RESEARCH ARTICLE

Understanding life sciences data curation practices via user research [version 1; peer review: 1 approved, 1 approved with reservations]

Aravind Venkatesan *, Nikiforos Karamanis*, Michele Ide-Smith, Jonathan Hickford, Johanna McEntyre

EMBL European Bioinformatics Institute, Cambridge, CB10 1SD, UK

* Equal contributors

V1 First published: 11 Sep 2019, 8(ELIXIR):1622
<https://doi.org/10.12688/f1000research.19427.1>
Latest published: 11 Sep 2019, 8(ELIXIR):1622
<https://doi.org/10.12688/f1000research.19427.1>

Abstract

Background: Manual curation is a cornerstone of public biological data resources. However, it is a time-consuming process that urgently needs supportive technical solutions in the face of rapid data growth. Supporting scalable curation is a part of the mission of the Elixir Data Platform. Thus far, we have established infrastructure capable of ingesting and aggregating text-mined outputs from multiple providers and making these available via an API. This public API is used by Europe PMC to display specific entities and relationships on full text articles (via the SciLite application).

Methods: To ensure that the future development of this infrastructure meets the needs of curators, we carried out a user research project to understand and identify common workflow patterns and practices via an observational study. Building on these outcomes, we then devised a curator community survey to more specifically understand which entity types, sections of a paper and tools are of top priority to address.

Results: The main challenges faced by curators included the following: a) There is a need for ways to prioritise and identify relevant papers for curation as the volume of literature is large; b) Finding specific information can prove difficult; quick ways of filtering articles based on specific entities, such as experimental methods, species and other important entities, such as genes, cell lines and tissue samples, are required; and c) Transferring information from the search/annotation tools to the various curation workflows was also challenging.

Conclusions: This study lays the foundation for identifying actionable items to orient the current infrastructure towards meeting the needs of curation community, by improving text-mined annotation quality and coverage and other engineering solutions; and reusing text-

Open Peer Review

Approval Status

	1	2
version 1		
11 Sep 2019	view	view

1. Lynette Hirschman , The MITRE Corporation, Bedford, USA
2. Cecilia N. Arighi , University of Delaware, Newark, USA

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

mined annotations and other metadata in Europe PMC for article triage. Furthermore, this study presents an opportunity to explore customisation of triage/ranking systems to suit different curation contexts.

Keywords

Database curation, User research, Observational study, Curator survey, Annotation Infrastructure, Europe PMC

This article is included in the **ELIXIR** gateway.

This article is included in the **EMBL-EBI** collection.

Corresponding author: Aravind Venkatesan (avenkat@ebi.ac.uk)

Author roles: **Venkatesan A:** Data Curation, Formal Analysis, Investigation, Project Administration, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Karamanis N:** Data Curation, Formal Analysis, Investigation, Methodology, Validation, Writing – Review & Editing; **Ide-Smith M:** Data Curation, Formal Analysis, Investigation, Methodology, Validation, Visualization; **Hickford J:** Supervision, Validation; **McEntyre J:** Conceptualization, Funding Acquisition, Investigation, Supervision, Validation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This research was part of ELIXIR-EXCELERATE project, funded by the European Commission under Grant Agreement number 676559.

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

Copyright: © 2019 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, Karamanis N, Ide-Smith M *et al.* **Understanding life sciences data curation practices via user research [version 1; peer review: 1 approved, 1 approved with reservations]** F1000Research 2019, 8(ELIXIR):1622 <https://doi.org/10.12688/f1000research.19427.1>

First published: 11 Sep 2019, 8(ELIXIR):1622 <https://doi.org/10.12688/f1000research.19427.1>

Introduction

Biological databases play a key role in knowledge discovery in life science research. A major contributor towards the maintenance of these databases is the process of manual curation. Curation is a high value task as experts carefully examine the relevant scientific literature and extract the essential information, such as biological functions and relationships between biological entities, generating the corresponding database records in a structured way. The advances in high-throughput technologies have resulted in tremendous growth of biological data, consequently increasing the number of research papers being published. As a result, the demand for high-quality curation that makes use of these resources has never been higher, but this demand can present challenges for curators in finding, and assimilating scientific conclusions described in the literature.

