Check for updates

SOFTWARE TOOL ARTICLE

SciLite: a platform for displaying text-mined annotations as a means to link research articles with biological data

[version 1; peer review: 3 approved with reservations]

Aravind Venkatesan[1*], Jee-Hyub Kim[1*], Francesco Talo[1], Michele Ide-Smith[1], Julien Gobeill[2], Jacob Carter[3], Riza Batista-Navarro[3], Sophia Ananiadou[3], Patrick Ruch[2,4], Johanna McEntyre[1]

[1]Literature Service group, European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge, UK
[2]SIB Text Mining, Swiss Institute of Bioinformatics, Geneva, Switzerland
[3]National Centre for Text Mining (NaCTeM), Manchester Institute of Biotechnology, Manchester, UK
[4]Bibliomics and Text Mining Group (BiTeM), HES-SO, Geneva, Switzerland

[*] Equal contributors

v1 **First published:** 12 Dec 2016, **1**:25
https://doi.org/10.12688/wellcomeopenres.10210.1
Latest published: 10 Jul 2017, **1**:25
https://doi.org/10.12688/wellcomeopenres.10210.2

Abstract

Biological databases are fundamental to biological research and discovery. Database curation adds highly precise and useful information, usually extracted from the literature through experts reading research articles. The significant amount of time and effort put in by curators, against the backdrop of tremendous data growth, makes manual curation a high value task. Therefore, there is an urgent need to find ways to scale curation efforts by improving data integration, linking literature to the underlying data.
As part of the development of Europe PMC, we have developed a new platform, SciLite, that overlays text-mined annotations on research articles. The aim is to aid Europe PMC users in finding key concepts more easily and provide links to related resources or tools, bridging the gap between literature and biological data.

Keywords

Open Access , Biocuration , Text-Mining , Data Integration , Semantic Web , Web Annotations , RDF , SPARQL , SciLite

Open Peer Review

Approval Status ✔ ? ✔

	1	2	3
version 2 (revision) 10 Jul 2017	✔ view		✔ view
version 1 12 Dec 2016	? view	? view	? view

1. **Lynette Hirschman**, MITRE Corporation, Bedford, USA
2. **Lee Harland**, Scibite Limited, Cambridge, UK
3. **Diana Maynard**, University of Sheffield, Sheffield, UK

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Johanna McEntyre (mcentyre@ebi.ac.uk)

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by the Wellcome Trust [098231], [108758]; and the ELIXIR-EXCELERATE grant, funded by the European Commission within the Research Infrastructures programme of Horizon 2020 [676559].
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2016 Venkatesan A *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Venkatesan A, Kim JH, Talo F *et al.* **SciLite: a platform for displaying text-mined annotations as a means to link research articles with biological data [version 1; peer review: 3 approved with reservations]** Wellcome Open Research 2016, **1**:25 https://doi.org/10.12688/wellcomeopenres.10210.1

First published: 12 Dec 2016, **1**:25 https://doi.org/10.12688/wellcomeopenres.10210.1

1. Introduction

Biological databases are critical for life sciences research, aiding scientists in the process of knowledge discovery. For example, the data within EMBL-EBI data resources were recently estimated to make research more efficient, minimally by £1bn per year (Beagrie & Houghton, 2016). One of the major contributing factors to this value is the structured rich metadata around deposited datasets, gathered both on submission and post-deposition, largely through the efforts of professional data curators who extract pertinent information, making the data more useful to the scientific community. Trained expert curators read numerous scientific articles to annotate data with the corresponding information, such as, biological functions, molecular interactions and gene-disease associations. Biocuration follows a formalised workflow that often involves: a) finding domain relevant articles; b) identifying mentions of the bio-entities of interest in those articles, e.g., proteins, genes, diseases, and accession numbers; c) identifying molecular events and evidences such as, entity interactions and experimental methods; and d) coordinating with system developers to update the information and annotations in the databases (Dauga, 2015; Hirschman *et al.*, 2012). Additionally, curators often collaborate with developers and researchers in developing standards for data collection, nomenclature, vocabularies/ontologies and metadata. This workflow is highly interconnected and precise where the databases are both dependent and required for this practice. Evidently, this process is a time consuming and challenging task, as the bio-entities and biological concept descriptions a curator is in search of could be spread across the article or over different articles (identified through references). For instance, curators often have to identify interacting partners for a large set of proteins or protein complexes, furthermore, categorizing them based on interaction evidences such as, physical and/or causal interactions, often inferred through implicit author statements. The International Society for Biocuration (ISB) (Bateman, 2010) was founded in 2009, to co-ordinate the aforementioned activities and share new methods in curation. Considering the exponential growth in data, curation needs to be supported by sophisticated computational approaches to make it scalable and sustainable in the long term. This is of prime importance considering the key role played by biological databases.

To this end, text-mining methodologies offer one approach to enhance biocuration workflows. Automated information extraction and literature analysis using text-mining has increased in sophistication over the last decade, with the ability to process full text articles, retrieving accession numbers, molecular interactions and gene-disease associations (Rebholz-Schuhmann *et al.*, 2012). Dedicated tools and pipelines have been developed for retrieving articles based on a given article category, tagging bio-entities of interest in articles, and identifying co-occurrences of entities in texts based on set of relations (Ananiadou *et al.*, 2015; Kafkas *et al.*, 2016; Piñero *et al.*, 2015; Pletscher-Frankild *et al.*, 2015). There are a number of text-mining based tools that have been developed to facilitate automated extraction of various article types and biological concepts, such as Textpresso (Müller *et al.*, 2004), iHOP (Fernández *et al.*, 2007), Whatizit (Rebholz-Schuhmann *et al.*, 2008), EAGLi (Gobeill *et al.*, 2009; Gobeill *et al.*, 2015) EVEX (Landeghem & Ginter, 2011), PubTator (Wei *et al.*, 2013) and Argo (Rak *et al.*, 2014). Among these tools, Textpresso has been significantly adapted by data providers and the curation community (Druzinsky *et al.*, 2016; Van Auken *et al.*, 2012). Collaboration between curators and the text-mining community has been instituted by the BioCreative (Critical Assessment of Information Extraction systems in Biology) workshop series, which broadly focuses on improving text-mining outputs in terms of precision and recall to assist curators. Additionally, infrastructure initiatives such as BeCalm and OpenMinTed have been established with an aim to orchestrate various text-mining efforts to make textual annotations more accessible.

Having described the importance of information extraction by using text-mining methods in the context of aiding manual curation, the other essential aspect that plays a key role is effective data integration. The assertions made by curators are a product of information collected from various sources. Therefore, efficient integration of information from disparate data sources is necessary to aid the process of curation. Accordingly, the emergence of the Semantic Web (SW) is starting to have a significant impact on data integration, querying, and knowledge sharing in the life science domain (Antezana *et al.*, 2009; Chen *et al.*, 2013; Jupp *et al.*, 2014; Williams *et al.*, 2012). SW depends on a set of web technologies specifically designed to facilitate machine interoperability, such as Resource Description Framework (RDF), Web Ontology Language (OWL) and the SPARQL Query Language (SPARQL). These technologies are based on the Unique Resource Identifier (URI), which typically uses the Hypertext Transfer Protocol (HTTP) to identify resources or concepts and the relationship between them, affording machine interoperability across the Web. In particular, the graph-based data model of RDF is widely used for data integration across computing platforms due to the flexibility it offers in modelling domain knowledge and integrate data from multiple sources.

Taken together, the two approaches provide the necessary mechanism towards information extraction and seamless integration. By the use of canonical URIs provided by data generators or reference URIs provided by the Identifiers.org registry (Juty *et al.*, 2012), it is possible to represent text-mined entities in RDF and thereby linking those entities to the corresponding data sources. When such information is presented effectually it will aid users in identifying the main concepts, plausibly beginning to reduce the burden of extracting the essence of a given article.

Europe PMC (Europe PMC Consortium, 2015) is a repository for life science literature, hosted at the EMBL-EBI. The repository provides access to over 31 million abstracts and 3 million full-text articles (at the time of writing). Europe PMC was established with an aim to leverage text-mining solutions towards data discovery in scientific articles. In this article, we describe a new platform, SciLite, developed as part of Europe PMC that allows text-mined annotations from any source to be displayed on full text articles. The aim of this platform is to capitalise on the various text-mined outputs in useful ways for the scientific community. Additionally, the open architecture of SciLite allows for collaborative efforts to extend the core functionality by developing applications, making deeper links between the literature and data for clear provenance of curatorial statements.

2. Methods

2.1. Architecture

SciLite was conceived as a platform that allows text-mined annotations from any provider to be highlighted on scientific articles. The platform consists of a data integration pipeline that consolidates various text-mined annotations (refer to section 2.2) and a front-end component to display those annotations on full text research articles browsed within Europe PMC webpages. Currently, the SciLite pipeline operates on full text articles with the license type: CC0, CC-BY, or CC-BY-NC (~775000 articles at the time of writing).

The annotations generated by the SciLite pipeline (see below) are converted to RDF and stored in a triple store (OpenLink Virtuoso version 7.2). The third party annotations are loaded as separate RDF graphs. To account for newly submitted articles, the main annotation pipeline is executed daily, updating the RDF store with new annotations. The RDF store is queried on demand, fetching all the relevant annotations for a given article (refer to section 2.5). Figure 1 provides an overview of the SciLite pipeline. Furthermore, we are investigating the use of MongoDB to further optimise the performance with respect to retrieving and displaying annotations. In such a case, we will maintain a dedicated MongoDB instance for highlighting purposes and a RDF instance.

Should it be desirable, there is also provision for third-party mining software to annotate in-coming Europe PMC content on a daily basis via the installation of virtual machines on the Embassy Cloud.

