted, and those which resides in others are listed in a panel along with the article name and the corpus to which it belongs to. In that way the user is provided with an instrument to navigate across different articles and corpus through the following up of a specific idea.

IV. Evaluation

There have been several separate evaluations of the modules described in this paper. As a way to verify the quality of the core text mining functionalities, the underlying text mining pipeline has been used to perform several tasks which have been formally evaluated within the context of community-organized text mining evaluations campaigns ("shared tasks"), such as BioCreative [18]. Some of most interesting results include best results in the detection of protein-protein interactions in BioCreative 2009 [19], top-ranked results in several tasks of BioCreative 2010 [20], best results in the triage task of BioCreative 2012 [21].

The usefulness of ODIN as a curation tool, leveraging upon the results of the text mining system, has been demonstrated through an experiment aimed at making more efficient the

ISSN 2395-8618

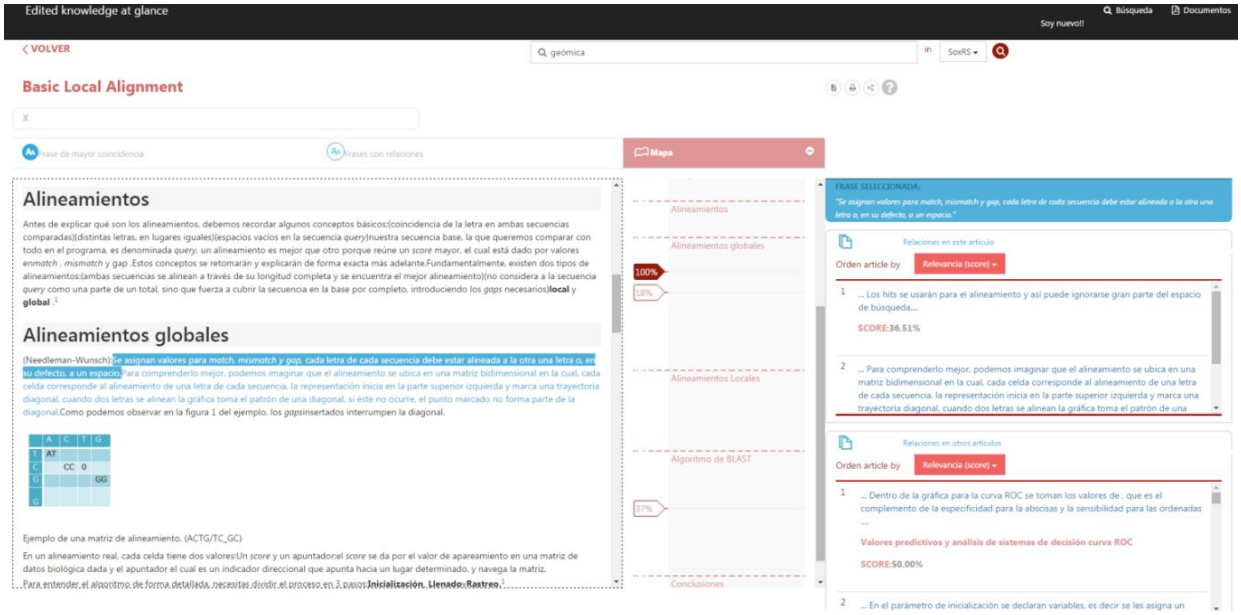

Fig. 3. Preliminary version of the interface of the system

identification by the curators of specific types of entities (*"growth conditions"*) which had not been curated before, and were therefore not part of the reference database. An initial set of articles, likely to be relevant for the given task, was identifies through conventional IR techniques. The curators used ODIN filters to restrict their view of the selected articles to the set of sentences satisfying a given logical condition (e.g. containing an entity of type *"transcription factor"* and an entity of type *"effect"*). The manual analysis of the selected sentences allowed them to identify the missing information in 75% of the cases, but having to inspect only about 10% of the articles, thus providing a considerable improvement in efficiency, as described in section II.

