 doi:10.20944/preprints201610.0012.v1

Article

Bio-Resource Exchange: Study of Prevalence of Antibody Donation and Development of a Web Portal to Facilitate it

Sandeep Subramanian[1] and Madhavi Ganapathiraju[1,2,*]

[1] Language Technologies Institute, Carnegie Mellon University

[2] Department of Biomedical Informatics, and Intelligent Systems Program, University of Pittsburgh

* Correspondence: madhavi@pitt.edu and madhavicmu@gmail.com; Tel.: +1-412-648-9552

Abstract: Bio-molecular reagents like antibodies required in experimental biology are expensive and their effectiveness, among other things, is critical to the success of the experiment. Although such resources are sometimes donated by one investigator to another through personal communication between the two, there is no previous study to our knowledge on the extent of such donations, nor a central platform that directs resource seekers to donors. In this paper, we describe, to our knowledge, a first attempt at building a web-portal titled Bio-Resource Exchange that attempts to bridge this gap between resource seekers and donors in the domain of experimental biology. Users on this portal can request for or donate antibodies, cell-lines and DNA Constructs. This resource could also serve as a crowd-sourced database of resources for experimental biology. Further, in order to index donations outside of our portal, we mined scientific articles to find instances of donations of antibodies and attempted to extract information about these donations at the finest granularity. Specifically, we extracted the name of the donor, his/her affiliation and the name of the antibody for every donation by parsing the acknowledgements sections of articles. To extract annotations at this level, we propose two approaches – a rule based algorithm and a bootstrapped relation learning algorithm. The algorithms extracted donor names, affiliations and antibody names with average accuracies of 57% and 62% respectively. We also created a dataset of 50 expert-annotated acknowledgements sections that will serve as a gold standard dataset to evaluate extraction algorithms in the future.

Keywords: data exchange; resource donations; text mining

 doi:10.20944/preprints201610.0012.v1

1. Introduction

Antibodies and other such wet-lab reagents are vital resources in a variety of experiments in molecular biology. These resources are expensive, and also, their quality is crucial for the success of the experiment. It would be extremely valuable if these reagents if available in spare quantities in one lab, are donated to others when required. This donation will be even more useful if the donor lab has tested the quality of the resource. For example, a research group that studies the protein HMGB1 extensively, will tend to have a reliable and well-tested antibody for it, and could potentially donate some of it to colleagues or collaborators who may need it. Such donations, where possible, can help unfunded junior investigators to be able to carry out experiments. Further, such acts of generosity could spark collaborations between research groups and can serve as a means to connect researchers with similar expertise.

Even in the strongly connected world that we are in today, researchers usually depend on contacting a vendor whose information is readily available online, being unaware of a group that may have the reagents in close proximity within their organization. Increasingly, there has been a trend towards open resource sharing. Open source software, open data sharing and open access of journals, for example, have become pervasive and have accelerated advancement of science.

In these open sharing environments, what are the factors that drive people to do social good? While several individuals have altruistic motives such as contributing to the advancement of science and encouraging junior investigators, there are others who build a reputation for being highly visible donors and build goodwill for future reciprocations. How feasible is it to share material resources among research groups, given that they cannot be shared simply over the Internet?

In this work, we studied the extent to which researchers share biological reagents, specifically antibodies, by parsing the acknowledgements sections of papers available in Pubmed Central. Encouraged by what we found, we developed a web portal to connect donors with seekers of reagents to facilitate and promote sharing of such resources. The portal can serve as a means for people to find locally available resources for their experiments.

The amount of bio-medically relevant content is increasing at an unprecedented rate; two new articles are published on PubMed every minute [1]. Therefore, Information extraction from text documents has seen several advancements and active involvement over the past decade [2-4]. The BioCreative and BioNLP workshop initiatives were created to evaluate text mining and information extraction approaches. Tasks ranging from named entity recognition (NER) on genes, drugs and chemical compounds to protein-protein interaction extraction from PubMed have been a part of these initiatives [5, 6]. Further, GENIA [7] has datasets pertinent to text mining of bio literature and has played an important role in the advancement of Biomedical Natural Language Processing.

