SOFTWARE TOOL ARTICLE

REVISED SciLite: a platform for displaying text-mined annotations as a means to link research articles with biological data [version 2; referees: 2 approved, 1 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

v2 **First published:** 12 Dec 2016, **1**:25 (doi: 10.12688/wellcomeopenres.10210.1)
Latest published: 10 Jul 2017, **1**:25 (doi: 10.12688/wellcomeopenres.10210.2)

Abstract

The tremendous growth in biological data has resulted in an increase in the number of research papers being published. This presents a great challenge for scientists in searching and assimilating facts described in those papers. Particularly, biological databases depend on curators to add highly precise and useful information that are usually extracted by reading research articles. Therefore, there is an urgent need to find ways to improve linking literature to the underlying data, thereby minimising the effort in browsing content and identifying key biological concepts.
As part of the development of Europe PMC, we have developed a new platform, SciLite, which integrates text-mined annotations from different sources and overlays those outputs on research articles. The aim is to aid researchers and curators using Europe PMC in finding key concepts more easily and provide links to related resources or tools, bridging the gap between literature and biological data.

Open Peer Review

Referee Status: ✓ ? ✓

	Invited Referees		
	1	**2**	**3**
REVISED **version 2** published 10 Jul 2017	✓ report		✓ report
version 1 published 12 Dec 2016	? report	? report	? report

1 **Lynette Hirschman**, MITRE Corporation, USA

2 **Lee Harland**, Scibite Limited, UK

3 **Diana Maynard**, University of Sheffield, UK

Discuss this article

Comments (0)

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

Author roles: Venkatesan A: Formal Analysis, Investigation, Software, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Kim JH**: Conceptualization, Formal Analysis, Methodology, Software, Validation; **Talo F**: Conceptualization, Investigation, Methodology, Software; **Ide-Smith M**: Investigation, Methodology, Project Administration, Validation, Visualization; **Gobeill J**: Formal Analysis, Investigation, Software, Validation; **Carter J**: Formal Analysis, Investigation, Software, Validation; **Batista-Navarro R**: Formal Analysis, Methodology, Software, Validation; **Ananiadou S**: Investigation, Supervision, Validation; **Ruch P**: Conceptualization, Investigation, Supervision, Validation, Writing – Review & Editing; **McEntyre J**: Conceptualization, Funding Acquisition, Supervision, Validation, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

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 2; referees: 2 approved, 1 approved with reservations]** Wellcome Open Research 2017, **1** :25 (doi: 10.12688/wellcomeopenres.10210.2)

Grant information: Funding for Europe PMC is provided by twenty-seven European-based funders of life science research: http://europepmc.org/Funders/, under Wellcome Trust grants [098321] and [108758] awarded to EMBL-EBI. This work was also supported by 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.

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

REVISED **Amendments from Version 1**

The current version of the article incorporates the suggestions made by the reviewers. The changes are as follows:

Section:

- Section 1. Introduction: The section has being reworked.
- Section 2. Methods:
  - Sub-section 2.1 now provides an overview of the architecture of SciLite
  - Sub-section 2.2 has been expanded to include new annotations types (gene-disease relationships and manually curated protein-protein interactions).
  - Sub-section 2.3 provides an overview of the Web Annotation model.
  - Sub-section 2.4 has been reworked.
  - Sub-section 2.5 now includes a new Figure (4) that provides an overview on the feedback mechanism.
- Section 3 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.

Figures:
All the figures previously included have been updated and a new figure (Figure 4) has been introduced.

Supplementary material: The supplementary material provides details on the classes and relations used to model the annotations as RDF along with sample SPARQL queries.

See referee reports

1. Introduction

Biological databases underpin life sciences research, aiding scientists in the process of knowledge discovery. 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 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 bioentities 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 software 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. Curation is a time consuming and challenging task, as the bio-entities and biological concept descriptions sought be spread within an article or over multiple articles. For instance, curators may have to identify interacting partners for a large set of proteins or protein complexes, categorizing them based on evidences for the interaction, 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 curation activities and share new methods. Considering the exponential growth in data, there is increasing pressure for curation to be better supported by sophisticated computational approaches to make curation efforts sustainable in the long term.

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, and retrieve entities and relationships, such as 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 to triage articles for curation (Druzinsky *et al.*, 2016; Van Auken *et al.*, 2012). Collaboration between curators and the text-mining community has been fostered 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 more recently been established with an aim to orchestrate various text-mining efforts.

In addition to the above examples, other tools annotate entities in scientific articles and link them to the corresponding data sources. For instance, Reflect (O'Donoghue *et al.*, 2010) and EXTRACT (Pafilis *et al.*, 2016) are tools that use real-time tagging and augmented browsing approach to tag bio-entities, such as genes, proteins and small molecules described in papers. Whereas, Utopia documents (Attwood *et al.*, 2010), is a desktop application that links explicit and implicit information on static (PDF) versions of articles to the corresponding online resources. Furthermore, PubAnnotation is an online resource developed by the Data Base Centre for Life Sciences (DBCLS), which serves as a repository for text annotations. The tool uses text alignment functions to create new annotations or map existing annotations on scientific articles.