2.2. Text-mined annotations generated by SciLite

- **Core named entities:** The Europe PMC text-mining pipeline identifies concepts such as gene/protein names, organisms, diseases, GO terms, chemicals and accession numbers. The pipeline is mainly a cascade of three modules:
    - Section tagger (Kafkas *et al.*, 2015): a rule-based module for identifying different sections in articles such as Introduction, Methods and Results.
    - Sentence splitter: an in-house module to identify sentence boundaries.
    - Named entity taggers: the module is based on the dictionary-based approach (Rebholz-Schuhmann *et al.*, 2008) combined with a machine-learning based filter (Chang *et al.*, 2007), for filtering out potential false positives in annotations.

2.3. Third party annotations

- **Biological event annotations:** The National Centre for Text Mining (NaCTeM), Manchester, UK, extracted annotations on approximately 150,000 open-access articles as part of the Europe PMC project in 2015. For the initial phase, phosphorylation events from this set have been included.

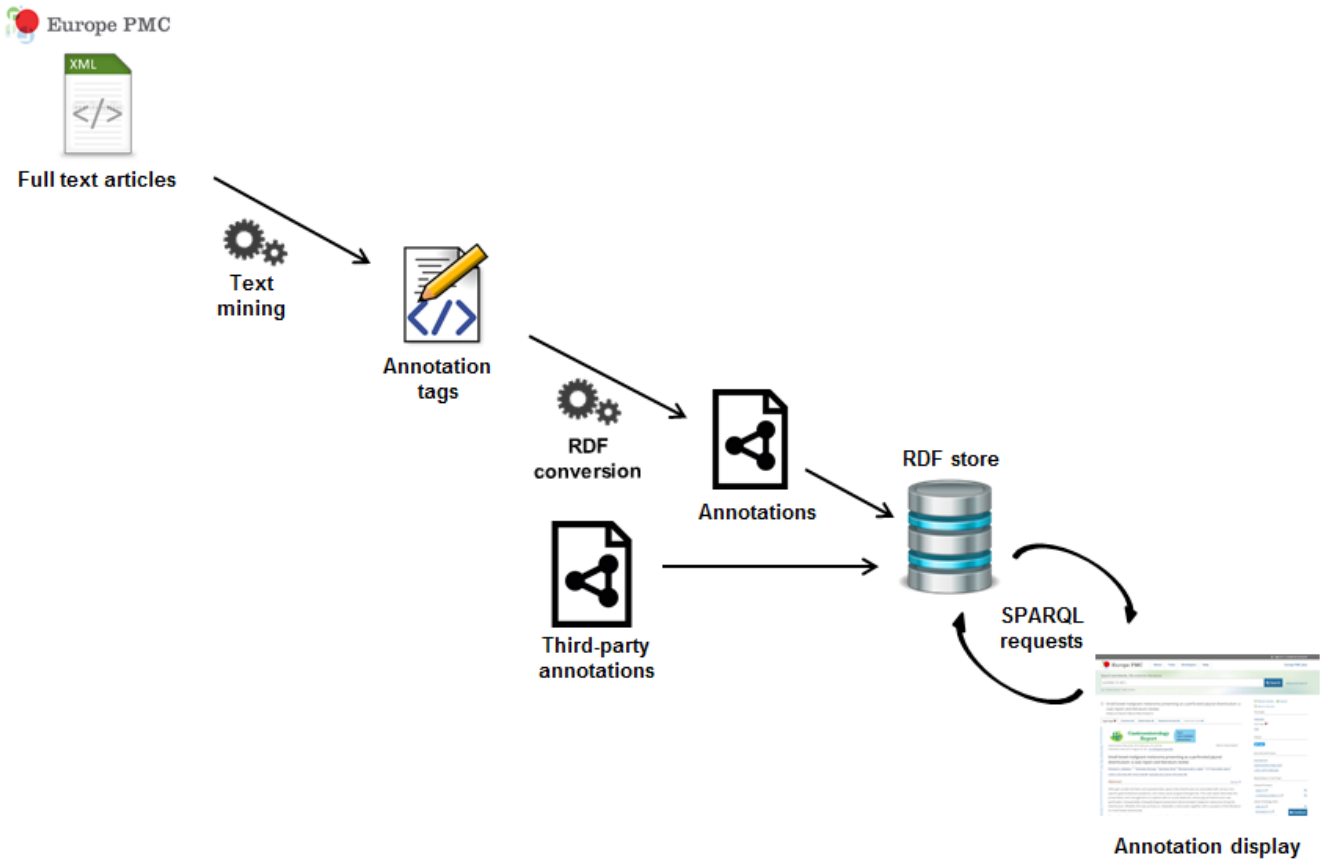


Figure 1. The figure illustrates the SciLite annotation pipeline.

- **Gene function annotations:** Gene Reference into Function (GeneRIF), are annotations on gene function extracted from articles and applied to genes (in Entrez Gene) by MeSH indexers. GeneRIFs can be simple "cut-and-paste" text snippets from abstracts or full text, or sometimes are a more complex synthesis of text fragments. The GeneRIF mappings have been contributed by the Bibliomics and Text Mining group at the University of Applied Sciences, Geneva, Switzerland as part of the Elixir-EXCELERATE project. A dedicated application has been developed to identify GeneRIFs that combine extracts from several sections of the original article (Gobeill *et al.*, 2008). The current dataset contains about over 13,000 GeneRIF statements.

2.4. Converting annotations to RDF

The text-mined outputs are converted to RDF in the Web Annotation Data Model. The data model specification has recently been proposed by the World Wide Web Consortium (W3C) that provides structured models for sharing annotations across different software platforms. The specifications accommodate representation of different kinds of web annotations such as, text quotes/fragments, links, video segments and images. The model is being widely adopted across academia, publishing houses and technology organisations. This is evident through initiatives such as the Annotating All Knowledge (AAK) project and the IAnnotate conference series.

For the current set of annotations incorporated in SciLite, we adopted two types of web annotation models: the Text Quote Selector and the Fragment Selector. The selection of the aforementioned types were based on the kind of the dataset:

- **Text Quote Selector:** We have to represent entities extracted from specific sentences recognised by the text-mining modules from the Europe PMC and NaCTeM datasets. Therefore, this model was suitable to specify the exact tagged concepts such as gene names, diseases and biological events. Figure 2 illustrates a sample representation of an annotation identified by the SciLite text-mining module.

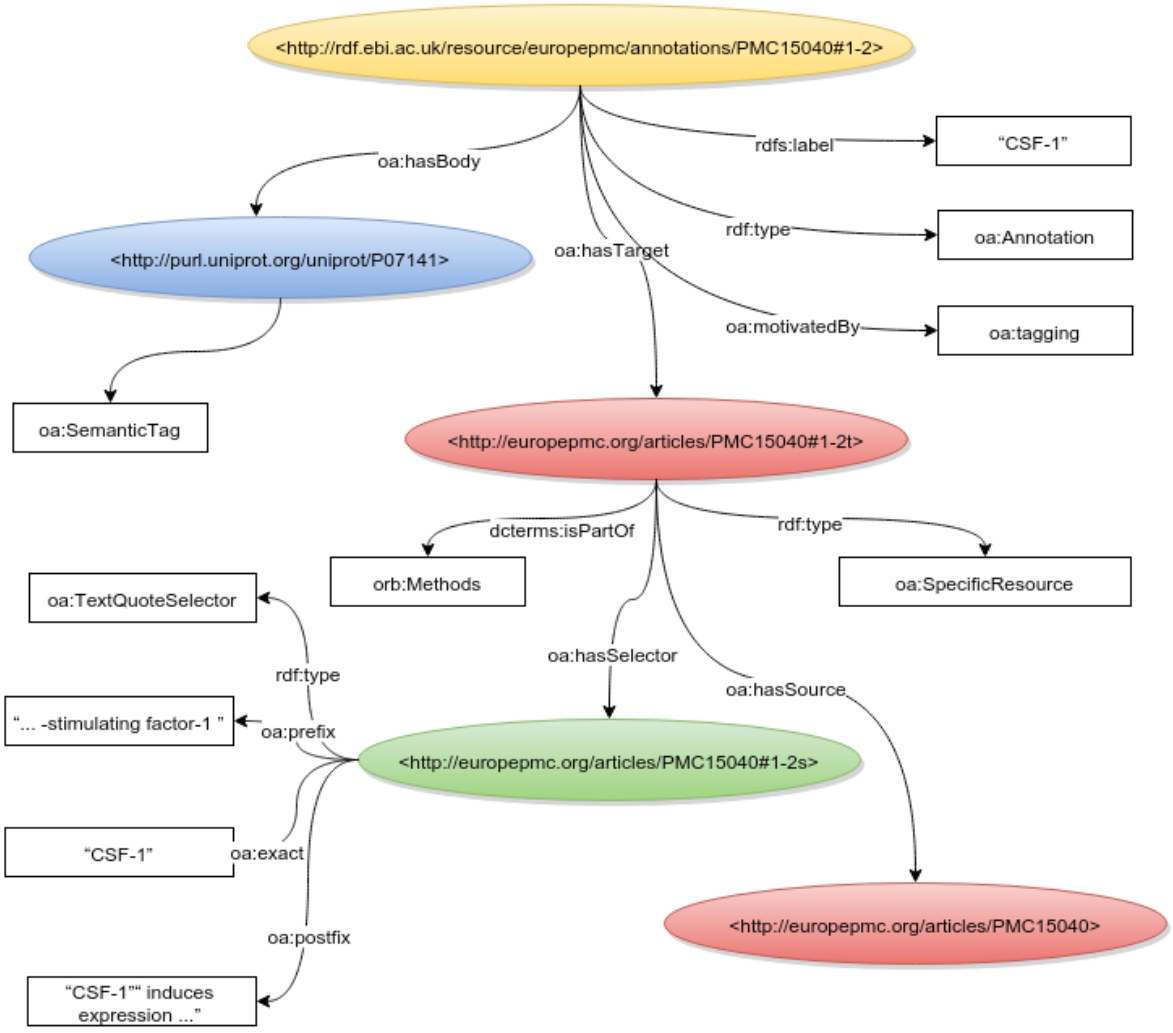


Figure 2. This figure is an illustration of a sample annotation generated by SciLite text-mining pipeline in the Text Quote Selector model.