Text mining, machine learning and analytics promise to provide better ranking of reading lists, classification of articles, and identification of assertions with their biological context and evidence buried within the text of articles. To this end, many life science knowledgebases now include text mining (to varying degrees) in curation workflows. For example, databases, such as neXtProt^{1,2} and FlyBase³ have integrated text mining algorithms into their respective curation workflows to retrieve a ranked list of relevant articles and tag entities of interest. Furthermore, tools like PubTator⁴ and TextPresso⁵ are other examples of text mining tools that have been adopted by some curation communities. On the other hand, databases that mainly rely on manual curation, such as IntAct⁶ and DisProt⁷, are exploring possibilities to leverage text mining approaches to select articles for further curation.

Broadly speaking, the curation community recognises the potential of text mining in article triage and the identification of entities/concepts for curation. Nevertheless, text mining pipelines adopted thus far have been engineered to cater to specific domains or projects and wide uptake is lacking; curators often continue to use manual curation methods. This mainly stems from the wide variety of very precise information required by curators. The challenge is therefore to produce robust systems that both address the immediate and specific needs of curators as well as scale across multiple curation groups. In order to do this, we need to know the immediate challenges faced by curators with respect to selection and prioritisation of articles to curate. A clear understanding of the requirements will help build new systems and/or re-orient existing systems that cater to the needs of the curation community.

In this report we describe the outcomes of a user research project, conducted to understand curation practices and priorities for article selection. The project comprised of two parts, a) an observational study, to understand how curators proceed with selecting articles to curate, to identify commonalities in curator requirements; and b) a community survey, to specifically identify the immediate priorities of curators, such as entity types and sections of interest in an article, to name a few. The aim of this study is to identify specific actions for the Elixir Data

Platform in the future, optimising and extending existing systems and infrastructural components. In the subsequent sections we present the main findings from our investigation.

Methods

Observational study

We initially drafted an interview guide (list of questions available as *Extended data*⁸) and a preliminary curator persona, reflecting our initial hypotheses about curators and their work practices. Following this we selected five different curation teams. The selection criteria was based on the type of curation the group was involved with, i.e., extracting biological information from scientific literature and integrating it into a biological database, contrary to groups that process raw data submissions (such as sequencing data). Out of these, two teams were based at EMBL-EBI, the other three teams were based in Norway, Switzerland and Italy. For teams that were situated at EMBL-EBI, the interviews were held in person and for the other teams the interviews were conducted over conference calls. The sessions were conducted over a period of two months: between March and April 2018. We observed three project leaders and four curators from the selected teams. The participants belong to teams that focused on curating very specific experimental evidence primarily on proteins such as protein-protein interactions, the role of a protein in a complex, protein disruption, human protein functions and transcription factor regulation. One team is focussed on the annotation of human genes relevant to a particular disease and another one on curating publications reporting associations of genetic variants with diseases.

For each session we followed an iterative user research process^{9,10} (as outlined in [Table 1](#)):

- The participants were asked to proceed with their daily curation work¹¹ and were observed on how:
 - they select the entities (either come from a spreadsheet or partially curated data record) they wish to curate,
 - perform searches (including the query parameters) to retrieve the initial set of publications,
 - the criteria they used to either discard or select an article,
 - the information from the selected article is transferred to the respective curation platform.
- Using the “What? So What? Now What?” method¹, we transcribed our notes from these sessions and identified the most important observations, patterns and their implications.

Additionally, we further carried out two follow-up interviews with one project lead and one curator from the same team at EMBL-EBI to clarify particular curation tasks.

¹<http://www.myddelton.co.uk/blog/what-so-what-now-what>

Table 1. Overview of user research activities. The table provides an overview of the observation study on curation practices.