We have to acknowledge that articles are read and discovered in different contexts, for example, via locally saved PDFs, literature indexing services and databases, and via publisher's websites. While the above listed tools can support information extraction and in some cases integration, the annotations generated are hard to share across different platforms. In part this is due to the lack of a common standard to exchange annotations. Although formats like BioC (Comeau *et al.*, 2013) and Biological Expression Language (BEL) are used to represent text-mined outputs, they are highly specific to the domain and their uptake is limited beyond the life sciences text mining community. Recently, the World Wide Web Consortium (W3C) proposed the Web Annotation Data Model as a standard to share annotations across different platforms. Based on the Resource Description Framework (RDF), the model can represent text quotes/fragments, links, video segments and images, which are located in a document via a prefix-suffix tagging approach that offers opportunities to mix and show both human and machine-generated annotations on any instance of the same content.

In this article, we describe a new platform, SciLite, developed as part of Europe PMC (Europe PMC Consortium, 2015) 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 text-mining advances from the research community, bringing the results to a broader audience of readers and developers who might use the results to address additional challenges around information retrieval, visualization and integration. In the context of ELIXIR, the European Infrastructure for life sciences data, of particular interest is the reuse of text mining outputs to make deeper links between the literature and biological data resources, providing more seamless navigation and clear evidence behind curatorial statements.

2. Methods

2.1. Architecture

SciLite is a platform that allows text-mined annotations from any provider to be highlighted on scientific articles. The platform consolidates various text-mined annotations (see section 2.2) and displays those annotations on full text research articles browsed within Europe PMC webpages. Currently, SciLite operates on full text articles with the license type: CC0, CC-BY, or CC-BY-NC (~900000 articles at the time of writing).

Annotations received from contributors are modelled according to the W3C standard Web Annotation Data Model (see section 2.3 for more information). The annotations are stored in a MongoDB database. MongoDB is performant and fetching annotations for a given PMCID is straightforward. Additionally, annotations are stored as RDF in a triple store (OpenLink Virtuoso version 7.2). RDF is more powerful for graph-based queries; offering an opportunity to explore SciLite annotations in conjunction with other RDF graphs in the Linked Data cloud. The SciLite RDF endpoint is publicly available (see section 5) and contains over 1.4 billion triples (as of May 2017). Details on classes, relations and sample queries are available in the Supplementary material. Furthermore, we plan on providing a public RESTful API within the next few months.

SciLite is very flexible regarding the frequency of data deposition by contributors: anything from a static dataset to daily updates can be accommodated. Once stored in the database, the SciLite application uses an API to retrieve annotations for a given article. The article view provides the opportunity for reader feedback on the quality of annotations, which can be reported back to the source text-mining algorithm for potential future improvement (see section 2.5 for more information). Figure 1 gives an overview of the workflow by which text-mined annotations can be accommodated and viewed on the SciLite platform.

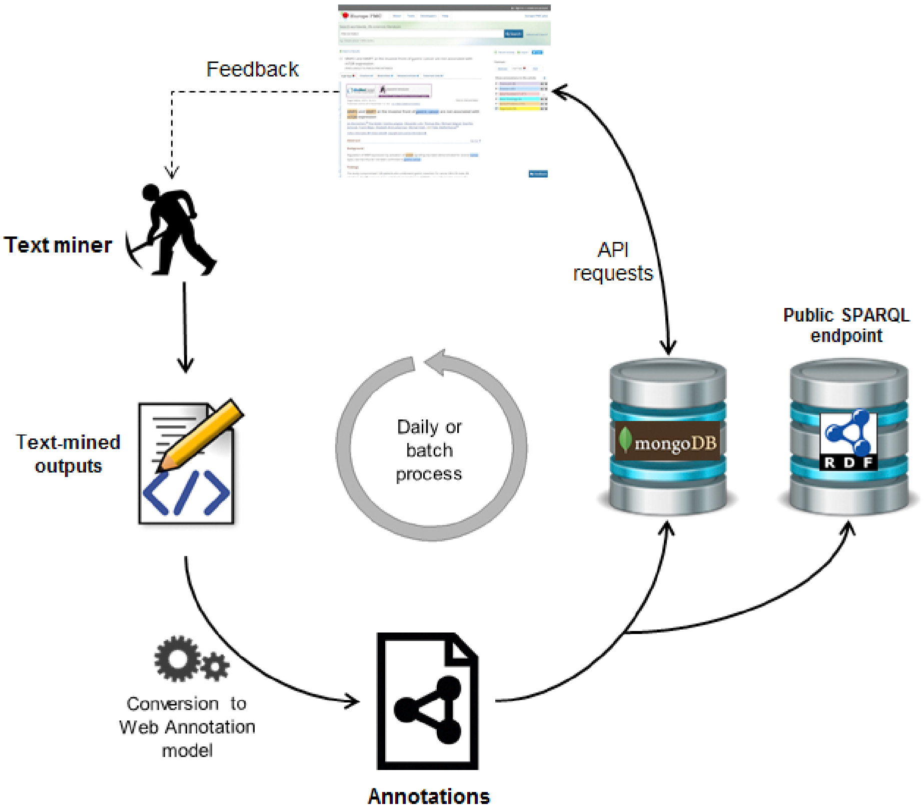


Figure 1. Overview of how text mining results are incorporated into SciLite.