- **Fragment Selector:** To represent annotations from the GeneRIF dataset, we opted for the Fragment Selector. In this case, the entire sentence had to be represented as opposed to a specific entity in a sentence. Figure 3 presents a sample representation of GeneRIF annotations in the Fragment Selector model.

2.5. User interface

To make annotations available to readers, we have implemented a workflow wherein the RDF triple store is queried on request by the browser, fetching all the relevant annotations for a given article. The retrieved annotations are highlighted (color coded) on the Europe PMC website. Additionally, the annotations are interactive; clicking the highlighted annotation opens a popup-box containing additional information about the annotation (see Figure 4), such as source of annotation, link to the related database, and in certain cases molecular structure of a highlighted protein. The highlighting algorithm consists of the following steps:

- Retrieving all the annotations for that specific PMC ID from the triple store using a SPARQL request.
- Sorting the annotations from the response according to their position in the text (ascending order of occurrence). This is done in order to optimize the performance of the

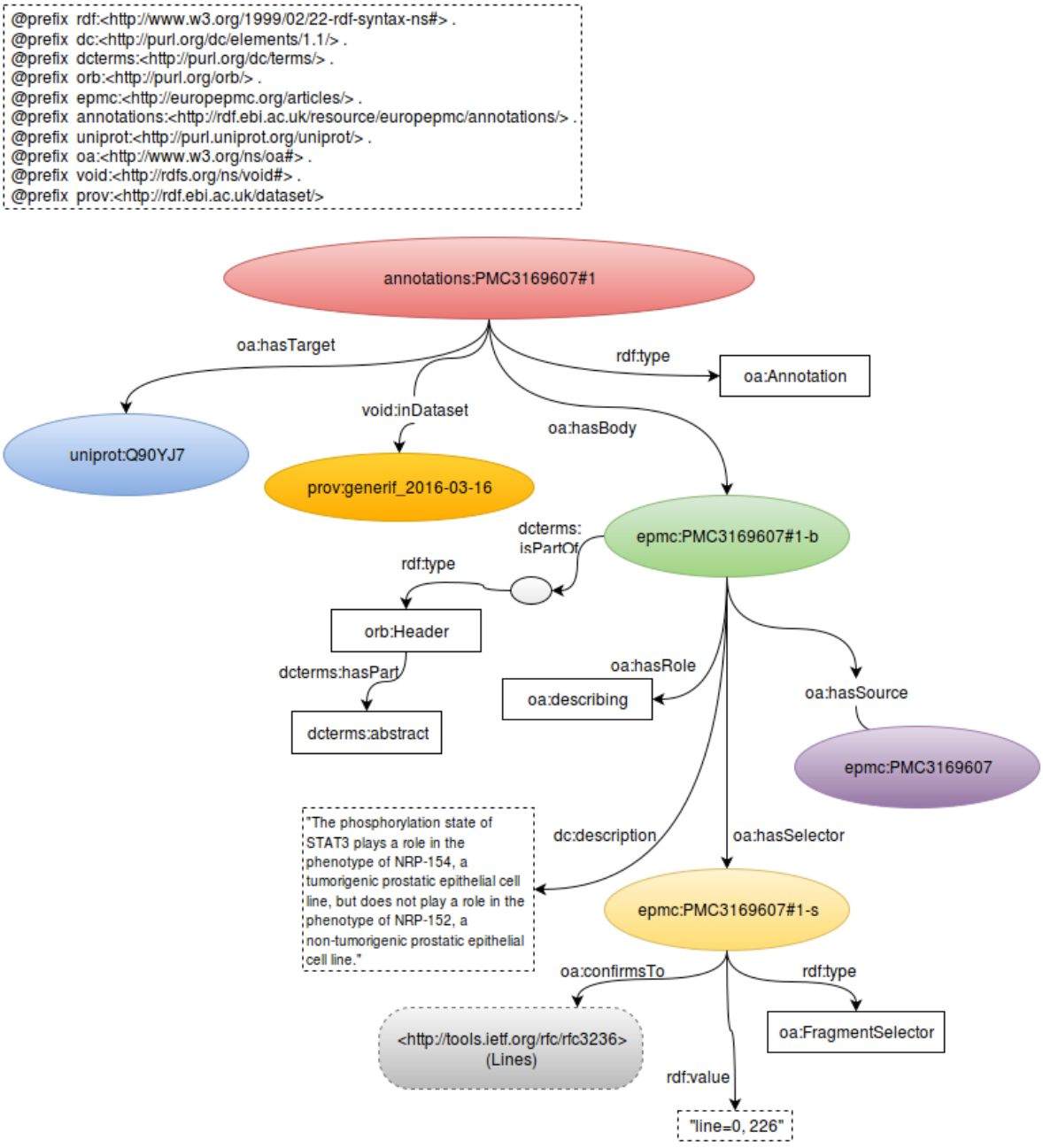
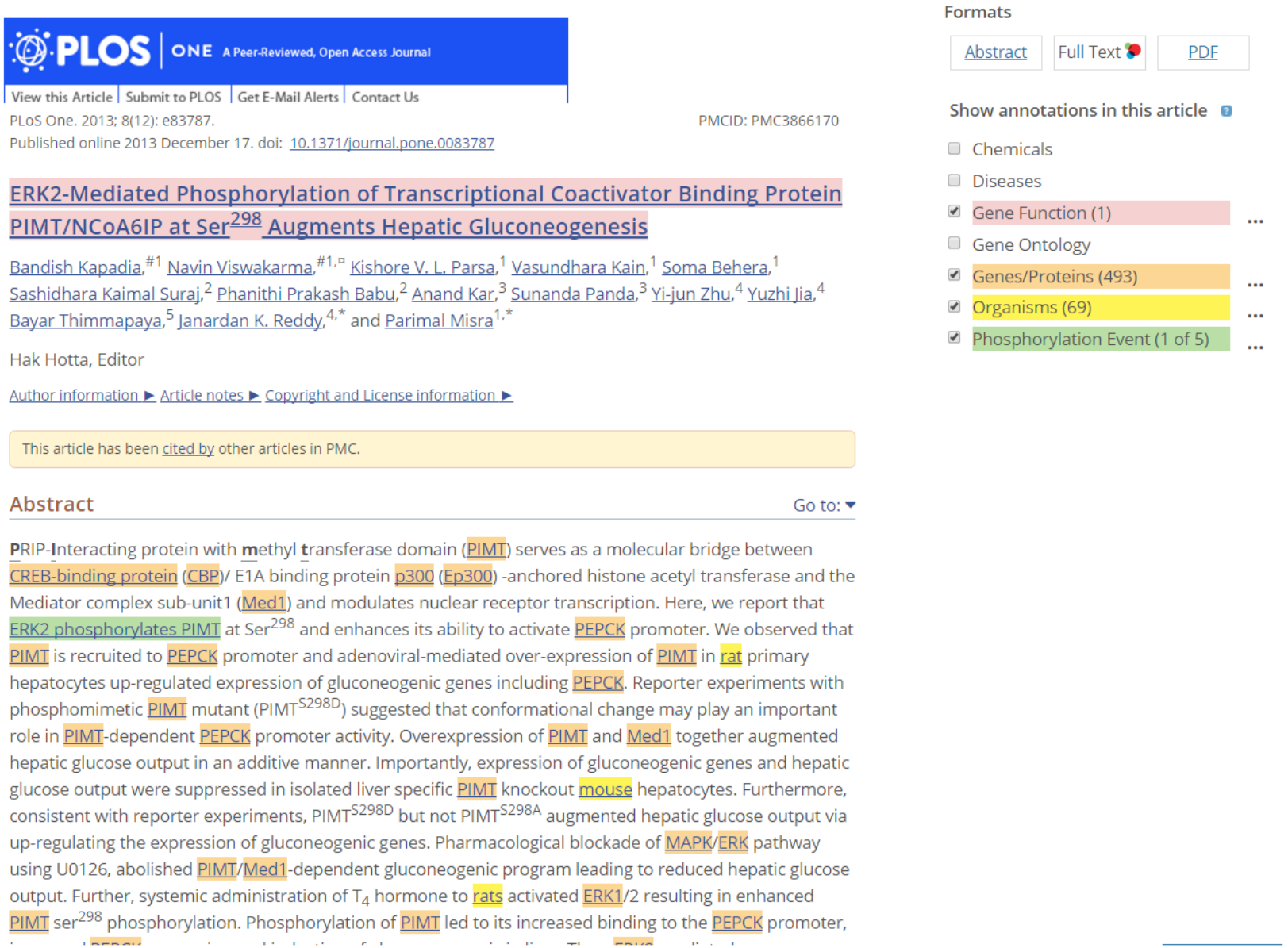


Figure 3. This is an illustration of sample GeneRIF annotations represented in the Fragment Selector model.



PLOS ONE A Peer-Reviewed, Open Access Journal

View this Article | Submit to PLOS | Get E-Mail Alerts | Contact Us

PLoS One. 2013; 8(12): e83787. PMCID: PMC3866170

Published online 2013 December 17. doi: 10.1371/journal.pone.0083787

ERK2-Mediated Phosphorylation of Transcriptional Coactivator Binding Protein PIMT/NCoA6IP at Ser298 Augments Hepatic Gluconeogenesis

Bandish Kapadia,[#1] Navin Viswakarma,[#1,¤] Kishore V. L. Parsa,[1] Vasundhara Kain,[1] Soma Behera,[1] Sashidhara Kaimal Suraj,[2] Phanithi Prakash Babu,[2] Anand Kar,[3] Sunanda Panda,[3] Yi-jun Zhu,[4] Yuzhi Jia,[4] Bayar Thimmapaya,[5] Janardan K. Reddy,[4,*] and Parimal Misra[1,*]

Hak Hotta, Editor

Author information ▶ Article notes ▶ Copyright and License information ▶

This article has been cited by other articles in PMC.