More recently, we tested the sentence similarity component using three set of scientific articles:

1) 42 articles of SoxRS: oxidative stress in *Escherechia coli* K12
2) 35 articles of *Salmonella typhimurium* pathogenicity island SP1
3) 10 articles of role of EZH2 gene in cancer

We had six domain experts that worked with these sets (two per set). The goal for the test exercise was to read the articles looking for specific information, as it's done in the curation process. Then to extract and save all the information they could find within 2 hours of reading the literature. One of the experts from each set had access to the system, the other didn't and used the PDF files instead. The users with access to the system were able to review more articles thus they extracted more sentences in total with similar information. The users with the files couldn't finish all articles but they extracted more sentences per reviewed article.

The general opinion form the experts was that the system could be very powerful if the similarity is improved to detect more topic-related sentences and also made some suggestions to the web interface in order to be more intuitive. The system has proved to be useful for the curation process, we are now working to add more capabilities, improve the interface design by implementing User eXperience (UX) techniques, and integrate all components in a single unified system. Figure 4 shows a schematic representation of the system that we are now in the process of implementing.

V. Conclusion

The work described in this paper takes place in the context of a large-scale NIH-funded[3] four-year project, started in 2015, which has as one of its goals the implementation of a process of digital assisted curation, which involves the integration of advanced text mining techniques within the curation pipeline of a biomedical database. Human experts will be able to leverage upon the best results of text mining technologies in order to improve the effectiveness of the curation process without sacrificing its quality.

This paper describes some of the components that will be used in order to reach that goal: an advanced text mining pipeline for entity extraction and relation detection, a customizable flexible user-interface, and a way to interlink information across several papers. The new system will constitute a very powerful curation tool that will allow semiautomatic data annotation, and a new way of knowledge discovery reducing reading time without affecting understanding.