Should it be desirable, the EMBL-EBI Embassy Cloud is available for groups to operate their text-mining workflows on a daily basis.

2.2. Annotation types

- **Named entities:** Annotations supplied by the Europe PMC text-mining pipeline identifies concepts, such as gene/protein names, organisms, diseases, GO terms, chemicals and accession numbers. This is achieved through a cascade of three modules: section tagger (Kafkas *et al.*, 2015), sentence splitter and named entity taggers, a module that applies the dictionary-based approach (Rebholz-Schuhmann *et al.*, 2008) and a machine-learning based filter (Chang *et al.*, 2007), for filtering out potential false positives in annotations.

- **Biological events:** 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, over 400 phosphorylation events from this set are included.

- **Relationships:**Target-disease associations are supplied by two providers:
    - ○ Open Targets (Koscielny *et al.*, 2017): Open Targets is a platform for accessing potential drug targets associated with disease. The dataset included in SciLite currently contains over 2 million text mined gene-disease associations.
    - ○ DisGeNET (Piñero *et al.*, 2015): The DisGeNET platform contains collections of genes and gene variants associated with human diseases. The platform contains data from various sources that includes curated sources, animal models and scientific articles. From this, a subset of 7000 gene-disease associations are integrated.

- **Text phrases:**
    - ○ Gene Reference into Function (GeneRIF): GeneRIF refers to the function of genes that are extracted from articles by MeSH indexers and added to the NCBI's Entrez Gene records. GeneRIFs can be simple "cut-and-paste" text snippets from abstracts or full text, but sometimes are a more complex synthesis of text fragments. The text mining group at the Swiss Institute of Bioinformatics (SIB), Geneva, Switzerland has mapped GeneRIFs back onto article full text and supplied the results to SciLite 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 over 32,000 GeneRIF statements.
    - ○ Molecular interactions: Text snippets describing protein-protein interactions extracted manually from scientific articles were supplied by IntAct (Orchard *et al.*, 2014). Over 100 high-quality curated statements have been included in SciLite.

2.3. Representing annotations in SciLite

The text-mined outputs are represented in the Web Annotation Data Model. The model consists of a *Target* - the text describing an entity and a *Body* – a link (to the data source) for the tagged entity; the representation may vary according to the type of annotation. For the current set of annotations, we adopted the Text Quote Selector and the Fragment Selector models. The former is used to represent named entities and biological events (see Figure 2) and the latter to represent relationships and text phrases (see Figure 3).

2.4. User interface

To make annotations available to readers, the MongoDB database is queried using API requests fetching all the relevant annotations for a given article. The retrieved annotations are highlighted (colour coded) on the Europe PMC website (see Figure 5). Additionally, the annotations are interactive; clicking the highlighted annotation opens a popup-box containing additional information about the annotation (see Figure 6), such as source of annotation and link to the related database. The algorithm used to highlight annotations consists of the following steps:

- Retrieving all the annotations for a specific PMCID from the annotation database using an API request.

- Sorting the annotations from the API response according to their position in the text (ascending order of occurrence). This is done in order to optimize the performance of the searching process in the article text.

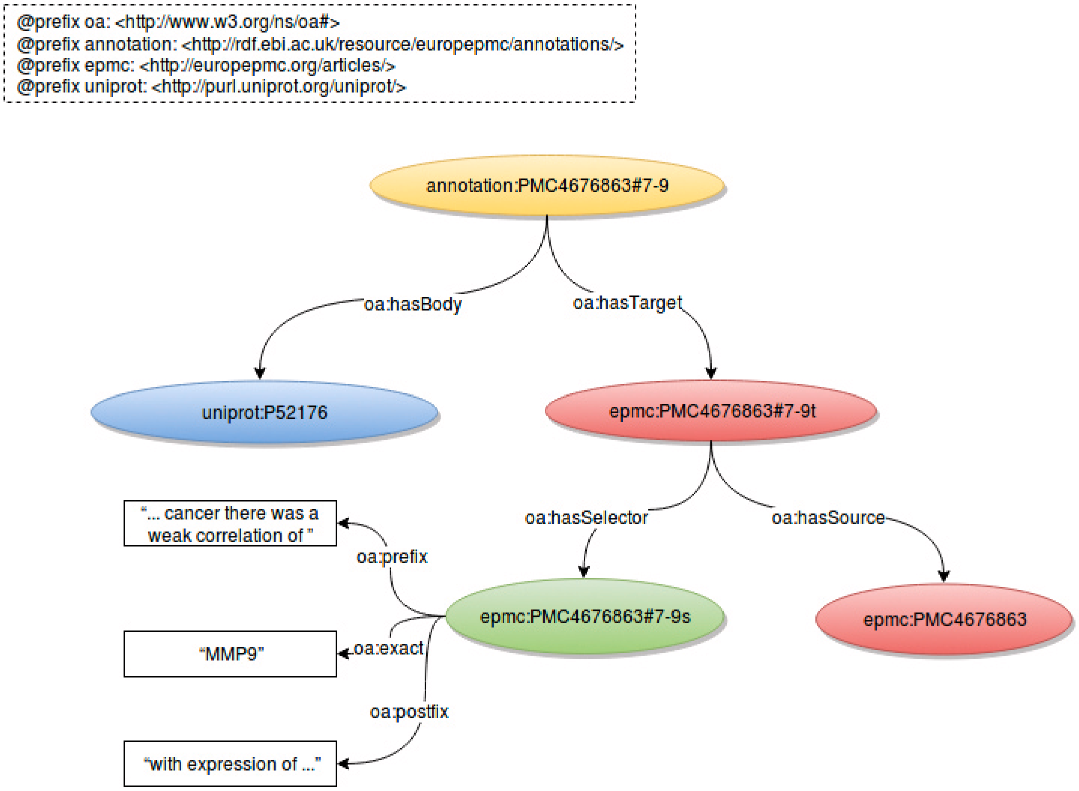
- A listener is associated with each annotation tag to display the corresponding information in the popup window.