Riloff and Jones [8] pioneered an information extraction algorithm that iteratively learns rules for extracting relevant information and in turn uses the information to learn new rules. This approach to learning is often referred to as bootstrapped EM (Expectation Maximization) amongst the machine learning community and is still being used in practice till date [9-11]. Some of the biggest and most successful information extraction systems like Never Ending Language Learner (NELL) [12] have used bootstrapped EM effectively even in the biomedical domain [13]. We adopt this as one of our methods to extract information from literature. The NLP research community has largely stuck to machine learning approaches for information extraction until very recently when rule based systems have seen some resurfacing, while the industry has always stuck to the latter [14]. Rule based information extraction systems have the advantage of being interpretable and can be fine-tuned easily [11]. In this work, we experiment with using a purely hand-engineered rule based extraction system and compare its performance with bootstrapped relation learning system.

 doi:10.20944/preprints201610.0012.v1

2. Materials and Methods

Researchers acknowledge donations from others by thanking them in the acknowledgements section of their published work. In this particular work, we focused on studying acknowledgements pertaining to antibody donations. We mined full-text articles from PubMed Central to extract information at coarse and fine granularities. At the coarse level, we extracted the entire acknowledgement section if a case-insensitive search on the entire acknowledgement section contains the word "antibody" or "antibodies". Authors however, tend to acknowledge multiple things in this section such as manuscript reading, instrument usage and their grant providers for funds. For example, the acknowledgement "We thank Peter Merrifield and Stefano Schiaffino for providing antibodies. This work was supported by grants from the Medical Research Council of Canada. K.E.M.H is a Killam Scholar of the Montreal Neurological Institute." contains information extraneous to the task we are focused on. We therefore had to develop extraction algorithms that can carefully extract out donor names, donor affiliations and antibody names from entire acknowledgement sections.

A. Data Acquisition

PubMed Central provides full-text access to all of its open access papers. As of April 2015, it consists of 5,104 journals with a total of 1,000,148 open access papers. These papers are available for download, free of cost and formatted in XML. We parsed these to extract the acknowledgements section of every paper and searched for a reference to an antibody donation within it. Since generating the entire XML parse tree of every paper was computationally expensive, we ran a shallow parse using regular expressions to parse out just the acknowledgements section.

A crude extraction approach for this task was done using a case insensitive search for the word "antibody" and "antibodies" in the acknowledgements section, which returned 6,533 instances across all papers in all journals. Only a very small fraction of these did not contain a reference to an antibody donation. For example, the sentence "We'd like to thank Doris Thelian for her expert advice on antibody cocktails and flow cytometry data analysis" has absolutely no reference to donation of an antibody.

We then analyzed the extracted paragraphs using information extraction algorithms that we will describe in detail in the subsequent sections to determine the antibody donor name, affiliation and the name of the antibody.

B. Rule Based Extraction

Rule based systems can easily exploit the formal and consistent nature of writing in scientific articles. Rule based information extraction systems that search key-word context windows have been employed with success in the past. The context in which a word occurs (i.e.) the words surrounding it has been exploited in a plethora of tasks. Most recently, distributional representations of words have been learned from word contexts [15-17]. We formulate heuristically determined search rules for information extraction within these word contexts

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 October 2016 doi:10.20944/preprints201610.0012.v1

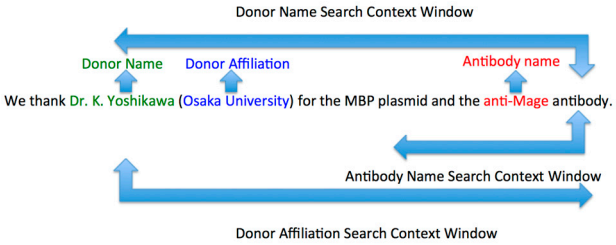

Fig. 1 Rule based extraction

We observed that a vast majority of sentences that corresponded to authors acknowledging donations were written in active voice (Fig. 1). This allows us to search the left context of the word "antibody" for the name of the antibody that was donated (the donor being the subject of a sentence written in active voice will appear to the left of the antibody in a subject-verb-object language like English). However, there are exceptions to this rule; for example: "We thank Dr. Y. Nishiyama for the antibody to UL7." Is written in passive voice, which would require searching the right context instead. We assert that the first word within the left context window of the word "antibody" that is not in the English dictionary or a named entity is the name of the antibody that was donated. If no such word is found in the left context, we then proceed to search the right context. The size of the left and right context windows is set at 4 words, determined after examining the paragraphs from the high-level extraction step. Further, we also search the left context window for any tags (primary, secondary, monoclonal, polyclonal) that may associated with the antibody. While an NER system for antibodies would have been ideal, biomedical NER systems such as BANNER [18] are incapable of tagging antibodies nor is there a corpus from which a supervised one can be trained.