Activity	Purpose	Participants	Materials	Outputs	Analysis
Stakeholder interviews	Learn about curation practices	3 project leads and 4 curators from 5 different curation teams	Interview guide	Interview notes	Observations - Patterns - Implications
Follow-up interviews	Clarify certain aspects of curation	1 project lead and 1 curator from the same team	Clarification questions	Interview notes	
Stakeholder workshop	Validate learnings from interviews	3 project leads and 4 curators from 4 different curation teams	Guide and notes	HCW themes	Transcribed HCW themes and feedback
			Draft curator persona	Curator persona with feedback	Revised curator persona
			Draft curation process and example screenshots	Curation process with feedback	Revised curation process
					Curation experience map
Report on curation practice	Consolidate and share validated learnings from our research	Previous participants and another researcher	Draft report Revised persona & curation experience map	Revised materials incorporated into this report	

Furthermore, we conducted a stakeholder workshop¹² with three project leads and four curators from four different curation teams to validate our main learnings from the interview sessions. As some curators took part both in the interviews and the workshop, overall we have engaged with 12 curators and team leads from seven different teams. The participants were presented with the preliminary curator persona¹³ and a workflow outlining the curation process. The participants were invited to give us their feedback on these drafts and express their challenges or pain points as How Can We (HCW) questions². Their feedback was used to revise the curator persona and the curation process workflow and was consolidated into the curation experience map.

Community survey

Based on the interview guide used for the observational study, we formulated questions to understand the immediate challenges. The survey consisted of 15 questions (see *Extended data*⁸) ranging from, for instance, the section of the article the curators were most interested in; the types of biological entities curators look for; and whether it helps to know that a given article has been curated/accounted for in another database. The community survey was conducted online and was developed using *Typeform*. The survey was promoted via the mailing lists of various consortia, such as ELIXIR, International Society for Biocuration (ISB) and Alliance of Genome Resources. These widely known consortia provide a forum for developers, researchers and curators to streamline and standardise the maintenance of biological resources. The survey was conducted between December 2018 to January 2019.

²<http://www.designkit.org/methods/3>

Ethical issues

We confirm that we have obtained consent to use data from the participants as per the [Europe PMC privacy notice](#). The privacy notice is formulated in accordance with [EMBL's data protection framework](#). The consent was part of the survey form and the participants can take the survey on accepting the terms and conditions for data re-use.

Results and discussion

Observational study

Curator persona. We created a persona called Ashley (see *Underlying data*⁸) to present the curators' needs, sentiments, tasks and pain points from their own perspective in more detail to help us empathise with them¹³:

Ashley curates with **precision** and **attention to detail**, while trying to be as efficient as possible. Ashley is looking for **very specific information** about an experiment that the authors of a paper do not always report in a lot of detail. Ashley appreciates being able to ask a team mate when the “detective work” bears no fruit.

Apart from these “organic, informal discussions”, Ashley works independently during “triage”, “annotation” and to fill in the curation record in the Editor. During the latter stage Ashley tries to “translate from the author's language to the curator's language” using the appropriate **identifiers** and **Control Vocabulary (CV) terms** for species, proteins, methods and other important entities so that the curated evidence is referred to **precisely** and **consistently** and the annotations in the curation record are self explanatory outside the context of the paper.

This is cumbersome as a particular type of evidence is not always referred to in the same way and in enough detail in the literature. Moreover, the Editor is not integrated with the search and annotation tools, so Ashley spends a lot of time going back and forth between the paper and the Editor, translating the curatable text from the paper into CV terms, switching browser tabs and consulting notes from the online research and team discussions.

Curation experience map. The curation experience map in Figure 1 presents the identified pain points in the context of the main curation activities. As shown in the map, curation consists of four stages:

- Deciding which entity (primarily protein in this case) to curate. What to curate often depends on the curator's background and the project that they are working on.
- "Triaging" the literature to identify relevant publications.

- "Annotating" a relevant publication to identify the precise curatable information in detail, including determining the **species** and the relevant **experimental method**.

- Filling in the curation record based on the curatable information in the publication (which is often done in parallel with annotating the paper).

In a typical curation scenario, curators:

- Search PubMed for a protein and scan the titles in the search results to identify relevant **experimental** papers during "triage".
- If a **title** indicates that the paper is relevant (e.g. by mentioning the protein and/or **species** of interest), then they skim read the **Methods** and the **Results** of the paper. They are particularly interested in **Figures**, **Tables** and their **Legends**, which is where they usually find the key (curatable) information.

Curating papers "manually, one protein at a time"

Figure 1. The curation experience map presents the pain points in the context of the main curation activities.

- These sections are read more thoroughly during the “annotation” stage to identify the **exact experimental context** that needs to be curated. They may “glance through” the Abstract or skip it altogether.