2.5. Improving annotation accuracy

To reduce false positives, we have set up a semi-automated process that enables a user to report an annotation error, which triggers the removal of that specific annotation. These reports can be collected periodically to improve the text-mining workflows (see Figure 4). The handling of error reports involves two steps:

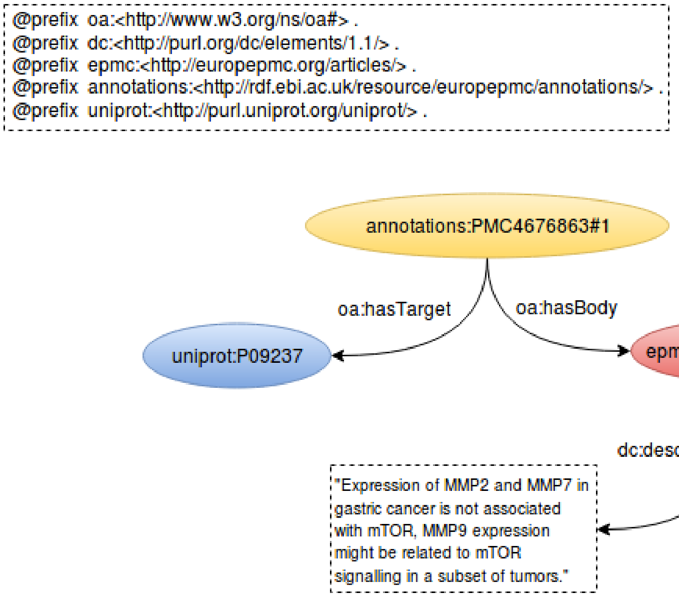
- Once an error report is received from a user the particular annotation is deleted from the annotation databases. This is a daily procedure that engenders user trust.

- The report may be used to refine the text-mining algorithms. Often ambiguous terms are reported that are incorrect in the context of the sentence. This often occurs with three letter entity names. For instance, *EPO,* would be tagged as Erythropoietin protein while the sentence could be referring to the *European Patent Office (EPO)*. This step is carried out over a longer period of time, allowing the providers to examine the reports prior to the inclusion of those exceptions in their algorithms.



```
@prefix oa: <http://www.w3.org/ns/oa#>
@prefix annotation: <http://rdf.ebi.ac.uk/resource/europepmc/annotations/>
@prefix epmc: <http://europepmc.org/articles/>
@prefix uniprot: <http://purl.uniprot.org/uniprot/>
```

Figure 2. The figure illustrates a sample annotation of protein *MMP9* described in an article (PMC4676863): the figure lists the vocabularies used to represent the text-mined annotations. The annotation consists of a link for the tagged entity (Body - UniProt: P52176) and the mentions of the entity (Target) in the text snippet. The text is represented by: prefix – the text that occurs before the tagged entity; exact – tagged entity itself (*MMP9*); and postfix – the text snippet that occurs after the tagged entity.



```
@prefix  oa:<http://www.w3.org/ns/oa#> .
@prefix  dc:<http://purl.org/dc/elements/1.1/> .
@prefix  epmc:<http://europepmc.org/articles/> .
@prefix  annotations:<http://rdf.ebi.ac.uk/resource/europepmc/annotations/> .
@prefix  uniprot:<http://purl.uniprot.org/uniprot/> .
```

"Expression of MMP2 and MMP7 in gastric cancer is not associated with mTOR, MMP9 expression might be related to mTOR signalling in a subset of tumors."

Figure 3. An illustration of a sample GeneRIF (gene function) annotation (PMC4676863): the figure lists the vocabularies used to represent the annotation. The annotation consists of: Body - text phrase about protein mTOR and a target - data source link for the described protein (UniProt: P09237).