While extracting the name of the donor, we do not fix the size of the context window that we search in. Instead, we keep searching the left context of the antibody name until a named entity labeled as a person is encountered. We found that the name of a donor is typically located far away from the antibody. We used MIT's Information Extraction Library[1] (MITIE) for NER that identifies named entities and provides labels for them like person, organization and location. Another observation about the nature of acknowledgements in biomedical literature was that a person's affiliation almost always occurred immediately after his/her name within brackets. We used this to label the donor's affiliation as the closest organization extracted by our NER in his/her right context but still on the left context of the word "antibody" or "antibodies".

C. **Bootstrapped relation learning**

While rule-based extraction systems are capable of extracting entities with high precision, they require rules to be explicitly defined. This also prevents them from being easily adapted to new domains. Bootstrapping alleviates this problem by automatically learning phrases/relations that identify entities of interest from seeded ground-truth annotations. The following subsection describes the bootstrapping algorithm that we used to automatically learn extraction rules.

[1] https://github.com/mit-nlp/MITIE

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 October 2016 doi:10.20944/preprints201610.0012.v1

We used the idea of bootstrapping to identify antibody names only, and default to using the same context based approach as described in the Rule Based Extraction to identify the donor names, affiliations and antibody names. The algorithm is as follows:

- Seed an initial set of antibody names.
- **E-step**: For every sentence that contains any of the seeded antibody names, run a constituency parse to extract the leaves of immediate parent noun phrase as shown in Fig. 2. and replace the seeded antibody with a wildcard expression like ??.
- These phrases constitute the learned relations. (Fig. 3)
- **M-step**: Extract new antibody names using these learned relations by pattern matching any of these relations with every sentence.
- Repeat the E-step and the M-step iteratively.

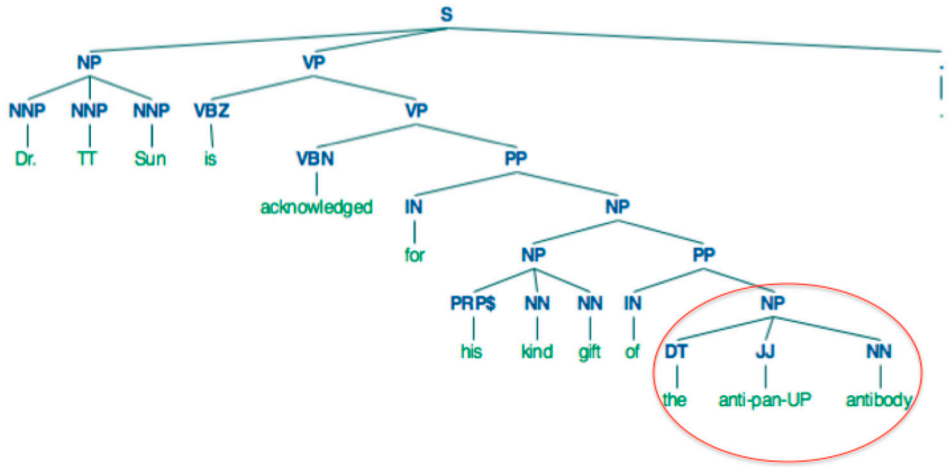

Fig. 2 Constituency parse of a sentence to find an extraction rule

Bootstrapped learning algorithms such as [2] will iteratively learn new relations and ground-truth in an Expectation-Maximization (EM) like approach. The E-step constitutes extracting antibody names either from the initial seed or from the relations learned thus far. The M-step constitutes learning new relations from the current set of antibody names extracted. We observed that 2 EM iterations gave us the best performance. More iterations introduced noisy extraction rules.