Abstract

Go to:

PRIP-Interacting protein with methyl transferase domain (PIMT) serves as a molecular bridge between CREB-binding protein (CBP)/ E1A binding protein p300 (Ep300) -anchored histone acetyl transferase and the Mediator complex sub-unit1 (Med1) and modulates nuclear receptor transcription. Here, we report that ERK2 phosphorylates PIMT at Ser^{298} and enhances its ability to activate PEPCK promoter. We observed that PIMT is recruited to PEPCK promoter and adenoviral-mediated over-expression of PIMT in rat primary hepatocytes up-regulated expression of gluconeogenic genes including PEPCK. Reporter experiments with phosphomimetic PIMT mutant ($PIMT^{S298D}$) suggested that conformational change may play an important role in PIMT-dependent PEPCK promoter activity. Overexpression of PIMT and Med1 together augmented hepatic glucose output in an additive manner. Importantly, expression of gluconeogenic genes and hepatic glucose output were suppressed in isolated liver specific PIMT knockout mouse hepatocytes. Furthermore, consistent with reporter experiments, $PIMT^{S298D}$ but not $PIMT^{S298A}$ augmented hepatic glucose output via up-regulating the expression of gluconeogenic genes. Pharmacological blockade of MAPK/ERK pathway using U0126, abolished PIMT/Med1-dependent gluconeogenic program leading to reduced hepatic glucose output. Further, systemic administration of T_4 hormone to rats activated ERK1/2 resulting in enhanced PIMT ser^{298} phosphorylation. Phosphorylation of PIMT led to its increased binding to the PEPCK promoter,

Formats

Abstract | Full Text | PDF

Show annotations in this article

- Chemicals
- Diseases
- Gene Function (1) ...
- Gene Ontology
- Genes/Proteins (493) ...
- Organisms (69) ...
- Phosphorylation Event (1 of 5) ...

PLoS One

PLoS One

Figure 4. **The screenshot shows the front-end rendering of various annotation types for a given article on Europe PMC.**

Wellcome Open Research 2016, 1:25 Last updated: 13 NOV 2024

searching process in the article text. For this purpose, the following variables are required:

- *Start position*: contains the current offset in the article text from which the algorithm starts to look for the next annotation. At the beginning its value is 0.
- *Annotations offset*: contains the offsets in the article text where the different annotations' exact match have been retrieved.

- Looping through the annotations:
  - Finding the first occurrence of the exact term, with the specific prefix and postfix (refer Figure 2), in the text of the article starting from *Start position* value.
  - The information about the current annotation is stored into the *Annotations offset* data structure, in order to populate the popup window that will appear once the user will click on the highlighted text into the page.

- Additionally, the *Annotations offset* data structure is extended to accommodate all the overlapping annotations in the longest single piece of text. Therefore, a single item in the *Annotations offset* data structure can contain multiple annotations at this stage. All the items in the *Annotations offset* data structure are sorted by start offset (ascending order).

- Looping through the modified *Annotations offset* data structure and sequentially modifying the article text, highlighting the correspondent piece of text. In particular, for every item of the *Annotations offset*, an opening html tag <*span*> is injected into the start offset position and the relative closing html tag </*span*> is injected into the end offset position. This span has a specific class attribute that will define the colour of the text to be highlighted.

A listener is associated dynamically to the "on click" event of the html *span* tag to display a popup window containing all the necessary information regarding the current annotation.

2.6. Improving annotation accuracy

To reduce false-positives, we have setup a semi-automated process where the user could report a particular erroneous annotation. These reports are collected to improve our text-mining algorithm. This process involves two steps:

a. Once we receive a report from a user we conduct a *quick fix* by deleting the particular annotation from the RDF store. This ensures the erroneous annotation is removed immediately and helps user trust and satisfaction.

b. The error report is used to refine the text-mining algorithm. This step is carried out relatively over a longer period of time. Considering certain reports could be subjective, it is important to examine the reports prior to the inclusion of those exceptions in the algorithm.

3. Discussion

We have adopted a user-driven approach towards the development of SciLite, conducting user research on cognitive and functionality aspects of the platform. So far, we have tested SciLite with 13 users (6 curators and 7 researchers). Overall we received very positive feedback. Most people found at least 3 annotation types useful. Everyone preferred annotations to be turned off by default. Some users commented that seeing different annotation types highlighted in close proximity in the text was useful, as it suggested a possible relationship between those terms e.g. gene-disease. Initially, when users chose an annotation type, such as 'diseases', they found it hard to locate the highlighted terms in the text. To avoid such cases we introduced a term navigation feature (up/down buttons) that allow users browse through the highlighted terms quickly. We found that inaccuracies affected users' trust on the feature and reasoned that user input can be used to refine the precision of our text mining algorithm. Therefore, we have introduced a feedback mechanism where the users can validate or report an erroneous annotation (refer to section 2.6).

Furthermore, the SciLite RDF endpoint is publicly available (refer to section 5) and contains over 1.2 billion triples (as of October 2016). The endpoint is more suitable for computationally heavy analysis, where programs can consume the annotation information by making complex queries. The RDF store is part of the EBI RDF platform, this offers an opportunity to query SciLite in conjunction with other RDF graphs in the EBI RDF platform or other RDF endpoints. Since the annotations are made at the sentence level, it provides deeper provenance.

While highlighting biological terms is useful for skim-reading articles and the links for verifying the highlighted terms, we envision that SciLite is a starting point for the development of future applications that have the potential to improve full text searching, filtering and integration with data. An initial step towards this has been made for Protein Data Bank (PDB) accession numbers with the application: BioJS. For a given PDB accession number, the application fetches the coordinate information and displays the corresponding 3D molecular structure in the popup-box, serving as an interactive visualiser (see Figure 5). Similar applications could be developed for other data types that pull relevant information and display it in the context of the article without the user having to leave the page.

Europe PMC has a wide user base; hence it will be a challenging task to provide annotation types that serve all users. SciLite's open setup allows contribution from other providers, increasing the coverage in annotation types. We welcome contributions from text-mining and other associated communities and encourage them to share annotations on the SciLite platform. We have setup a participation page to assist interested groups to submit annotation data. The page provides details on data requirements, submission format and examples.

4. Conclusion

There is an increasing need to develop tools that bridge the gap between literature and data for the benefit of the scientific community. With SciLite, we have taken the initial step in bringing literature closer to the underlying biological data in a very

resolution [19]. It consists of a dimer with each protomer adopting the typical thiolase fold decorated with specific structural features in the form of a cap (Fig. 1C). The structures of wild-type KasA (416 residues, MW 43.3 kDa), the other fatty acyl elongation β-ketoacyl synthase, and of the acyl enzyme mimic C171Q, both unliganded and with bound thiolactomycin (TLM), were also resolved to high resolution [20]. In line with their high sequence homology, KasA and KasB are structurally similar and superposition of the wild-type apo-dimers (PDB codes 2WGD and 2GP6, respectively) led to a root mean square deviation value of 1.1 Å for 814 aligned Cα atoms sharing 66% sequence identity. The active site, containing the Cys-His-His catalytic triad, is located in the core domain. As shown for KasA [20], TLM binds close to the active site in the malonyl-binding pocket and the hydrophobic acyl-binding channel of the substrate is connected to the malonyl-binding pocket and also directly acces
Thr334 and Thr336 together with Ile235, Phe2
these residues being strictly conserved in KasA
[19] (Fig. 1C, right panel). Thr334 is located clos
catalytic triad. Their side chains are aligned wh
distance of 3.6 Å. Replacement of Thr334 and
side of the tunnel. In contrast, Thr→Asp replac
[17], [21] are very likely to induce a profound p
potential. In addition, the carboxyl group of th
the NE2 atoms of the two catalytic histidines. T
might lead to severe impairment of the enzym
Moreover, this was confirmed by analysis of th
proteins (Figure S2 in Text S1), and was consist
of Asp or Ala at position 334 and 336 does not
profiles of the different KasB derivatives were

Loss of acid-fast staining in a *M. tuberculo*

To study the effect of the two KasB phosphory
replaced either with phosphomimetic (asparta
have shown that acidic residues such as aspar
with regard to functional activity [15], [17], [21
transfer single point mutant alleles, respective
(Table S1 in Text S1, Fig. 2A). These strains con
accD6. The introduction of the *sacB* and *hyg* ca
presence of the point mutation(s) was verified
CDC1551 was constructed using the same plas



Formats

Abstract | Full Text | PDF

Show annotations in this article

- Accession Numbers (1 of 2) ...
- Chemicals (144) ...
- Diseases
- Gene Function (1) ...
- Gene Ontology
- Genes/Proteins (162) ...
- Organisms
- Phosphorylation Event

Figure 5. A screenshot showing the 3D molecular structure of a given PDB accession number.

Wellcome Open Research 2016, 1:25 Last updated: 13 NOV 2024

transparent way. We believe that the impact of SciLite will be more pronounced with community-wide participation. Indeed, we plan in the near future to include other types of annotations such as gene-disease associations and molecular interactions. Additionally, we plan to engage with application developers to extend the functionality of SciLite based on the core annotations.

5. Data and software availability

Annotation data: The RDF data generated by the SciLite platform is available for querying at: http://www.ebi.ac.uk/europepmc/rdf/sparql

Latest source code: https://github.com/EuropePMC/Biojs.Annotator/tree/Biojs.Annotator_1.0

Archived source code as at the time of publication: Biojs.Annotator version 1.0 - DOI: https://doi.org/10.5281/zenodo.183819 (Talo & EuropePMC, 2016).