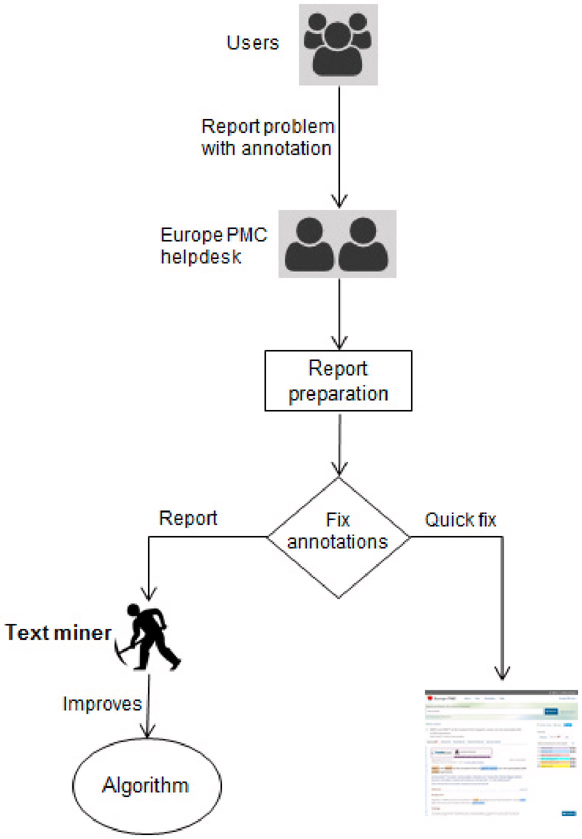


Figure 4. An illustration of the semi-automated feedback mechanism to improve annotations. Erroneous annotations reported by users is used to prepare a report by the helpdesk at Europe PMC. This report is used to perform: a) a quick fix by deleting the particular annotation; b) further the reports are used to refine the text-mining algorithms in the longer term.

The current process does not take into account false negatives, i.e. when an annotation was missed by an algorithm. A mechanism to handle false negatives would be a desirable future development.

3. Discussion

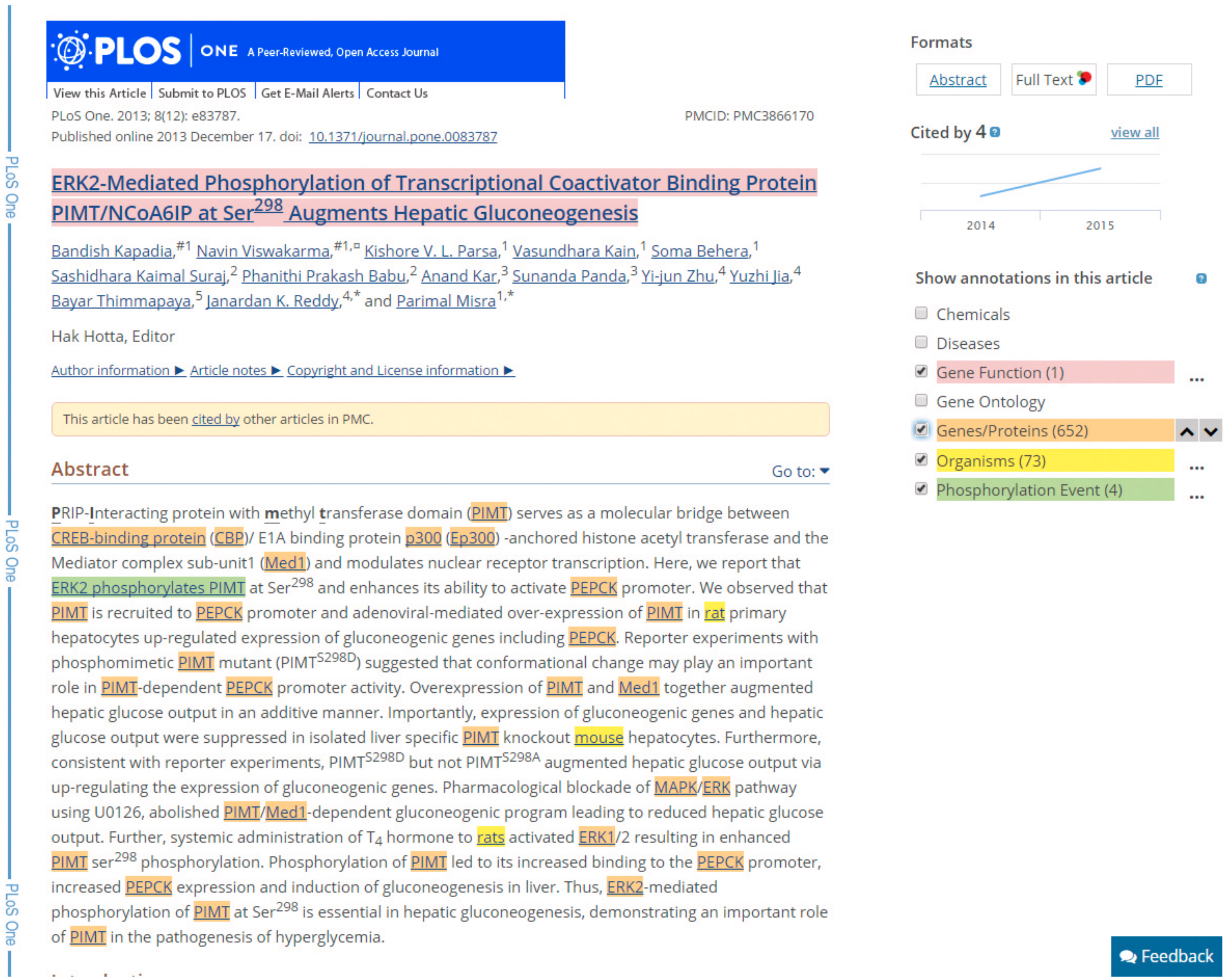
We have adopted a user-driven approach towards the development of SciLite, conducting user research on cognitive and functional aspects of the platform. We have tested SciLite with users that have a range of backgrounds over different geographical locations. A total of 17 users participated, including curators, senior researchers, clinicians and doctoral students. The test sessions were unguided, that is, the users were observed for their behaviour whilst (1) visually scanning an article of interest, (2) discovering annotation types and features and (3) reacting to the annotations. At the end of the session, the subjects were asked to rate the feature:

- Would they like to use such a feature?
- Was the feature easy to use?
- Trust in data quality and confidence in using the feature.