A pain point during “triage” is **identifying relevant publications**: The curators reported that most publications returned in a PubMed search are usually not relevant (i.e. they are false positives). Additionally, because they are often looking for **very specific** and at times **underreported** experimental evidence, some searches return very **few papers** or **no papers** at all.

Ambiguities in the paper about **species**, **proteins**, and relevant **experimental methods** may slow down curation significantly during “triage” and “annotation”. The curators highlighted that identifying the **species** as their main pain point as this task may take up to “75% of the curation effort”, and in the end may turn out to be irretrievable from the paper.

Furthermore, to get clarity on the details of an experiment, curators would look up specific references in the paper and do further research online. If this “detective work” is not successful, the curators will either not annotate the paper or will provide fewer annotations. The curators would also discuss unresolved questions such as the annotation of unusual data or what to do when there is no curatable data with their teammates during “organic, informal discussions”. As a last resort they would contact the authors directly; however, authors often do not respond to requests for clarification.

If there are no matching CV terms to annotate the paper new ones are requested. This can delay annotating the paper because the ontology staff are often different to those curating the papers and sometimes requesting new terms results in prolonged discussions between curators and ontologists.

It was observed that curators use different tools at each stage, which are not integrated with each other. During “triage” they would search PubMed for relevant publications and then look at a particular paper on the publisher’s site. Annotating a publication may involve downloading or printing the pdf version of the paper and highlighting curatable text. To fill in the curation record they use a bespoke tool which they call “the Editor”, which presents a template that needs to be filled in with molecule names and experimental context, supported by standardised identifiers, controlled vocabularies and ontologies as well as free text describing the experimental evidence. Most of the times the assertions that go in the database are not in the paper in the same words.

Summary of the pain points in curation workflows. The identified pain points were formulated as How Can We (HCW) questions as follows:

- HCW identify **relevant publications** for curation in search results or a list of references during “triage”?

- HCW identify **species**, **experimental methods**, **molecules** (primarily **proteins**) and other important entities (such as **cells** and **tissues**) in a publication during “triage” and “annotation”?
- HCW help curators fill in the curation record more efficiently?

Community survey

The survey received 42 respondents in total, covering a number of European countries, such as the United Kingdom, France, Italy and Switzerland. The majority of the participants identified themselves as ‘Scientific curator’ with over 5 years experience in curation. Broadly speaking respondents mainly curate peer-reviewed articles (43.6%), followed by review (25.5%) and preprint (16%) articles. **Figure 2** shows their preference in the type of articles for curation.

Figure 3 shows the section of articles of interest to the curators: the majority look for method sections in articles. The other sections of importance are the figures/tables and their legends. Apart from these, the supplementary data seems to be a section of importance. Furthermore, as shown in **Figure 4**, the types of entities curators look for in articles were diverse with preference given mainly to: genes/protein curation and their functions, database accession numbers, experimental methods and gene mutations.

Respondents were asked if it was useful to know a given article was already curated by another database: the results indicate that it was useful (see **Figure 5**). A follow-up question was asked as to (if useful) why it was useful to know if an article is already annotated by another database, the majority of the responses ranged from: avoiding duplication if the curators belong to the same consortium, as a means of validation (in case of ontology terms), and consistency in annotations.

Outcomes of the user research project

Text mining approaches are sophisticated and play a vital role in addressing big data questions, the results of which can contribute to supplying “leads” on key papers for curation. However, curators require a wide variety of very precise information. Addressing each of those specific requirements will be a complex task, but text mining systems can certainly provide underlying services based on broad commonalities in the requirements that can prove useful to curators. To this end, this effort has proved to be useful in terms of understanding the main challenges faced by curators. While the sample size of the community survey was small, when analysed in conjunction with the observational study we found significant commonalities on the work practices. For instance, in the survey when the respondents were asked for their biggest challenge while curating, the majority responses indicated finding relevant papers, and identifying specific information that includes genes and species.

Figure 2. The pie chart shows the article types that are of interest to curators.

Figure 3. The figure shows the article sections that are of interest to curators.

Figure 4. The graph provides an overview of the entity types of interest.

Figure 5. The figure shows the response for the question: How useful is it to know if the article has been curated by another database?