Some of the relations learned by this algorithm starting with 40 antibody names as seeds are:

- the mouse ?? antibody
- rabbit ?? antibodies
- ?? monoclonal antibodies
- ?? antibody
- antibody to ??

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 October 2016 doi:10.20944/preprints201610.0012.v1

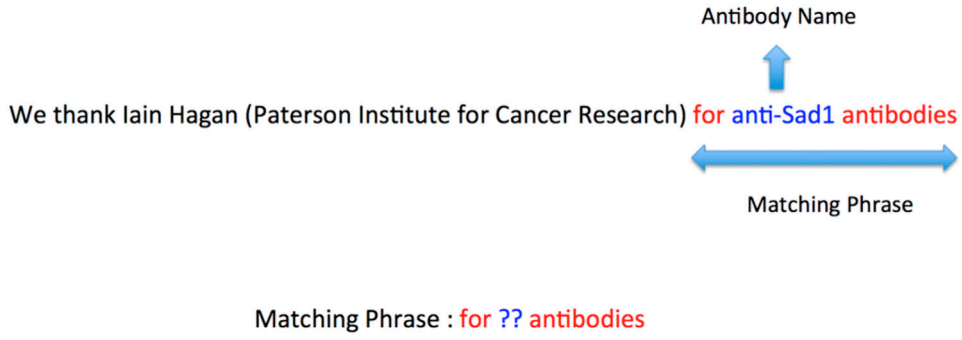

Fig. 3 Example extraction rule

D. Human Annotations

Since there is no dataset with ground-truth annotations for evaluating these algorithms, we sought human annotations for 50 randomly sampled acknowledgement sections. Biologists familiar with the domain were asked to manually annotate donor names, his/her affiliation, the name of the antibody and any of its attributes. We also asked the annotators to identify other bio-resources (e.g. cell-lines) that they could find in the acknowledgements and annotate them with labels describing the resource and the resource name for future work along this line. Further, we also asked them to annotate people or organizations in the acknowledgements that were not part of a donation of a bio-resource for potential application in NER tasks.

Example annotations of sentences describing only antibody donations are shown in Fig. 4 and Fig. 5 and annotations of sentences containing other bio-resources are shown in Fig. 6.

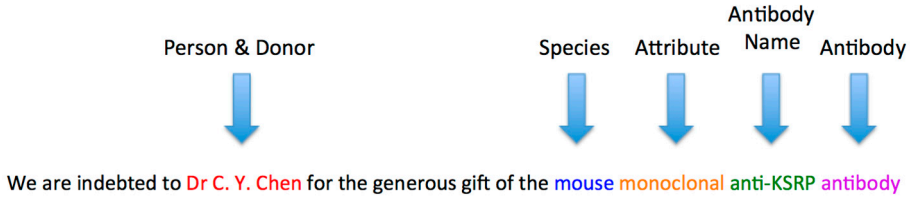

Fig. 4 Example annotation of an antibody donation

Fig. 5 Example annotation of an antibody donation

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 October 2016 doi:10.20944/preprints201610.0012.v1

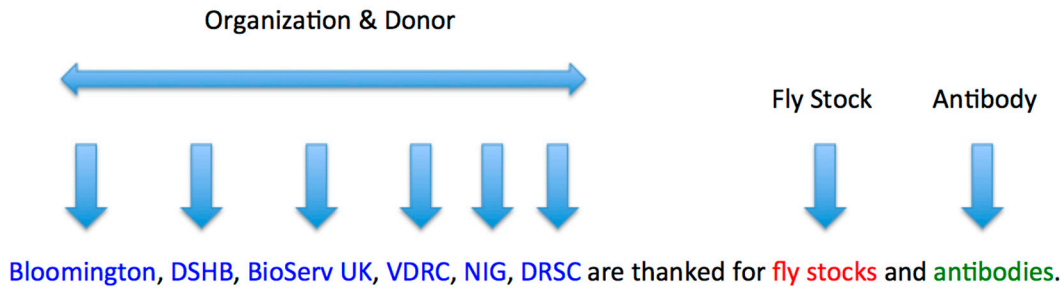

Fig. 6 Example annotation of a fly stock donation

Five biologists participated in the annotation process. Fifty abstracts were annotated overall, of which 18 were annotated by at least 2 individuals. Inter-annotator agreement was 75%. We foresee these annotations being used as ground-truth for other researchers to evaluate their algorithms on the same task. More generally, these annotations could be used to train information extraction and named entity recognition systems. The annotations are formatted in XML, a snippet of which can be seen in Fig. 7.