License:Apache version 2.0

Author contributions

AV contributed to the data integration aspects of SciLite and wrote the manuscript; JHK, FT designed and implemented SciLite text-mining pipeline and front-end component; MIS worked on project management, conducted user research, front-end design of annotations on Europe PMC; JC and RBN designed and worked on NaCTeM annotation extraction; SA guided and conceived NaCTeM annotation extraction; JG designed and extracted GeneRIF dataset; PR conceived extraction of GeneRIF dataset within Elixir-EXCELERATE project. JM conceived SciLite and reviewed the manuscript. All the authors approved the final manuscript.

Competing interests

No competing interests were disclosed.

Grant information

This work was supported by the Wellcome Trust [098231], [108758]; and the ELIXIR-EXCELERATE grant, funded by the European Commission within the Research Infrastructures programme of Horizon 2020 [676559].

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

We would like to acknowledge the valuable inputs provided by: Dr. Paolo Ciccarese and Dr. Timothy Clark, Massachusetts General Hospital, Boston, USA, Dr. Senay Kafkas and the EBI RDF platform team.

References

Ananiadou S, Thompson P, Nawaz R, *et al.*: **Event-based text mining for biology and functional genomics.** *Brief Funct Genomics.* 2015; **14**(3): 213–30.
PubMed Abstract | Publisher Full Text | Free Full Text

Antezana E, Kuiper M, Mironov V: **Biological knowledge management: The emerging role of the Semantic Web technologies.** *Brief Bioinform.* 2009; **10**(4): 392–407.
PubMed Abstract | Publisher Full Text

Bateman A: **Curators of the world unite: the International Society of Biocuration.** *Bioinformatics.* 2010; **26**(8): 991.
PubMed Abstract | Publisher Full Text

Beagrie N, Houghton J: **The Value and Impact of the European Bioinformatics Institute.** 2016.
Reference Source

Chang YM, Kuo CJ, Huang HS, *et al.*: **Analysis and Enhancement of Conditional Random Fields Gene Mention Taggers in BioCreative II Challenge Evaluation.** In *LBM (Short Papers).* 2007; **7**: 1.
Reference Source

Chen H, Yu T, Chen JY: **Semantic Web meets Integrative Biology: a survey.** *Brief Bioinform.* [Accessed September 21, 2016] 2013; **14**(1): 109–25.
PubMed Abstract | Publisher Full Text

Dauga D: **Biocuration: a new challenge for the tunicate community.** *Genesis.* 2015; **53**(1): 132–142.
PubMed Abstract | Publisher Full Text

Druzinsky RE, Balhoff JP, Crompton AW, *et al.*: **Muscle Logic: New Knowledge Resource for Anatomy Enables Comprehensive Searches of the Literature on the Feeding Muscles of Mammals.** *PLoS One.* 2016; **11**(2): e0149102.
PubMed Abstract | Publisher Full Text | Free Full Text

Europe PMC Consortium: **Europe PMC: a full-text literature database for the life sciences and platform for innovation.** *Nucleic Acids Res.* [Accessed June 7, 2016] 2015; **43**(Database issue): D1042–8.
PubMed Abstract | Publisher Full Text | Free Full Text

Fernández JM, Hoffmann R, Valencia A: **iHOP web services.** *Nucleic Acids Res.* 2007; **35**(Web Server issue): W21–6.
PubMed Abstract | Publisher Full Text | Free Full Text

Gobeill J, Gaudinat A, Pasche E, *et al.*: **Deep Question Answering for protein annotation.** *Database.* 2015; **2015**: pii: bav081.
PubMed Abstract | Publisher Full Text | Free Full Text

Gobeill J, Tbahriti I, Ehrler F, *et al.*: **Gene Ontology density estimation and discourse analysis for automatic GeneRiF extraction.** *BMC Bioinformatics.* 2008; **9**(Suppl 3): S9.
PubMed Abstract | Publisher Full Text | Free Full Text

Gobeill J, Patsche E, Theodoro D: **Question answering for biology and medicine.** In *2009 9th International Conference on Information Technology and Applications in Biomedicine.* 2009; 1–5.
Publisher Full Text

Hirschman L, Burns GA, Krallinger M, *et al.*: **Text mining for the biocuration workflow.** *Database (Oxford).* 2012; bas020.
PubMed Abstract | Publisher Full Text | Free Full Text

Jupp S, Malone J, Bolleman J, *et al.*: **The EBI RDF platform: linked open data for the life sciences.** *Bioinformatics.* 2014; **30**(9): 1338–9.
PubMed Abstract | Publisher Full Text | Free Full Text

Juty N, Le Novère N, Laibe C, *et al.*: **Identifiers.org and MIRIAM Registry: community resources to provide persistent identification.** *Nucleic Acids Res.* [Accessed September 14, 2015] 2012; **40**(Database issue): D580–6.
PubMed Abstract | Publisher Full Text | Free Full Text

Kafkas Ş, Pi X, Marinos N, *et al.*: **Section level search functionality in Europe PMC.** *J Biomed Semantics.* 2015; **6**(1): 7.
PubMed Abstract | Publisher Full Text | Free Full Text

Kafkas S, Dunham I, McEntyre J: **Literature Evidence in Open Targets– a target validation platform.** In *Phenotype Day, ISMB.* Orlando, Florida, US. 2016.
Reference Source

Landeghem SV, Ginter F: **EVEX: a PubMed-scale resource for homology-based generalization of text mining predictions.** *Proceedings of BioNLP.* 2011; 28–37.
Reference Source

Müller HM, Kenny EE, Sternberg PW: **Textpresso: An ontology-based information retrieval and extraction system for biological literature.** *PLoS Biol.* 2004; **2**(11): e309.
PubMed Abstract | Publisher Full Text | Free Full Text

Piñero J, Queralt-Rosinach N, Bravo À, *et al.*: **DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes.** *Database*

(Oxford). 2015; **2015**: bav028.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Pletscher-Frankild S, Pallejà A, Tsafou K, *et al.*: **DISEASES: text mining and data integration of disease-gene associations.** *Methods.* 2015; **74**: 83–89.
PubMed Abstract | **Publisher Full Text**

Rak R, Batista-Navarro RT, Rowley A, *et al.*: **Text-mining-assisted biocuration workflows in Argo.** *Database (Oxford).* [Accessed September 26, 2016]. 2014; **2014**: pii: bau070.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Rebholz-Schuhmann D, Arregui M, Gaudan S, *et al.*: **Text processing through Web services: calling Whatizit.** *Bioinformatics.* [Accessed June 7, 2016] 2008; **24**(2): 296–8.
PubMed Abstract | **Publisher Full Text**

Rebholz-Schuhmann D, Oellrich A, Hoehndorf R: **Text-mining solutions for biomedical research: enabling integrative biology.** *Nat Rev Genet.* 2012; **13**(12): 829–839.
PubMed Abstract | **Publisher Full Text**

Talo' F, EuropePMC: **EuropePMC/Biojs.Annotator: Biojs.Annotator 1.0 release.** *Zenodo.* 2016.
Data Source

Van Auken K, Fey P, Berardini TZ, *et al.*: **Text mining in the biocuration workflow: applications for literature curation at WormBase, dictyBase and TAIR.** *Database (Oxford).* 2012; **2012**: bas040.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Wei CH, Kao HY, Lu Z: **PubTator: a web-based text mining tool for assisting biocuration.** *Nucleic Acids Res.* 2013; **41**(Web Server issue): W518–22.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Williams AJ, Harland L, Groth P, *et al.*: **Open PHACTS: semantic interoperability for drug discovery.** *Drug Discov Today.* 2012; **17**(21–22): 1188–1198.
PubMed Abstract | **Publisher Full Text**

Open Peer Review

Current Peer Review Status: ? ? ?

Version 1

Reviewer Report 17 January 2017

https://doi.org/10.21956/wellcomeopenres.10999.r19404

© **2017 Maynard D.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Diana Maynard
Department of Computer Science, University of Sheffield, Sheffield, UK

This paper describes an interesting and useful platform for enhancing research articles with enriched biological metadata, and the ensuing visualisation. It is good to see that both the software and data are also freely available. The description of the architecture and rationale are clearly presented, but I think the paper needs some refocusing. There are many existing similar platforms for displaying such metadata, and thus the concept itself is far from novel. It is unclear how the SciLite approach in principle differs from such existing platforms. The second paragraph of the Introduction provides some references to these, which is used as evidence of the importance of this kind of tool, along with the current development of new infrastructures such as OpenMinTed, but some explanation of how SciLite could (or could not) be integrated with these would be useful.

What is interesting in this work is the particular focus on the biological data and the underlying approach to annotation, but this is treated rather superficially in the paper.
Something which concerns me a little is that the approach to annotation itself seems rather shallow – this may simply due to the brevity of description and the focus more on architectural aspects. In the Introduction, Semantic Web Technologies are mentioned, but I recommend going a little bit beyond the concept that “ontologies are useful” and explaining in more detail how they help with the task. From the description of the annotation process and visualisation, this is not apparent, although linking terms to ontologies offers great potential for enhanced visualization and exploration for the end-user (though it’s not clear whether this is offered in the platform, and if not, why not). The authors claim that “when such information is presented effectually it will aid users in identifying the main concepts” but it is not clear how this is possible in the current platform. Note also that “effectually” should be replaced with “effectively”.

I would suggest expanding Section 2.2 to explain better the methodology for named entity extraction. From the description, the approach sounds rather simplistic, but I suspect that this description hides some complexity. A fuller description of the actual dictionary-based approach and spurious entity filtering techniques would be useful. The reference given is only a summary of dictionary-based approaches in general, but I couldn’t see anywhere a description of or link to the

actual technique used in SciLite. It is also not clear how major issues such as entity disambiguation and variation are dealt with, in order to ensure correct linking to the ontology.

In Figure 2, please make it clearer what the original text is from which the figure is derived. Figure 4 does not actually show the popup containing additional information about the annotation, but the text describing it indicates that it does. It would be useful to see this popup also.