Overall the feedback was positive. While users' annotation preferences differed based on their background, most people found at least three annotation types (e.g.: gene/proteins, diseases, organisms) useful. Everyone preferred annotations to be turned off by default as the appearance of different colours immediately on viewing the article was distracting. Some users commented that viewing different annotation types highlighted in close proximity in the text was useful, as it suggested a possible relationship between those terms e.g. chemical-disease. In an earlier version of SciLite, users found it hard to locate the highlighted terms in the text. In response to this feedback, we introduced a term navigation feature (up/down buttons) that allows users browse through the highlighted terms quickly. We also found that inaccuracies affects user trust. This observation resulted in the development of the feedback mechanism described in section 2.5. Other useful suggestions addressed the information that appears in the popup window, range of annotation types (e.g.: cell types, software), and the ability to refine a search based on the type of highlighted annotation. The receipt of user feedback and subsequent improvement of SciLite in response is an iterative process that will continue in the future.

3.1. Engagement with the text mining community

The goal of the BioCreative v.5 BeCalm task was to provide the means to benchmark text mining services from different providers, addressing not only the efficacy of the text mining


PLoS One
PLOS ONE A Peer-Reviewed, Open Access Journal
View this Article | Submit to PLOS | Get E-Mail Alerts | Contact Us
Formats
Abstract
Full Text
PDF
Cited by 4
view all
2014
2015
Show annotations in this article
☐ Chemicals
☐ Diseases
☑ Gene Function (1)
☐ Gene Ontology
☑ Genes/Proteins (652)
☑ Organisms (73)
☑ Phosphorylation Event (4)
Go to:
Feedback

PLoS One. 2013; 8(12): e83787.
Published online 2013 December 17. doi: 10.1371/journal.pone.0083787

PMCID: PMC3866170

ERK2-Mediated Phosphorylation of Transcriptional Coactivator Binding Protein PIMT/NCoA6IP at Ser^{298} 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 Misra1,*

Hak Hotta, Editor

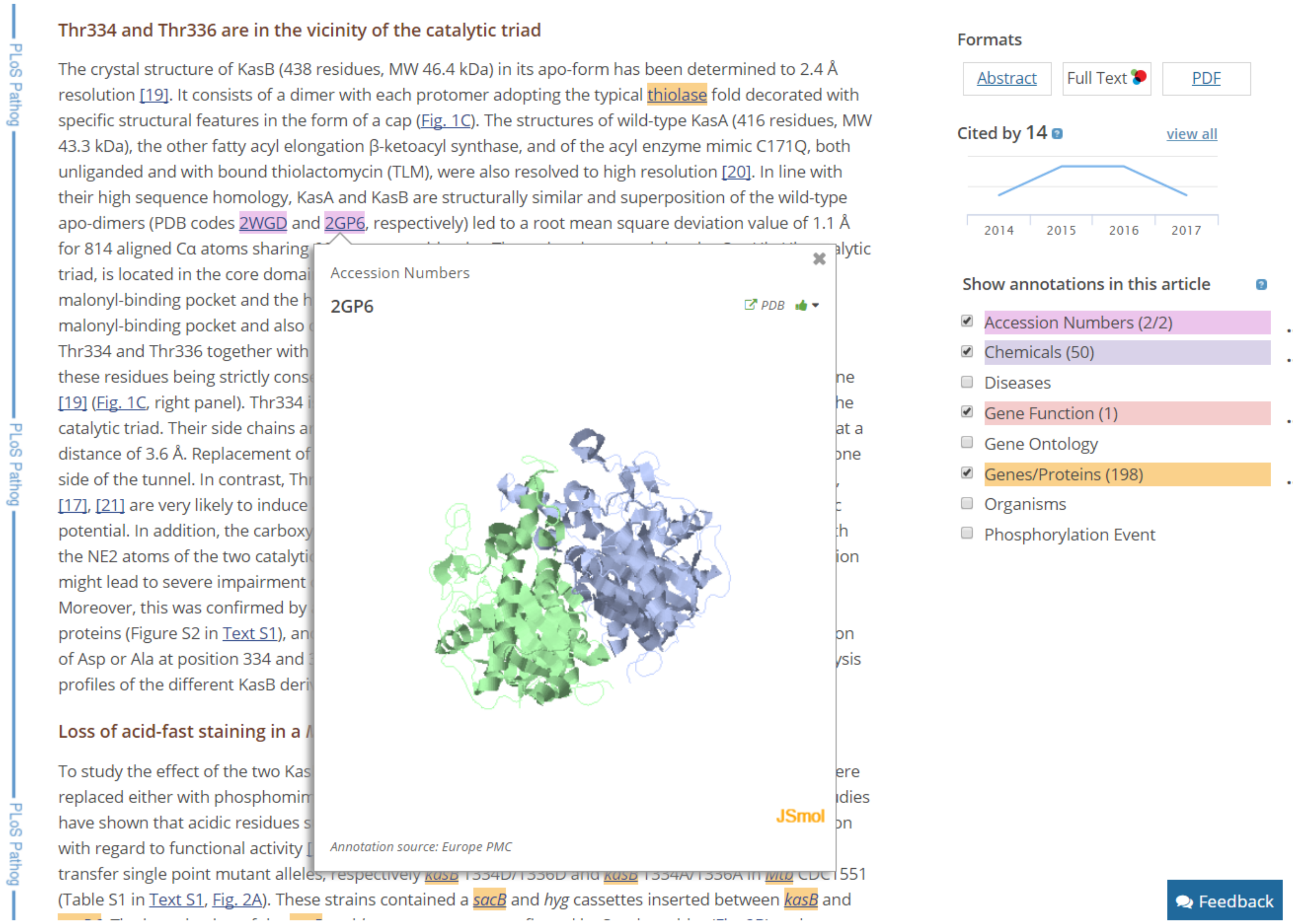
Author information ▶ Article notes ▶ Copyright and License information ▶