[3]National Institute of Health, US

ISSN 2395-8618

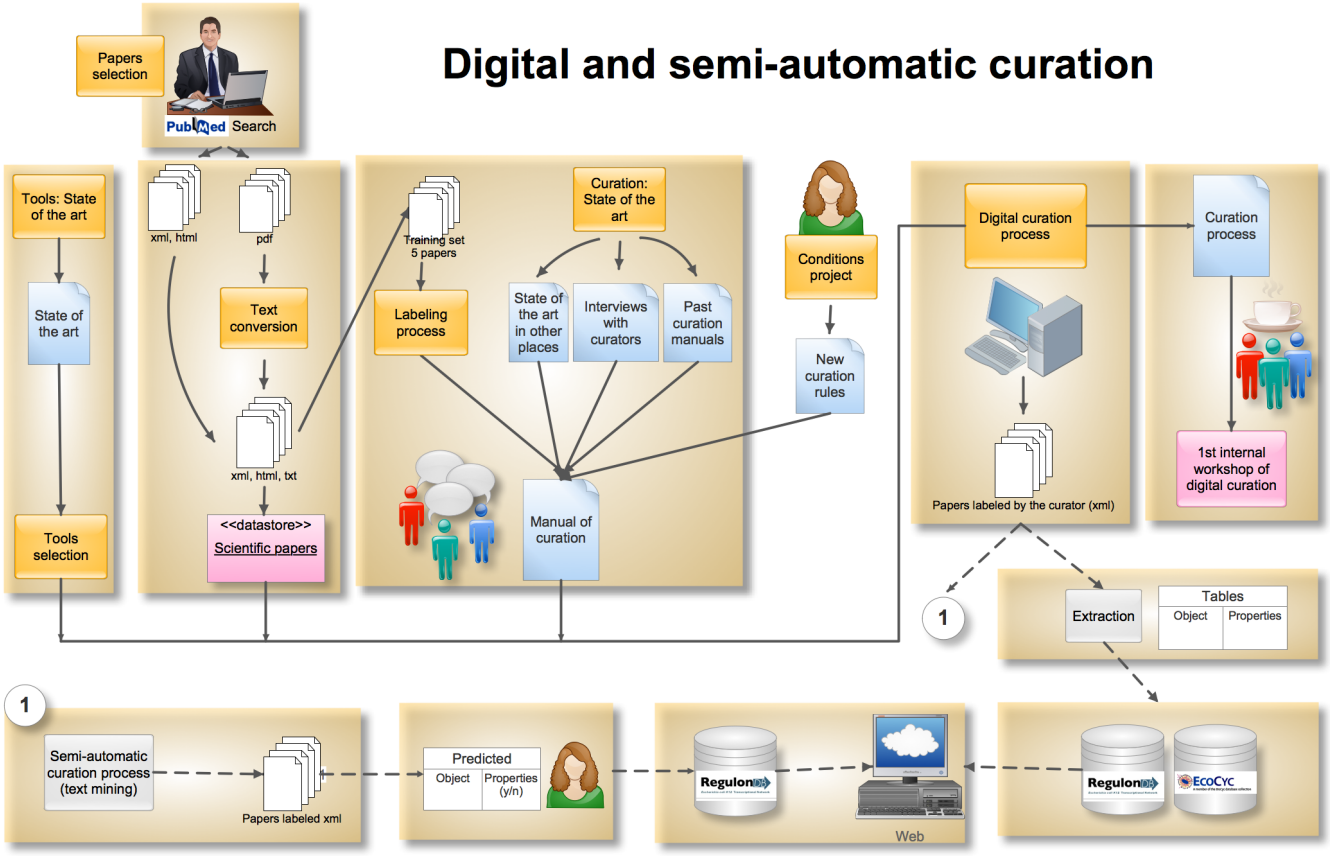

Fig. 4. Diagram illustrating the process of digital curation

ACKNOWLEDGMENT

The development of OntoGene/ODIN has been supported by the Swiss National Science Foundation (grant 105315-130558/1, PI: Fabio Rinaldi) and by the Data Science Group at Hoffmann-La Roche, Basel, Switzerland. The development of RegulonDB is supported by NIH grant 1R01GM110597 to Julio Collado Vides.

REFERENCES

[1] L. D. Stein, "The case for cloud computing in genome informatics," *Genome Biology*, vol. 11, no. 5, p. 207, May 2010. [Online]. Available: http://dx.doi.org/10.1186/gb-2010-11-5-207

[2] A. M. Huerta, H. Salgado, D. Thieffry, and J. Collado-Vides, "RegulonDB: A database on transcriptional regulation in escherichia coli," *Nucleic Acids Research*, vol. 26, no. 1, pp. 55–59, 1998. [Online]. Available: http://dx.doi.org/10.1093/nar/26.1.55

[3] H. Salgado, M. Peralta-Gil, S. Gama-Castro, A. Santos-Zavaleta, L. Muniz-Rascado, J. S. Garcia-Sotelo, V. Weiss, H. Solano-Lira, I. Martinez-Flores, A. Medina-Rivera, G. Salgado-Osorio, S. Alquicira-Hernandez, K. Alquicira-Hernandez, A. Lopez-Fuentes, L. Porron-Sotelo, A. M. Huerta, C. Bonavides-Martinez, Y. I. Balderas-Martinez, L. Pannier, M. Olvera, A. Labastida, V. Jimenez-Jacinto, L. Vega-Alvarado, V. D. Moral-Chavez, A. Hernandez-Alvarez, E. Morett, and J. Collado-Vides, "RegulonDB v8.0: Omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more," *Nucleic Acids Res.*, vol. 41, no. D1, pp. D203–D213, 2013.

[4] F. Rinaldi, "The ontogene system: An advanced information extraction application for biological literature," *EMBnet.journal*, vol. 18, no. Suppl B, pp. 47–49, 2012. [Online]. Available: http://journal.embnet.org/index.php/embnetjournal/article/view/546/755

[5] F. Rinaldi, S. Clematide, Y. Garten, M. Whirl-Carrillo, L. Gong, J. M. Hebert, K. Sangkuhl, C. F. Thorn, T. E. Klein, and R. B. Altman, "Using ODIN for a PharmGKB re-validation experiment," *Database: The Journal of Biological Databases and Curation*, vol. 2012, pp. 1–12, 2012. [Online]. Available: http://database.oxfordjournals.org/content/2012/bas021.full

[6] F. Rinaldi, S. Clematide, H. Marques, T. Ellendorff, R. Rodriguez-Esteban, and M. Romacker, "Ontogene web services for biomedical text mining," *BMC Bioinformatics*, vol. 15, no. Suppl 14, p. S6, 2014.

[7] F. Rinaldi, S. Clematide, and G. Schneider, "Odin: Advanced text mining in support of the curation process," in *Pacific Symposium on Biocomputing (PSB)*, Big Island, Hawaii, Jan. 2012.

[8] F. Rinaldi, A. P. Davis, C. Southan, S. Clematide, T. R. Ellendorff, and G. Schneider, "ODIN: a customizable literature curation tool," in *Proceedings of the Fourth BioCreative Challenge Evaluation Workshop*, vol. 1, 2013, pp. 219–223.