Section 2.5 goes into quite some detail about the interface, but most of this is rather standard technology used in most annotation/text mining GUIs (for instance, the mechanism for sorting the annotations to display) and could easily be omitted. In the last bullet point in this section, “correspondent” should be “corresponding”.

In Section 2.6 (and throughout the paper), “setup” should be two words when used as a verb. In bullet point b, it is not clear what is meant by “relatively” here – I think this word could just be deleted.

In Section 3, the presentation of the evaluation could do with being more detailed, as this is an important aspect of the platform. What does it really mean to say that “most people found at least 3 annotation types useful”? This is rather vague. How many annotation types were there in total? Did they find the actual annotations useful too? What about the ontologies and the popup information? How did they use the information shown to them, and how did it enhance their experience? What did they find not useful? Did they have suggestions for improvement? How accurate were the annotations? Are there any plans to provide relational information also (as it seems to be indicated by the users that this would be useful)? There are tools that already can provide this kind of information, see for example GATE’s Prospector tool:

V. Tablan, K. Bontcheva, I. Roberts, and H. Cunningham. Mímir: An open-source semantic search framework for interactive information seeking and discovery. Journal of Web Semantics: Science, Services and Agents on the World Wide Web, 2014.

A final point – throughout the paper, the authors have added a comma after the phrase “such as”. The comma should precede this phrase. For example, “web annotations such as, text…” should be “web annotations, such as text…”

In summary, the work is interesting, but I think it could be enhanced greatly by a considerable rewrite with more focus on the underlying methodology and evaluation parts.

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 04 Jul 2017

Aravind Venkatesan

Dear Diana Maynard,
Thank you for providing your inputs with regard to our paper. We have reworked the manuscript, addressing your suggestions.

Overview of the changes made are as follows:

- Section 1. Introduction: The section has being restructured and now acknowledges contributions made by other annotation tools.
- Section 2. Methods:
  - Sub-section 2.1 (Architecture) now provides an overview of the architecture of SciLite (includes new Figure 1)
  - Sub-section 2.2 (Annotation types) has been expand to include new annotations types (gene-disease relationships and manually curated protein-protein interactions).
  - Sub-section 2.3 describes the Web Annotation model (includes new Figures: 2 and 3).
  - Sub-section 2.4 (User interface) has been reworked to provide concise description of the highlighting process.
  - Sub-section 2.5 now includes a new Figure (4) that provides an overview on the feedback mechanism.
- Section 3. Discussion: The section now provides details on how the user research was conducted. The section has been restructured with two new sub sections:
  - 3.1 – Engagement with text-mining community
  - 3.2 – Future directions.
- Supplementary material: The supplementary material provides details on the classes and relations used to model the annotations as RDF along with sample SPARQL queries.

We think that the manuscript is much improved as a result of your feedback. Please see a point-by-point response, below:

**The description of the architecture and rationale are clearly presented, but I think the paper needs some refocusing. There are many existing similar platforms for displaying such metadata, and thus the concept itself is far from novel. It is unclear how the SciLite approach in principle differs from such existing platforms. The second paragraph of the Introduction provides some references to these, which is used as evidence of the importance of this kind of tool, along with the current development of new infrastructures such as OpenMinTed, but some explanation of how SciLite could (or could not) be integrated with these would be useful.*

The Introduction section now acknowledges the state of the art and describes the complementary role of SciLite.
The novelty of SciLite is that it allows annotations from multiple sources to be included and displayed in Europe PMC, for the purposes of reuse for readers, and to integrate literature with data on an infrastructural level. Europe PMC is a database that is updated daily with new content and already has a large user base, so rather than expecting users to visit another interface, the text-mining results are incorporated into their usual search

behaviour. SciLite complements OpenMinTed in that the outputs from such infrastructures, which is focussed on the needs of text miners rather than end users, can be made widely and publically available. This aspect has been included in section 3.1.

**What is interesting in this work is the particular focus on the biological data and the underlying approach to annotation, but this is treated rather superficially in the paper.*
Something which concerns me a little is that the approach to annotation itself seems rather shallow – this may simply due to the brevity of description and the focus more on architectural aspects. In the Introduction, Semantic Web Technologies are mentioned, but I recommend going a little bit beyond the concept that "ontologies are useful" and explaining in more detail how they help with the task. From the description of the annotation process and visualisation, this is not apparent, although linking terms to ontologies offers great potential for enhanced visualization and exploration for the end-user (though it's not clear whether this is offered in the platform, and if not, why not). The authors claim that "when such information is presented effectually it will aid users in identifying the main concepts" but it is not clear how this is possible in the current platform. Note also that "effectually" should be replaced with "effectively".

We acknowledge that some of the original text was misleading. In the current version of the manuscript parts of the introduction has been reworked, focussing on sharing of text-mined annotations for the benefit of the end-user. In this regard, we have now added a paragraph to review the state of the art, highlighting the advantages of using the Web Annotation Data model and the complementary role played by SciLite.

**I would suggest expanding Section 2.2 to explain better the methodology for named entity extraction. From the description, the approach sounds rather simplistic, but I suspect that this description hides some complexity. A fuller description of the actual dictionary-based approach and spurious entity filtering techniques would be useful. The reference given is only a summary of dictionary-based approaches in general, but I couldn't see anywhere a description of or link to the actual technique used in SciLite. It is also not clear how major issues such as entity disambiguation and variation are dealt with, in order to ensure correct linking to the ontology.*

The main aim of SciLite is to be a platform for sharing the text mining outputs multiple sources, in this sense, describing text mining methodology used by contributors is beyond the objectives of SciLite. Since the Europe PMC text mining pipeline also serves as a provider for SciLite we briefly mention the steps involved, citing the articles in which the methodology of these steps are described.

**In Figure 2, please make it clearer what the original text is from which the figure is derived.*

We have simplified Figures 2 and 3. The figure legends now describe the sample model with links to the source articles that contains the original annotation.

**Figure 4 does not actually show the popup containing additional information about the annotation, but the text describing it indicates that it does. It would be useful to see this popup also.*

The correction has now been made, the text now refers to Figure 6 which includes a popup

window and an example link has been included for the same.

**Section 2.5 goes into quite some detail about the interface, but most of this is rather standard technology used in most annotation/text mining GUIs (for instance, the mechanism for sorting the annotations to display) and could easily be omitted. In the last bullet point in this section, "correspondent" should be "corresponding".*

We agree with the point being made. We have now reworked the section providing concise description of the highlighting process.

**In Section 2.6 (and throughout the paper), "setup" should be two words when used as a verb.*

We have now made the necessary corrections throughout the article.

**In bullet point b, it is not clear what is meant by "relatively" here – I think this word could just be deleted.*

Step b requires the providers to make necessary improvements to their algorithms and this is more time consuming than step a, which is a "quick fix". The text has been rephrase, we hope the sentence reads better.

**In Section 3, the presentation of the evaluation could do with being more detailed, as this is an important aspect of the platform. What does it really mean to say that "most people found at least 3 annotation types useful"? This is rather vague. How many annotation types were there in total? Did they find the actual annotations useful too? What about the ontologies and the popup information? How did they use the information shown to them, and how did it enhance their experience? What did they find not useful? Did they have suggestions for improvement? How accurate were the annotations?*

We agree with the point being made with regards to user research. Section 3 has been expanded, providing details on how the usability tests were conducted, aspects of SciLite that the users liked and how we improved SciLite based on the feedback we received.

**Are there any plans to provide relational information also (as it seems to be indicated by the users that this would be useful)?*

SciLite now includes gene-disease relationship provided by two contributors (Open Targets platform and DisGeNET) and protein-protein interactions from IntAct. To accommodate the latest development we have expanded sub-section 2.2

**There are tools that already can provide this kind of information, see for example GATE's Prospector tool:*
V. Tablan, K. Bontcheva, I. Roberts, and H. Cunningham. Mímir: An open-source semantic search framework for interactive information seeking and discovery. Journal of Web Semantics: Science, Services and Agents on the World Wide Web, 2014.

We have now reworked the manuscript focusing on SciLite as a platform for bringing

various text mined annotations to the wider scientific community. In this sense, we believe that Mímir (as a search framework) is broader in its application when compared to SciLite.

**A final point – throughout the paper, the authors have added a comma after the phrase "such as". The comma should precede this phrase. For example, "web annotations such as, text..." should be "web annotations, such as text..."*

We have edited the text for grammatical corrections and hope that the text now reads significantly better!

Competing Interests: No competing interests were disclosed.

Reviewer Report 13 January 2017

https://doi.org/10.21956/wellcomeopenres.10999.r18964

© 2017 Harland L. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Lee Harland
Scibite Limited, Cambridge, UK

* This article describes the "SciLite" tool for enriching biomedical articles with annotations derived from text mining and other sources and the benefits such an approach has for users. It is well written and referenced and outlines the principles in a consistent way. The application links at the bottom of the article pointed to valid working websites. I think it valid to review both the paper and the tool itself and all of the functionality I tried on the website worked very well and there were no software errors, validating that the system is running as described in the paper.

In terms of areas for consideration:

* The introduction is fine, the major omission being coverage of previous work on in-line biological annotation, I would strongly support the acknowledgement of the reflect.ws tool which was perhaps the most prominent one here - and note that many biological web page annotators, both commercial and non-commercial exist (e.g. EXTRACT https://academic.oup.com/database/article/doi/10.1093/database/baw005/2630147/EXTRACT-interactive-extraction-of-environment which works very well on the Europe PMC website). The introduction should clarify the advantages the authors approach brings over 3rd party, javascript bookmarklet approaches.

* Its not clear what the actual tagging methodology is 2.2 states "Named entity taggers: the module is based on the dictionary-based approach (Rebholz-Schuhmann *et al.*, 2008) combined with a machine-learning based filter (Chang *et al.*, 2007), for filtering out potential false positives in annotations.". Does this use these existing tools or have the authors invented new ones and which

vocabularies were used? I think those in the text mining field would be interested to know more about how the text mining is done [I would agree this is not a text mining methods paper, but this is very light on detail here]. Maybe I missed it but "Sentence splitter: an in-house module to identify sentence boundaries." could also do with some form of definition (what's it based on, were any improvements needed etc, the tool correctly doesn't split 'Nkx6.1' so there is some logic being coded there).