This article has been cited by other articles in PMC.

Abstract

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, increased PEPCK expression and induction of gluconeogenesis in liver. Thus, ERK2-mediated phosphorylation of PIMT at Ser^{298} is essential in hepatic gluconeogenesis, demonstrating an important role of PIMT in the pathogenesis of hyperglycemia.

Figure 5. The screenshot shows the front-end rendering of various annotation types for an article on Europe PMC.



Thr334 and Thr336 are in the vicinity of the catalytic triad

The crystal structure of KasB (438 residues, MW 46.4 kDa) in its apo-form has been determined to 2.4 Å 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 alytic
triad, is located in the core domai
malonyl-binding pocket and the h
malonyl-binding pocket and also
Thr334 and Thr336 together with
these residues being strictly cons ne
[19] (Fig. 1C, right panel). Thr334 i he
catalytic triad. Their side chains a at a
distance of 3.6 Å. Replacement of one
side of the tunnel. In contrast, Th
[17], [21] are very likely to induce
potential. In addition, the carboxy h
the NE2 atoms of the two catalyti ion
might lead to severe impairment
Moreover, this was confirmed by
proteins (Figure S2 in Text S1), an on
of Asp or Ala at position 334 and ysis
profiles of the different KasB deri

Loss of acid-fast staining in a

To study the effect of the two Kas ere
replaced either with phosphomim dies
have shown that acidic residues s on
with regard to functional activity
transfer single point mutant alleles, respectively kasB T334D/T336D and kasB T334A/T336A in Mtb CDC1551
(Table S1 in Text S1, Fig. 2A). These strains contained a sacB and hyg cassettes inserted between kasB and

Accession Numbers

2GP6

Annotation source: Europe PMC

Formats

Cited by 14

Show annotations in this article

- ☑ Accession Numbers (2/2) ...
- ☑ Chemicals (50) ...
- □ Diseases
- ☑ Gene Function (1) ...
- □ Gene Ontology
- ☑ Genes/Proteins (198) ...
- □ Organisms
- □ Phosphorylation Event

Figure 6. A screenshot showing the 3D molecular structure for a given PDB accession number.

itself, but also technical aspects of services offered. Ideally, the outcomes of BeCalm-based benchmarking would feed into SciLite, allowing "best of class" annotations to be made available in Europe PMC interfaces and APIs. Not only would this approach deliver state-of-the-art text mining results into a widely used interface, it would also counter a potential future user interface challenge of too many repeat annotations from different groups layered on articles. While this is technically possible, it could lead to a performance burden and is unlikely to be of interest to the wider scientific community.

Further extended collaborations within the text-mining community will address the challenge of providing different annotation types that serve different user needs. We welcome contributions encourage them to share annotations on SciLite. The SciLite participation page, which provides details on data requirements, submission formats and examples for interested groups.

3.2. Future development

Highlighting biological terms enables skim-reading of articles and the links are useful for verification of the term, but it is only one application of the text mining. We envision that the annotations store and basic SciLite visualisations described here are a basis for the development of further applications by third parties that have the potential to improve full text searching, filtering and integration with biological data. An initial example of this application layer is the use of the BioJS framework to display 3D interactive views of molecular structures identified through text mined PDB accession numbers. In this case, for a given text mined PDB ID, SciLite calls the BioJS module that fetches the structure coordinate information and another that then renders the coordinates using JSmol (see PMC4014462 and Figure 6).

4. Conclusion

As there is no let-up in the production of data and publication of articles, the need to find programmatic ways to bridge the gap between literature and data increases. SciLite is an initial step in this direction, but the impact of SciLite will be more pronounced with community-wide participation. We will continue to engage with the text mining and curation communities in particular to extend this work in the future.

5. Data and software availability

Annotation examples: Annotations available in SciLite can be accessed here.

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 the 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 co-authored the manuscript. All the authors approved the final manuscript.

Competing interests

No competing interests were disclosed.