The dataset of annotated acknowledgements is provided in Supplementary File 1.

```
<acknowledgements>
    <acknowledgement PMCID=5567>
        <content>We thank G. Meissner for the generous gift of anti-RyR2 antibody
        C3-33 and Y. Chen for providing the photomicrograph used in Figure 2.
        </content>
        <annotations>
            <annotation label="person,donor">G. Meissner</annotation>
            <annotation label="antibody name">anti-RyR2</annotation>
            <annotation label="antibody name">C3-33</annotation>
            <annotation label="antibody">antibody</annotation>
            <annotation label="person">Y. Chen</annotation>
        </annotations>
    </acknowledgement>
</acknowledgements>
```

Fig. 7 Human annotations formatted in XML

3. Results and Discussion

We studied frequent donors (people & organizations), frequently donated antibodies, donation trends across journals and over time. These results are presented for both approaches.

A. Rule Based Approach

The rule-based approach extracted a total of 7,589 antibody donations. The number of extracted donations exceeds the number of acknowledgement sections because the algorithm is capable of extracting multiple donations within the same acknowledgement section. Table. 1 contains the top 5 donor names irrespective of their affiliation.

Table 1. Top donors by name

Donor	Number of Donations
Keith Gull	32
Albert Einstein College of Medicine	15
Peter Davies	12
K. Gull	10
Hugo Bellen	10

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 October 2016 doi:10.20944/preprints201610.0012.v1

Table. 2 contains the top 5 donor-affiliation pairs.

Our approach suffers from some weaknesses – the NER system tagged "Albert Einstein College of Medicine" as a person. Also, it is incapable of identifying different ways of writing a donor name (Keith Gull vs K. Gull vs Gull, K) or affiliation (University of Oxford vs Oxford University).

Table 2. Top donors by donor-affiliation pairs

Donor	Affiliation	Number of Donations
Keith Gull	University of Oxford	6
Keith Gull	Oxford University	5
Gary Ward	University of Vermont	4
K. Mackie	Indiana University	3
Yoshihiko Funae	Oosaka City University	3

Table 3 contains the organizations that donated the most antibodies.

Table 3. Top donors by organization

Donor	Number of Donations
University of California	24
NIH	19
Rockefeller University	15
Harvard Medical School	15
University of Pennsylvania	12

Table 4 contains the most frequently donated antibodies.

Table 4. Most frequently donated antibodies

Antibody Name	Number of Donations
plasmids	111
autoantibody	31
DSHB	28
anti-tubulin	14
anti-GFP	14

Table 5 contains the journals that had the most references to an antibody donation. Note that these journals are completely open access, because of which all their articles are available in the data we processed.

Table 5. Journals with the most donations

Journal	Number of Donations
PLoS One	2,894
PLoS Genetics	667
PLoS Pathology	536
PLoS ONE	294
PLoS Biology	286

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 October 2016 doi:10.20944/preprints201610.0012.v1

Fig. 6 is a plot of the number of antibody donations from the year 2000 to early 2015. It shows the exponential increase in the number of donations over the years. However, it is important to note that this could also be a reflection of the fact that the number of open-access publications has also been increasing over the years. It must be noted that the last data-point on the x-axis of the graph is the year 2015, which is the cause for the drop in the number of donations owing to the fact that we collected papers only until early 2015.