Our research on curation practices so far indicates a need to better support curators on the following areas:

- Identify **relevant papers** for curation during “triage”. An efficient way towards article selection, where search results could be prioritised based on a set of parameters.
- Identify **species**, relevant **experimental methods**, **molecules** (primarily **proteins**) and other important entities (such as **cells** and **tissues**) in a publication during “triage” and “annotation”.
- Retrieving certain sections of articles such as Methods, Figures or Results.
- Integrate triage systems to the various curation workflows.

Conclusion

Contributions made by manual curation are vital to the maintenance of biological databases. To maximise the impact of this critically important process, the latest technological advancements need to be leveraged. Under the Elixir Data platform, we have established infrastructural elements to support scalable curation, which includes automated systems to ingest and aggregate from various sources, APIs to redistribute the annotations and an application called SciLite to display annotations on articles. However, a key challenge for scalable curation is to make use of such core components across different curation teams, whose requirements and workflows can be highly precise and vary widely. Consequently, this requires engagement with the curation community to derive actionable insights that may contribute towards service delivery. Therefore, this project lays the foundation needed to understand the commonalities shared among various curation workflows. Going forward, we will use the results of the project to feed into improvements to text-mined annotation quality and coverage, triage and browsing systems or other engineering solutions.

Data availability

The interview responses from the observation study have not been made public to protect the participants’ privacy. Please contact the corresponding author to apply for access to the data, providing details of the information required and the intended use of the data. Access to the data will be granted once permission from participants to share the data has been obtained.

Underlying data

Zenodo: Results of user research project to understand data curation practices. <https://doi.org/10.5281/zenodo.3209658>⁸.

This project contains the following underlying data:

- Curator persona.docx (the curator persona generated during the first part of the study).
- Curator survey results.xlsx (raw data taken from the survey given to each participant).

Extended data

Zenodo: Results of user research project to understand data curation practices. <https://doi.org/10.5281/zenodo.3209658>⁸.

This project contains the following extended data:

- Observation study - interview guide.docx (interview guide outlines the type of questions to be asked)
- Curator survey questions.docx (questionnaire given to each participant in the community survey).

Data are available under the terms of the [Creative Commons Attribution 4.0 International license](#) (CC-BY 4.0).

Acknowledgements

We are grateful to our participants, to Francisco Talo for his involvement and to Ane Møller Gabrielsen for her comments.

References

1. Lane L, Argoud-Puy G, Britan A, *et al.*: **neXtProt: a knowledge platform for human proteins**. *Nucleic Acids Res.* 2012; **40**(Database issue): D76–D83. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
2. Mottin L, Gobeill J, Pasche E, *et al.*: **neXtA5: accelerating annotation of articles via automated approaches in neXtProt**. *Database (Oxford)*. 2016; **2016**: pii: baw098. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
3. Karamanis N, Seal R, Lewin I, *et al.*: **Natural language processing in aid of FlyBase curators**. *BMC Bioinformatics*. 2008; **9**: 193. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
4. 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](#)
5. 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](#)
6. 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](#)
7. Piovesan D, Tabaro F, Mičetić I, *et al.*: **DisProt 7.0: a major update of the database of disordered proteins**. *Nucleic Acids Res.* 2017; **45**(D1): D219–D227. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
8. Venkatesan A, Karamanis N, Ide-Smith M, *et al.*: **Results of user research project to understand data curation practices**. [Data set]. Zenodo. 2019. <http://www.doi.org/10.5281/zenodo.3209659>
9. Gothelf J, Seiden J: **Lean UX: Designing Great Products with Agile Teams**. O'Reilly Media, Inc. 2016. [Reference Source](#)
10. Karamanis N, Pignatelli M, Carvalho-Silva D, *et al.*: **Designing an intuitive web application for drug discovery scientists**. *Drug Discov Today*. 2018; **23**(6): 1169–1174. [PubMed Abstract](#) | [Publisher Full Text](#)
11. Beyer H, Holtzblatt K: **Contextual Design: Defining Customer-Centered Systems**. Morgan Kaufmann. 1997. [Reference Source](#)
12. Gray D, Brown S, Macanuso J: **Gamestorming: A Playbook for Innovators, Rulebreakers and Changemakers**. O'Reilly Media, Inc. 2010. [Reference Source](#)
13. Cooper A: **The Inmates Are Running the Asylum: Why High Tech Products Drive Us Crazy and How to Restore the Sanity**. Sams-Pearson Education. 2004. [Reference Source](#)

Open Peer Review

Current Peer Review Status:

Version 1

Reviewer Report 14 October 2019

<https://doi.org/10.5256/f1000research.21298.r53748>

© 2019 Arighi C. 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.