[9] S. Gama-Castro, F. Rinaldi, A. López-Fuentes, Y. I. Balderas-Martínez, S. Clematide, T. R. Ellendorff, A. Santos-Zavaleta, H. Marques-Madeira, and J. Collado-Vides, "Assisted curation of regulatory interactions and growth conditions of OxyR in E. coli K-12," *Database: The Journal of Biological Databases and Curation*, vol. bau049, pp. 1–13, 2014. [Online]. Available: http://database.oxfordjournals.org/content/2014/bau049

[10] R. Mihalcea, C. Corley, and C. Strapparava, "Corpus-based and knowledge-based measures of text semantic similarity," in *Proceedings of the 21st National conference on Artificial Intelligence*, vol. 1, 2006, pp. 775–780. [Online]. Available: http://www.aaai.org/Papers/AAAI/2006/AAAI06-123.pdf

ISSN 2395-8618

[11] H. Schutze, "Dimensions of meaning," in *Proceedings Supercomputing 1992*, 1992, pp. 787–796. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=236684

[12] A. Wallecha, J. Correnti, V. Munster, and M. van der Woude, "Phase variation of ag43 is independent of the oxidation state of oxyr," *Journal of Bacteriology*, vol. 185, no. 7, pp. 2203–2209, 2003.

[13] M. Zheng, B. Doan, T. D. Schneider, and G. Storz, "Oxyr and soxrs regulation of fur," *Journal of Bacteriology*, vol. 181, no. 15, pp. 4639–4643, 1999. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC103597/

[14] M. Zheng, X. Wang, L. J. Templeton, D. R. Smulski, R. A. LaRossa, and G. Storz, "DNA microarray-mediated transcriptional profiling of the escherichia coli response to hydrogen peroxide," *Journal of Bacteriology*, vol. 183, no. 5, pp. 4562–4570, 2001. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC95351/

[15] K. Toutanova, D. Klein, and C. D. Manning, "Feature-rich part-of-speech tagging with a cyclic dependency network," in *In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL 2003)*, vol. 1, 2003, pp. 252–259. [Online]. Available: http://dl.acm.org/citation.cfm?id=1073478

[16] M. Porter, "Snowball: A language for stemming algorithms," 2001. [Online]. Available: http://www.snowball.tartarus.org/texts/introduction.html

[17] A. Peter, "Efficient java matrix library (EJML)." [Online]. Available: http://ejml.org

[18] M. Krallinger, M. Vazquez, F. Leitner, D. Salgado, A. Chatr-aryamontri, A. Winter, L. Perfetto, L. Briganti, L. Licata, M. Iannuccelli, L. Castagnoli, G. Cesareni, M. Tyers, G. Schneider, F. Rinaldi, R. Leaman, G. Gonzalez, S. Matos, S. Kim, W. Wilbur, L. Rocha, H. Shatkay, A. Tendulkar, S. Agarwal, F. Liu, X. Wang, R. Rak, K. Noto, C. Elkan, Z. Lu, R. Dogan, J.-F. Fontaine, M. Andrade-Navarro, and A. Valencia, "The protein-protein interaction tasks of biocreative III: Classification/ranking of articles and linking bio-ontology concepts to full text," *BMC Bioinformatics*, vol. 12, no. Suppl 8, p. S3, 2011. [Online]. Available: http://www.biomedcentral.com/1471-2105/12/S8/S3

[19] F. Rinaldi, G. Schneider, K. Kaljurand, S. Clematide, T. Vachon, and M. Romacker, "OntoGene in BioCreative II.5," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 7, no. 3, pp. 472–480, 2010.

[20] G. Schneider, S. Clematide, and F. Rinaldi, "Detection of interaction articles and experimental methods in biomedical literature," *BMC Bioinformatics*, vol. 12, no. Suppl 8, p. S13, 2011. [Online]. Available: http://www.biomedcentral.com/1471-2105/12/S8/S13

[21] F. Rinaldi, S. Clematide, S. Hafner, G. Schneider, G. Grigonyte, M. Romacker, and T. Vachon, "Using the OntoGene pipeline for the triage task of BioCreative 2012," *The Journal of Biological Databases and Curation*, vol. bas053, 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3568389/