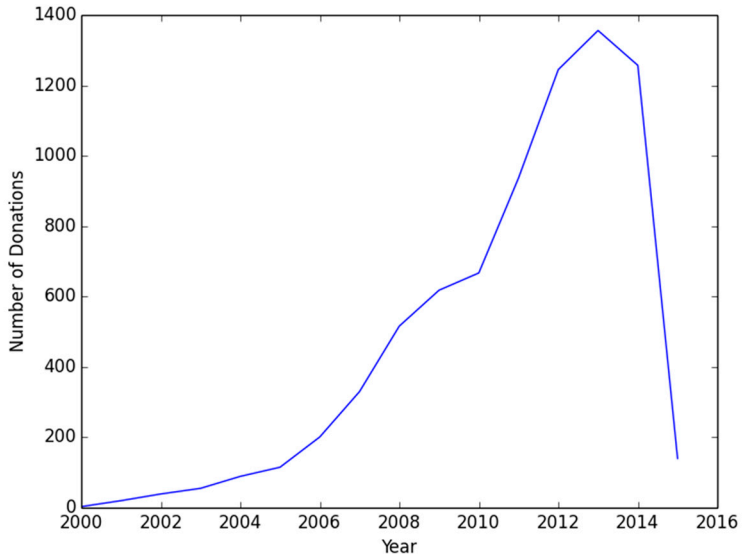

Fig. 6 Year vs number of donations extracted in that year

B. **Bootstrapped Relation Learning**

The bootstrapped relation-learning algorithm extracted a total of 7,864 antibody donations. Table 6 contains the top 5 donors independent of their affiliation.

Table 6. Top donors by donor-affiliation pairs

Donor	Number of Donations
Keith Gull	24
Albert Einstein College of Medicine	20
Erich Buchner	11
Charles Rice	11
K. Gull	10

Table 7 contains the top 5 donor-affiliation pairs

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 October 2016 doi:10.20944/preprints201610.0012.v1

Table 7. Top donors by name

Donor	Affiliation	Number of Donations
Dr. Charles Rice	Rockefeller University	9
Steven S. Gross	Weill Medical College	8
Harold Gainer	NIH	7
Keith Gull	University of Oxford	7
Gary Ward	University of Vermont	6

Table 8 contains the organizations that donated the most antibodies.

Table 8. Top donors by organization

Donor	Number of Donations
NIH	23
Harvard Medical School	22
Rockefeller University	21
University of California	20
University of Pennsylvania	20

Table. 9 below contains the antibody names that were donated the most frequently.

Table 9. Most frequently donated antibodies

Antibody Name	Number of Donations
plasmids	74
anti-mouse	30
anti-gfp	18
anti-tubulin	12
anti-actin	10

Table. 10 contains the journals that had the most references to an antibody donation

Table 10. Most frequently donated antibodies

Journal	Number of Donations
PLoS One	3174
PLoS Pathology	671
PLoS Genetics	577
PLoS Biology	306
PLoS ONE	301

The plot showing the temporal donation trend for this approach was identical to the previous approach and is therefore not included in this section.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 October 2016 doi:10.20944/preprints201610.0012.v1

4. Extraction Evaluation

A. Evaluation

We evaluated the performance of our algorithms, by comparing them to human labeled annotations. We report the accuracies in Table 11. It is evident that we are able to extract characteristics about an antibody using both our proposed approaches. The bootstrapped relation-learning algorithm achieves better performance than the simple rule-based approach at extracting donor and antibody names but doesn't do as well with extracting affiliations.

Table 11. Extraction results

Approach	Accuracy			
	Donor	Affiliation	Antibody Name	Mean
Rule Based	**50%**	**70%**	**50%**	**57%**
Bootstrapped Relation Learning	**57%**	**66%**	**64%**	**62%**

5. Bio-Resource Exchange Web Portal

We developed a resource-sharing web-portal called Bio-Resource Exchange (BRX) available at http://tonks.dbmi.pitt.edu/brx. It is built modularly, with the ability to be a generic resource-sharing platform. It allows users to make requests for or donate resources via a simple customizable form for each resource. At present, resources on BRX include antibodies, DNA constructs and cell-lines. The moment a form is filled in by a user, it appears on a bulletin board (analogous to a news feed on social networking websites) that is visible to all other users in the system. A user's news feed may be filtered based on the type of research he/she is looking for. It allows users to search for specific information, for example, particular antibodies or cell-lines, or posts by specific individuals. BRX also allows users to comment on posts and also puts them in touch with the author via email See Fig. 7.