Cecilia N. Arighi

Center for Bioinformatics and Computational Biology, University of Delaware, Newark, DE, USA

This work presents a study about the biocuration community and its literature-based curation practices. The work intends to identify pain points and commonalities in the curation workflow where ePMC infrastructure could work to assist this community.

- The study includes the detailed observation of 5 curation groups following their regular literature curation work and a survey targeted to the biocuration community to learn about the literature-based curation tasks. The approach is appropriate and well designed. However, it seems that the curation groups observed are biased toward protein/gene-centric curation, whereas there are other workflows, such as those for model organisms, chemicals, that may have a completely different approach to the literature curation. In fact, some model organism databases (like MGI) first do triage to select articles about the organism, then classify articles based on specific curation topics (phenotype, GO, etc). Then the conclusion from their workflow analysis could be different. Has anything come up from the survey indicating that other groups were represented in this work? For example, in the survey there is a question about database resource "Which data resource(s) do you curate?" But the result of this, which is important to learn about the databases represented, is not shown in the survey result document. I understand that the information in the survey result may be hidden to protect privacy of participants but showing the distribution of databases and/or type of databases represented (model organism vs. specific domain, like structure, PPI, etc) can shed light into bias or non-bias toward one type of curation.
- Another question is about the groups observed. In the introduction, a couple of groups are mentioned that have integrated text mining pipelines in their curation work. It would be important for the study to indicate if any of the curation groups that were observed or curators surveyed used text mining tools for their work, or to include some group in observation study that do use text mining to see in what capacity and if bottlenecks in literature curation are the same.
- The manuscript would be enriched by describing previous work done on this area

(biocuration workflows and bottlenecks in literature curation) and compare the conclusions in this manuscript with others. There are a few papers from BioCreative that looked into this matter.^{1, 2, 3}

- Finally, please consider using the term expert curation instead of manual curation. I think using the concept of manual curation in databases is not appropriate for modern times, as curators use tools to help them do all or part of their work, is not completely manual.
- Minor: Figures 2-5 should indicate number of participants who responded to the questions over total number of survey participants. The X-axis in Figures 3-4 need a label, same with Y-axis in Figure 5.

References

1. Hirschman L, Burns GA, Krallinger M, Arighi C, et al.: Text mining for the biocuration workflow. *Database (Oxford)*. 2012; **2012**: bas020 [PubMed Abstract](#) | [Publisher Full Text](#)
2. Lu Z, Hirschman L: Biocuration workflows and text mining: overview of the BioCreative 2012 Workshop Track II. *Database (Oxford)*. 2012; **2012**: bas043 [PubMed Abstract](#) | [Publisher Full Text](#)
3. Singhal A, Leaman R, Catlett N, Lemberger T, et al.: Pressing needs of biomedical text mining in biocuration and beyond: opportunities and challenges. *Database (Oxford)*. 2016; **2016**. [PubMed Abstract](#) | [Publisher Full Text](#)

Is the work clearly and accurately presented and does it cite the current literature?

Partly

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: My area of expertise is on biocuration, usability and text mining applied to biocuration.

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.

Reviewer Report 07 October 2019

<https://doi.org/10.5256/f1000research.21298.r53750>

© 2019 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

The MITRE Corporation, Bedford, MA, USA

This is a well-designed and informative examination of data curation practices in support of the Elixir Data Platform, with a focus on exploring curator needs and pain points to identify where automated tools might help. The article presents results from a series of studies based on interviews with curators and responses from a questionnaire filled out by over 40 curators from 11 countries.