* The use of RDF is interesting but no justification is given as to what advantages RDF provides over other database mechanisms. Why was RDF chosen over other systems? It seems that the queries are fairly straight forward (for an input ID give me annotations) and I'm not sure they require RDF necessarily. There are genuine other reasons I could think of but it would be good for the authors to describe this choice. Indeed the authors then state they are considering use of MongoDb (using RDF or not?). This seems to hint at some problems using RDF, but no information is given. I think the authors either need to describe why or perhaps remove this sentence entirely as these seems like more internal technical discussion?

* The system performed well for human centric articles but less so when looking at plant or bacterial data... This is a really hard problem in text mining and I don't expect the authors to solve it here but it would be good to perhaps outline whether the system could be tuned to address anything here as non-human area researchers will be using the tool.

* 2.6.a When error reported there is a "Quick fix" where the annotation is deleted. It is not clear if this is just for that 1 article or all articles as it may be an erroneous synonym? While 2.6.b hints to this "refine the text-mining algorithm." it gives no detail as to what this entails. Again, this is not a text mining methods paper, but examples would be useful here.

* Further, Figure 1 - doesnt touch on maintainence. When source vocabs change (e.g. as 2.6.b hints) do all articles need to be re-run? Are there version markers on the annotations? How regularly are older articles re-annotated? Is provenance applied to annotations, I didn't see any void/prov markers in Figure 2.

* Following on the provenance theme, the authors state: "Additionally, the open architecture of SciLite... data for clear provenance of curatorial statements.". I think this is shown in Figure 3 but only if you know what the RDF prefixes for void and prov mean, so perhaps should be described in the main article text for non RDF people.

* Finally, section (2) talks about the systems handling of false positives but does not mention false negatives at all which are equally irritating to the user. What is the proposed approach here?

* 3. "Everyone preferred annotations to be turned off by default." - I think thats a fairly important statement, did you ask why?

* The link to the Sparql end point http://www.ebi.ac.uk/europepmc/rdf/sparql provides only a default view and it is left to the user to figure out the schema and queries necessary. I am not sure of the best mechanism (supplementary data or footnote?) but an example query or two would be very helpful to get started

* In general, figures could do with a bit more description as to what they are showing.

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 04 Jul 2017

Aravind Venkatesan

Dear Lee Harland,
Thank you for providing your inputs with regard to our paper. We have reworked the manuscript, addressing your suggestions.
Overview of the changes made are as follows:

- Section 1. Introduction: The section has being restructured and now acknowledges contributions made by other annotation tools.
- Section 2. Methods:
  - Sub-section 2.1 (Architecture) now provides an overview of the architecture of SciLite (includes new figure 1)
  - Sub-section 2.2 (Annotation types) has been expand to include new annotations types (gene-disease relationships and manually curated protein-protein interactions).
  - Sub-section 2.3 describes the Web Annotation model (includes new figures: 2 and 3).
  - Sub-section 2.4 (User interface) has been reworked to provide concise description of the highlighting process.
  - Sub-section 2.5 now includes a new figure (4) that provides an overview on the feedback mechanism.
- Section 3. Discussion: The section now provides details on how the user research was conducted. The section has been restructured with two new sub sections:
  - 3.1 – Engagement with text-mining community
  - 3.2 – Future directions.
- Supplementary material: The supplementary material provides details on the classes and relations used to model the annotations as RDF along with sample SPARQL queries.

We think that the manuscript is much improved as a result of your feedback. Please see a point-by-point response, below:

** The introduction is fine, the major omission being coverage of previous work on in-line biological annotation, I would strongly support the acknowledgement of the reflect.ws tool which was perhaps the most prominent one here - and note that many biological web page annotators, both commercial and non-commercial exist (e.g. EXTRACT https://academic.oup.com/database/article/doi/10.1093/database/baw005/2630147/EXTRACT-interactive-extraction-of-environment which works very well on the Europe PMC website). The introduction should clarify the advantages the authors approach brings over 3rd party, javascript bookmarklet approaches.*

We have now added a paragraph in the Introduction section acknowledging the contributions made by other annotation tools. The section highlights the rationale behind the development of SciLite in comparison with other tools.

The main advantage of SciLite is that the platform aggregates annotations made by different tools. For instance, if desired annotations made by Reflect or Extract can be shared in SciLite. This allows users to view different types of annotations for a given article on Europe PMC website.

** Its not clear what the actual tagging methodology is 2.2 states "Named entity taggers: the module is based on the dictionary-based approach (Rebholz-Schuhmann et al., 2008) combined with a machine-learning based filter (Chang et al., 2007), for filtering out potential false positives in annotations.". Does this use these existing tools or have the authors invented new ones and which vocabularies were used? I think those in the text mining field would be interested to know more about how the text mining is done [I would agree this is not a text mining methods paper, but this is very light on detail here]. Maybe I missed it but "Sentence splitter: an in-house module to identify sentence boundaries." could also do with some form of definition (what's it based on, were any improvements needed etc, the tool correctly doesn't split 'Nkx6.1' so there is some logic being coded there).*

The purpose of SciLite is to be a platform for sharing the text mining work of others, so in a sense the text mining methodology used by contributors is not the business of SciLite. However, as the Europe PMC text mining pipeline also serves as a provider for SciLite we have briefly mentioned the steps involved and referred to the articles in which the methodology of these steps is described.

** The use of RDF is interesting but no justification is given as to what advantages RDF provides over other database mechanisms. Why was RDF chosen over other systems? It seems that the queries are fairly straight forward (for an input ID give me annotations) and I'm not sure they require RDF necessarily. There are genuine other reasons I could think of but it would be good for the authors to describe this choice. Indeed the authors then state they are considering use of MongoDb (using RDF or not?). This seems to hint at some problems using RDF, but no information is given. I think the authors either need to describe why or perhaps remove this sentence entirely as these seems like more internal technical discussion?*

We agree with the point being made. Based on our performance tests we find MongoDB to be more performant for retrieving annotations for a given PMCID. Whereas, RDF is a natural choice for linked data and is suitable for performing graph-based queries.
We have now restructured parts of the section to highlight this point, correspondingly we have changed Figure 1.

** The system performed well for human centric articles but less so when looking at plant or bacterial data... This is a really hard problem in text mining and I don't expect the authors to solve it here but it would be good to perhaps outline whether the system could be tuned to address anything here as non-human area researchers will be using the tool.*

Our approach to resolving the challenges of text mining across different fields and phylogenies is to engage the text mining community to provide their insights and solutions in SciLite. This key message of the article has been reworked to be much clearer (we acknowledge that some of the original phrasing was misleading).

** 2.6.a When error reported there is a "Quick fix" where the annotation is deleted. It is not clear if this is just for that 1 article or all articles as it may be an erroneous synonym?*

“Quick fix” is only for the specific instance in a specific article where an error was reported, as means of a quick response to user feedback. We have another mechanism which allows providers to correct their algorithms to reduce errors globally. This method could be improved further in the future to extend the deletion of a specific instance to “all instances in the article”, for example.

**While 2.6.b hints to this "refine the text-mining algorithm." it gives no detail as to what this entails. Again, this is not a text mining methods paper, but examples would be useful here.*

Error reports are submitted to the corresponding provider as refining text mining algorithms is not part of the SciLite workflow. However, improvements would entail, for instance, revising dictionaries by removing erroneous synonyms.

** Further, Figure 1 - doesnt touch on maintenance. When source vocabs change (e.g. as 2.6.b hints) do all articles need to be re-run? Are there version markers on the annotations? How regularly are older articles re-annotated? Is provenance applied to annotations, I didn't see any void/prov markers in Figure 2.*

In SciLite when new dataset is submitted, all annotations are checked and in cases where the annotations have changed the corresponding collections (in MongoDB) and the RDF triples are updated. When new data set is available the provenance graphs are updated.
Figure 1 has been replaced with a new Figure, the figure now provides an overview of the SciLite workflow, touching upon the maintenance aspect.
Figure 2 has been modified to refocus on the representation of annotations in the Web Annotation model. We hope the new figure is more understandable to the readers.

** Following on the provenance theme, the authors state: "Additionally, the open architecture of SciLite... data for clear provenance of curatorial statements.". I think this is shown in Figure 3 but only if you know what the RDF prefixes for void and prov mean, so perhaps should be described in the main article text for non RDF people.*

In the indicated sentence we were referring to annotations as textual evidence and were not referring to “provenance” in the context of RDF. We acknowledge that the usage of the phrase was misleading, we have now rephrased the sentence.

** Finally, section (2) talks about the systems handling of false positives but does not mention false negatives at all which are equally irritating to the user. What is the proposed approach here?*

You have hit on a tricky problem: how to report a false negative. Without an easy way for

people to, for example, highlight text and add a comment, it will be a cumbersome task to report false negatives. It is possible that solutions such as hypothes.is could help here, especially given that both SciLite and hypothes.is use the Web Annotation standard; we plan to explore these options in the future.

** 3. "Everyone preferred annotations to be turned off by default." - I think thats a fairly important statement, did you ask why?*

Based on the user feedback, we find that annotations highlighted (in different colours) by default are distracting to the user. Hence, the users preferred to choose the type of annotations that are of interest to them.
We have expanded the section explaining the process of testing SciLite with the users and describe the feedback we received.

** The link to the Sparql end point http://www.ebi.ac.uk/europepmc/rdf/sparql provides only a default view and it is left to the user to figure out the schema and queries necessary. I am not sure of the best mechanism (supplementary data or footnote?) but an example query or two would be very helpful to get started*

We agree with the point being made, we have now added Supplementary material with details of relations and classes used along with sample queries. We also plan on releasing a RESTful API on the annotations within the next few months

** In general, figures could do with a bit more description as to what they are showing.*

We have extended the figure legends to be more descriptive.