BRX was developed using the Django web framework with a MySQL backend database. Separate tables were created for each resource type to allow each of them to have different attributes using Django's ORM (Object-relation Mapping). Foreign keys to this table were made to store comments and email correspondences. The rest of the backend elements are designed to make the addition of a new resource extremely simple. Third-party authentication elements on BRX were built using an open source Oauth2 library[2]. For University of Pittsburgh users exclusively, we used LDAP (Lightweight Directory Access Protocol) authentication to let users sign in with their university email accounts. Front-end elements were built using twitter-bootstrap, Vanilla JavaScript and jQuery.

The results from mining literature haven't been incorporated into the website, adding them to the backend is the trivial task of inserting a few records into a MySQL database. In the future, the front-end could include leaderboards of universities, organizations and people who have donated the most resources to promote healthy competition.

[2] https://github.com/omab/python-social-auth

6. Acknowledgments

We thank Dr. Vishwajit L Nimgaonkar and researchers in his group Tulsi, Kodavali, Joel and Lora, and Srilakshmi Chaparala for annotating the sentences we collected. Srilakshmi along with Adam Handen also provided valuable inputs on the Bio-Resource Exchanged website. This work is funded by the BRAINS grant 1R01MH094564 from the National Institute of Mental Health of the National Institutes of Health (NIMH/NIH), USA.

7. References

1. Nawaz, R., P. Thompson, and S. Ananiadou, *Negated bio-events: analysis and identification.* BMC bioinformatics, 2013. **14**(1): p. 14.

2. Finkel, J.R., T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. in Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. 2005. Association for Computational Linguistics.

3. Banko, M., et al. Open information extraction for the web. in IJCAI. 2007.

4. Soderland, S., Learning information extraction rules for semi-structured and free text. Machine learning, 1999. **34**(1-3): p. 233-272.

5. Hirschman, L., et al., Overview of BioCreAtIvE: critical assessment of information extraction for biology. BMC bioinformatics, 2005. **6**(Suppl 1): p. S1.

6. Kim, J.-D., et al. Overview of BioNLP shared task 2011. in Proceedings of the BioNLP Shared Task 2011 Workshop. 2011. Association for Computational Linguistics.

7. Kim, J.-D., et al., *GENIA corpus—a semantically annotated corpus for bio-textmining.* Bioinformatics, 2003. **19**(suppl 1): p. i180-i182.

8. Riloff, E. and R. Jones. Learning dictionaries for information extraction by multi-level bootstrapping. in AAAI/IAAI. 1999.

9. Dempster, A.P., N.M. Laird, and D.B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm.* Journal of the royal statistical society. Series B (methodological), 1977: p. 1-38.

10. Gupta, S. and C.D. Manning, Spied: Stanford pattern-based information extraction and diagnostics. Sponsor: Idibon, 2014: p. 38.

11. Gupta, S. and C.D. Manning, Improved Pattern Learning for Bootstrapped Entity Extraction. CoNLL-2014, 2014: p. 98.

12. Carlson, A., et al. Toward an Architecture for Never-Ending Language Learning. in AAAI. 2010.

13. Movshovitz-Attias, D. and W.W. Cohen. Bootstrapping biomedical ontologies for scientific text using nell. in Proceedings of the 2012 Workshop on Biomedical Natural Language Processing. 2012. Association for Computational Linguistics.

14. Chiticariu, L., Y. Li, and F.R. Reiss. Rule-Based Information Extraction is Dead! Long Live Rule-Based Information Extraction Systems! in EMNLP. 2013.

15. Mikolov, T., et al. Distributed representations of words and phrases and their compositionality. in Advances in neural information processing systems. 2013.

16. Pennington, J., R. Socher, and C.D. Manning, *Glove: Global vectors for word representation*. Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014), 2014. **12**: p. 1532-1543.

17. Turian, J., L. Ratinov, and Y. Bengio. Word representations: a simple and general method for semi-supervised learning. in Proceedings of the 48th annual meeting of the association for computational linguistics. 2010. Association for Computational Linguistics.

18. Leaman, R. and G. Gonzalez. BANNER: an executable survey of advances in biomedical named entity recognition. in Pacific Symposium on Biocomputing. 2008. Citeseer.