These results are consistent with an earlier article from 2012, Text Mining for the Biocuration Workflow¹, which summarized the findings of a workshop held at the third International Biocuration Conference. Below is a short from the 2012 paper:

Curators wanted tools that were easy to use, easy to install and easy to maintain by the intended end user (ideally, a developer associated with the curation team, who will not necessarily be an expert in text mining or natural language processing). The tools do not have to be perfect, but they need to complement (not replace) the biocurator's function. A number of curation groups indicated that they would use the tools to do an initial batch processing, followed by biocurator validation, where the biocurator makes a yes/no decision and avoids having to type or look names up in a large database. Another important use was linking mentions of biological entities in text with the correct identifiers in biological databases, as well as linkage to the appropriate ontology terms. A number of curators felt that they would like text mining tools to aid in identifying and prioritizing papers for curation, to avoid wasting time on papers that did not have 'relevant' (e.g. curatable or novel) results. They also wanted tools to identify the sections of full-text papers containing curatable information.

It would be informative to do a comparison of the findings from the 2012 paper with this paper, to see whether curator needs have changed – and to what extent automated tools have been able to address some of the curator needs.

One of the paper's most interesting findings from the curator feedback is the need to identify species. This was flagged as a particular pain point, taking 'up to 75% of the curation effort'! This has also been a consistent stumbling block for text mining systems, because identification of species is essential to link a mention of a gene or protein to the correct accession number in databases such as UniProt or EntrezGene. This turns out to be a hard problem for text mining systems (as well as for curators) for several reasons: 1) mentions of species are often given as

background information and may well not be mentioned in the same sentence (or even section) as mentions of the gene or protein being studied; 2) authors may want to generalize findings to other species (especially to humans) even when the experiments have been done on other species; 3) information about the specific experimental constructs may be buried in the methods section – and may involve inserting a gene from one organism into the genome of a different organism.

One interesting omission is a discussion of the need for interactive curation tools. This has been a major theme of recent BioCreative evaluations (see, e.g., Overview of the interactive task in BioCreative V²). If an interactive system could show the curator a prioritized list of candidates, this could speed up several curation activities. Specifically, an interactive system could, e.g., show a ranked list of candidate papers for curation, with evidence highlighted; or show a paper (section) with a gene or protein mention highlighted together with a selectable list of candidate species, so the curator could quickly select the correct species in order to link to the correct accession number; or show highlighted evidence sentences supporting protein-protein interaction, for quick validation or rejection by the curator. Given a goal of providing tools to speed the curation workflow, interactive systems offer a promising approach by putting the human in the loop to augment text mining, where automated tools do not, on their own, provide sufficient accuracy.

Finally, in the Conclusion section, the authors discuss the need to tailor capabilities for different curation tasks or workflows. Identifying commonalities is indeed key, as the authors note; but there is also an urgent need to develop methods to quickly tailor tools to new tasks or specific requirements in a curation workflow. This is an underexplored area, but may be key to more widespread adoption of text mining tools.

Specific comments:

- Add references to some of the older background work in this area.
- The Conclusions section of the Abstract mentions “actionable items” but doesn’t provide specifics. The list on p. 9, col 1 top is informative, and could be included in the abstract, e.g., prioritizing papers; filtering articles based on specific entity types; and retrieving specific sections of articles.
- A table summarizing the different interactions with curation teams and curators would be helpful, along with the specific types of curation being done by the teams. This information is presented in the text at different points, but it is hard to keep track without a summary, since there were several rounds and types of interactions.
- In Figure 3, it would be useful to know the denominator, as well as the actual number of curators identifying specific sections.
- Figure 4 is very useful, but hard to read. In addition, it would be interesting to know what specific controlled vocabularies are in use for each of these types of information.
- Figure 5 – again, it would be useful to know the denominator (the total number of respondents for that question).

References

1. Hirschman L, Burns GA, Krallinger M, Arighi C, et al.: Text mining for the biocuration workflow. *Database (Oxford)*. 2012; **2012**: bas020 [PubMed Abstract](#) | [Publisher Full Text](#)
2. Wang Q, S Abdul S, Almeida L, Ananiadou S, et al.: Overview of the interactive task in BioCreative

V.Database (Oxford). 2016; **2016**. [PubMed Abstract](#) | [Publisher Full Text](#)

Is the work clearly and accurately presented and does it cite the current literature?

Partly

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Evaluation of text mining for biomedical applications, particularly curation.

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.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research