Competing Interests: No competing interests were disclosed.

Reviewer Report 05 January 2017

https://doi.org/10.21956/wellcomeopenres.10999.r18949

© 2017 Hirschman L. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Lynette Hirschman
Biomedical Informatics, MITRE Corporation, Bedford, MA, 01730, USA

The article describes the SciLite project and capability; SciLite opens exciting new possibilities by supporting the automated tagging (and indexing) of biological entities. The article presents the background behind the development of these capabilities and a description of both the capabilities and the underlying components.

Overall, the article provides a clear presentation of the rationale for SciLite, a useful description of the underlying approach and architecture, and a good bibliography referencing many of the key works related to automated extraction of bio-entities. However, there are several places which need clarification or elaboration, including 1) whether SciLite can extract the “essence” of an article; 2) explanations of the contents (and abbreviations) of Figs 2 and 3 – and how they relate to the text descriptions in the article; 3) some discussion of how to handle multiple sources for the same information, and 4) a link to (a demo of) SciLite, to allow interested people to explore this new capability. In addition, there are some detailed minor questions/suggestions at the end of the review.

1. Capturing the “essence” of an article: In the description of capabilities, there is one statement that requires either rethinking or further explanation: “When such information [referring to text-mined, linked entities] is presented effectually it will aid users in identifying the main concepts, plausibly beginning to reduce the burden of extracting the essence of a given article.” (p. 2, col. 2, para. 3) There is a danger that tagging of entities in full text articles can overwhelm the user, rather than giving them a sense of the essence of an article. It may be true that tagging entities in an **abstract** might make the key entities jump out; or that providing a summary table listing the most frequently mentioned entities might provide such an overview of “main concepts” – but simply tagging many types of entities (especially if there are false positives in the tagging) can quickly overwhelm the user.

2. Explaining Figs 2 and 3: These figures each need a paragraph or so of text to explain what is being illustrated, and how to read the graph displayed. For Fig 2., it would be very helpful to see the actual annotation generated (in its textual context); it would be good to gloss the various abbreviations in the graph (what is oa as a prefix? What is orb? What are the urls and their suffixes #1-2s and #1-2t? What are the things in boxes? In the box on the lower left, Fig 2, why are there multiple quote marks: “CSF-1””induces expression…” and what do these mean?

   For Fig 3, what are the items in the box with dashed lines in the upper left? What do the various terms dc, dcterms, orb, oa, void, prov mean? What is the actual example GeneRIF that is being constructed? And pretty much the same request as for Fig 2 – describe the contents of the figure in text, so that the reader can follow what is being shown.

3. Handling of multiple sources of the same kind of information – given that there are multiple taggers for certain entities types (e.g., genes/proteins), will there be a capability for users to choose which tagger they want? Will there be a way for users to select high precision vs high recall? Will there be any attempt to create an ensemble system from multiple taggers?

4. Please provide a link so that interested readers can explore SciLite.

Minor comments/questions:

p.2 col. 1, para 1 middle – I couldn’t follow the second part of the sentence, starting with “where…”: “This workflow is highly interconnected and precise where the databases are both dependent and required for this practice.”

p.2 col. 2 para 1: suggest replacing “instituted” with “fostered”.

p.2. col. 2 para 2 top- suggest removing the first clause (it’s redundant) – e.g., remove “Having described the importance of information extraction by using text-mining methods in the context of aiding manual curation” – and just start the paragraph with “The other essential aspect…”

p.2, col. 2 para 2 end: suggest changing “and integrate data” to “and integration of data”

p. 2, col. 2, para 3 l 2: suggest changing “mechanism toward” to “mechanism to support”

p.2. col. 2 last para: suggest changing “an aim to leverage text-mining solutions towards” to “an aim of leveraging text-mining solutions for”

p. 5, section 2.5 on User Interface – would it be appropriate to cite the Reflect project here, as a groundbreaking application to provide enriched metadata about objects in text? [Reflect: augmented browsing for the life scientist E Pafilis, SI O'Donoghue, LJ Jensen, H Horn, M Kuhn… - Nature biotechnology, 2009)

p. 7, col. 2, para 2 suggest changing “platform, this offers” to “platform, which offers”

p. 9, col 1 Author Contributions – insert “the” before SciLite text mining pipeline”

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 04 Jul 2017

Aravind Venkatesan

Dear Lynette Hirschman,
Thank you for providing your inputs with regard to our paper. We have reworked the manuscript, addressing your suggestions.
Overview of the changes made are as follows:

- Section 1. Introduction: The section has being restructured and now acknowledges contributions made by other annotation tools.
- Section 2. Methods:
  - Sub-section 2.1 (Architecture) now provides an overview of the architecture of SciLite (includes new figure 1)
  - Sub-section 2.2 (Annotation types) has been expand to include new annotations types (gene-disease relationships and manually curated protein-protein interactions).
  - Sub-section 2.3 describes the Web Annotation model (includes new figures: 2 and 3).
  - Sub-section 2.4 (User interface) has been reworked to provide concise description of the highlighting process.

- Sub-section 2.5 now includes a new figure (4) that provides an overview on the feedback mechanism.
- Section 3. Discussion: The section now provides details on how the user research was conducted. The section has been restructured with two new sub sections:
  - 3.1 – Engagement with text-mining community
  - 3.2 – Future directions.
- Supplementary material: The supplementary material provides details on the classes and relations used to model the annotations as RDF along with sample SPARQL queries.

We think that the manuscript is much improved as a result of your feedback. Please see a point-by-point response, below:

**Capturing the "essence" of an article: In the description of capabilities, there is one statement that requires either rethinking or further explanation: "When such information [referring to text-mined, linked entities] is presented effectually it will aid users in identifying the main concepts, plausibly beginning to reduce the burden of extracting the essence of a given article." (p. 2, col. 2, para. 3) There is a danger that tagging of entities in full text articles can overwhelm the user, rather than giving them a sense of the essence of an article. It may be true that tagging entities in an abstract might make the key entities jump out; or that providing a summary table listing the most frequently mentioned entities might provide such an overview of "main concepts" – but simply tagging many types of entities (especially if there are false positives in the tagging) can quickly overwhelm the user.*

We understand and agree with the point being made. Highlighting different annotation types could serve as an indicator (referred to in the text as "essence") with respect to the key concepts described in a given paper. We have now rewritten the paragraph to avoid confusion.

**Explaining Figs 2 and 3: These figures each need a paragraph or so of text to explain what is being illustrated, and how to read the graph displayed. For Fig 2., it would be very helpful to see the actual annotation generated (in its textual context); it would be good to gloss the various abbreviations in the graph (what is oa as a prefix? What is orb? What are the urls and their suffixes #1-2s and #1-2t? What are the things in boxes? In the box on the lower left, Fig 2, why are there multiple quote marks: "CSF-1""induces expression..." and what do these mean?*
For Fig 3, what are the items in the box with dashed lines in the upper left? What do the various terms dc, dcterms, orb, oa, void, prov mean? What is the actual example GeneRIF that is being constructed? And pretty much the same request as for Fig 2 – describe the contents of the figure in text, so that the reader can follow what is being shown.

We agree that the figures could be confusing to readers with limited exposure to RDF concepts. We now briefly describe the Web Annotation Model (Sub-section 2.3), further, we have simplified Figures 2 and 3. The figure legends now describes the sample model with links that point to the source articles that contains the exemplified annotations.

**Handling of multiple sources of the same kind of information – given that there are multiple taggers for certain entities types (e.g., genes/proteins), will there be a capability for users to*

choose which tagger they want? Will there be a way for users to select high precision vs high recall? Will there be any attempt to create an ensemble system from multiple taggers?

SciLite can host similar semantic types from multiple sources. The platform allows users to choose relevant sources of their interest. However, it is unlikely that this would be of interest to the wider scientific community. Ideally SciLite would consume annotations through systems like BeCalm, allowing users access to benchmarked “high quality” annotations.

**Please provide a link so that interested readers can explore SciLite.*

SciLite platform is integrated in Europe PMC and can be seen in action on any CC-BY or CC-BY-NC article. We have provided links to the illustrated examples in the figure legends of 2, 3 and 5 and a link (in sub-section 2.1 and section 5) that lists all the CC-BY and CC-BY-NC articles.

**Minor comments/questions:*
p.2 col. 1, para 1 middle – I couldn’t follow the second part of the sentence, starting with “where…”: “This workflow is highly interconnected and precise where the databases are both dependent and required for this practice.”
p.2 col. 2 para 1: suggest replacing “instituted” with “fostered”.
p.2. col. 2 para 2 top- suggest removing the first clause (it’s redundant) – e.g., remove “Having described the importance of information extraction by using text-mining methods in the context of aiding manual curation” – and just start the paragraph with “The other essential aspect…”
p.2, col. 2 para 2 end: suggest changing “and integrate data” to “and integration of data”
p. 2, col. 2, para 3 l 2: suggest changing “mechanism toward” to “mechanism to support”
p.2. col. 2 last para: suggest changing “an aim to leverage text-mining solutions towards” to “an aim of leveraging text-mining solutions for”
p. 5, section 2.5 on User Interface – would it be appropriate to cite the Reflect project here, as a groundbreaking application to provide enriched metadata about objects in text? [Reflect: augmented browsing for the life scientist E Pafilis, SI O'Donoghue, LJ Jensen, H Horn, M Kuhn… - Nature biotechnology, 2009)
p. 7, col. 2, para 2 suggest changing “platform, this offers” to “platform, which offers”
p. 9, col 1 Author Contributions – insert “the” before SciLite text mining pipeline”

We have made all the minor correction and added paragraphs to review the state of the art (including the Reflect application).

Competing Interests: No competing interests were disclosed.