Grant information

Funding for Europe PMC is provided by twenty-seven European-based funders of life science research: http://europepmc.org/Funders/, under Wellcome Trust grants [098321] and [108758] awarded to EMBL-EBI. This work was also supported by 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. We would like to thank the reviewers for their constructive suggestions.

Supplementary material

Supplementary File 1: Classes and relations used to model the annotations as RDF, with sample SPARQL queries.

Click here to access the data.

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

Attwood TK, Kell DB, McDermott P, *et al.*: **Utopia documents: linking scholarly literature with research data.** *Bioinformatics.* 2010; **26**(18): i568–i574.
PubMed Abstract | Publisher Full Text | Free 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

Comeau DC, Islamaj Doğan R, Ciccarese P, *et al.*: **BioC: a minimalist approach to interoperability for biomedical text processing.** *Database (Oxford).* 2013; **2013**: bat064.
PubMed Abstract | Publisher Full Text | Free 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 (Oxford).* 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; **2012**: bas020.
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

Koscielny G, An P, Carvalho-Silva D, *et al.*: **Open Targets: a platform for therapeutic target identification and validation.** *Nucleic Acids Res.* Oxford University Press; 2017; **45**(D1): D985–D994.
PubMed Abstract | Publisher Full Text | Free Full Text

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

O'Donoghue SI, Horn H, Pafilis E, *et al.*: **Reflect: A practical approach to web semantics.** *Journal of Web Semantics.* 2010; **8**(2–3): 182–189.
Publisher Full Text

Orchard S, Ammari M, Aranda B, *et al.*: **The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases.** *Nucleic Acids Res.* 2014; **42**(Database issue): D358–63.
PubMed Abstract | Publisher Full Text | Free Full Text

Pafilis E, Buttigieg PL, Ferrell B, *et al.*: **EXTRACT: interactive extraction of environment metadata and term suggestion for metagenomic sample annotation.** *Database (Oxford).* 2016; **2016**: pii: baw005.
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

Open Peer Review

Current Referee Status: ✓ ? ✓

Version 2

Referee Report 18 July 2017

doi:10.21956/wellcomeopenres.13087.r24114

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

The authors have taken the reviewers' comments into consideration and I am happy with the revised version, other than a few minor English mistakes which should be fixed.

p.3 on set of relations -> on sets of relations
with an aim to orchestrate -> with the aim of orchestrating
Whereas, Utopia documents -> On the other hand, Utopia documents
publisher's websites -> publishers' websites

p.4 "performant" is not an English word. I suggest "perform well" or equivalent.

p.5 Annotations....identifies concepts -> Annotations..identify concepts
a subset....are integrated -> a subset...is integrated
onto article full-text -> onto full-text articles
EPO, would be tagged -> EPO would be tagged

p.7 allows users browse -> allows users to browse
inaccuracies affects -> inaccuracies affect

p.10 The last 2 sentences in paragraph 2 (beginning "We welcome contributions") are both ungrammatical. Also, encourage whom? Contributors? The second sentence has no main verb.

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Referee Report 18 July 2017

doi:10.21956/wellcomeopenres.13087.r18409

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

This version has addressed the major comments from the reviewers. I have only a few minor suggestions for further revision.

In Figures 2 and 3, what is the significance of the colors? They don't seem to be consistent across the figures. Also, the presentation order for the children of a given node type seems to differ in the figures -- in Fig 2, the "hasBody" link is the left-hand child and hasTarget is the right-hand child; in Fig 3, those are reversed. Presumably order is not relevant, but it would be easier on the reader if the presentation order were consistent. Also, it would be useful to provide (an excerpt) of the actual text that gave rise to the annotation.

Section 3: there is reference to a "feature" that subjects were asked to rate. What does "feature" refer to?

Section 3 5 lines from end of paragraph: I couldn't understand the sentenced "...that appears in the popup window, range of annotation types..."

p. 10, l. 6 -- "While this is technically possible..." -- what is technically possible? Providing repeat annotations?

A few minor editing suggestions:

- The term "evidence" is a collective noun in English; use of the plural sounds awkward.
- p. 3 Introduction col 1, line 4 from bottom: this is hard (for me) to parse - does this mean: ...descriptions sought can be spread within an article..."?
- p. 3 col 2., parag 2, line 10: set should be "sets"
- p. 3 col 2 parag 3 l. 5 _ "an augmented browsing approach"
- p. 4 col 1 line 5: "their uptake is limited outside of the life sciences..."
- p. 4 col 2 parag 2 graph-based queries, offering an ... (change semi-colon to comma)
- p. 5 col 1 line 6 from bottom: "A dedicated applications has been developed ... that combines extracts (should be singular)
- p. 7 col 2 inaccuracies affect user trust (should be plural form)

.

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Referee Report 17 January 2017

doi:10.21956/wellcomeopenres.10999.r19404

?

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 have read this submission. I 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, EMBL European Bioinformatics Institute, UK

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.

Referee Report 13 January 2017

doi:10.21956/wellcomeopenres.10999.r18964

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 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 have read this submission. I 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, EMBL European Bioinformatics Institute, UK

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-interac 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.

Referee Report 05 January 2017

doi:10.21956/wellcomeopenres.10999.r18949

? **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 have read this submission. I 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, EMBL European Bioinformatics Institute, UK

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.