Corpora as evolving entities: embedding corpora in biomedical ontologies

Elizabeth T. HOBBS[a], Stephen M. GORALSKI[a], Ashley MITCHELL[a], Andrew SIMPSON[a], Dorjan LEKA[a], Emmanuel KOTEY[a], Matt SEKIRA[a], James B. MUNRO[b], Suvarna NADENDLA[b], Rebecca JACKSON[b], Aitor GONZALEZ-AGIRRE[c], Martin KRALLINGER[c,d], Michelle GIGLIO[b], and Ivan ERILL[a,1]

[a] *Department of Biological Sciences, University of Maryland Baltimore County, Baltimore, MD (USA)*

[b] *Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD 21201, USA*

[c] *Barcelona Supercomputing Center (BSC), 08034 Barcelona, Spain*

[d] *Centro Nacional de Investigaciones Oncológicas (CNIO), 28029, Madrid Spain*

Abstract. Biomedical corpora are essential for the development of text-mining tools to assist in and enhance the curation of biomedical knowledge stored in scientific text. Biomedical corpora lag behind other resources in the adoption of FAIR principles. This is due to their research-driven, narrow scope, limited funding and the lack of standards, and it is aggravated by rapid obsolescence. Here we report on the development of a novel corpus containing annotated evidence-based assertions in journal articles. The ECO-CollecTF corpus uses the Evidence and Conclusion Ontology (ECO) as a referential framework and is the first corpus to be released as an ontology embedding. Our work details the requirements for releasing a corpus embedded in an ontology, and outlines the many benefits of this approach, using one or more ontologies, for making corpora more findable accessible, reusable and interoperable, to delay obsolescence, and to promote scientific collaboration.

Keywords. biomedical, corpus, ontology, annotation, evidence, FAIR

1. Introduction

Life scientists are increasingly dependent upon the availability of standardized scientific knowledge to analyze high throughput experimental results [1,2]. Expert curators extract this information from journal articles and make it available as standardized knowledge [3–5]. However, manual curation cannot keep up with the rapid pace of publication. This has led to the use of text mining techniques to assist with curation [3–9]. Manually-constructed, gold standard biomedical corpora play a vital role in the development of text mining techniques. Biomedical corpora provide a computer-readable mapping between formal biomedical entities and written text, which can be used to train and fine tune text-mining systems [4,7,10–13].

[1] Corresponding Author, Ivan Erill, Department of Biological Sciences, University of Maryland Baltimore County (UMBC), 1000 Hilltop Circle, Baltimore, MD 21250 (USA); E-mail: erill@umbc.edu.

Databases, taxonomies and ontologies support the creation of computer-accessible, standardized knowledge by providing unique identifiers and well-defined concepts and concept relationships. In this context, the annotation process inherent to the creation of a corpus can be formally defined as the establishment of a set of mappings between segments of human readable text and objects in a database, taxonomy or ontology. Many corpora have been created to satisfy a range of goals, such as recognizing disease and species names [14–17], identifying biomolecular interactions [7,18], recognizing anatomical entities [19,20], specifying negation and speculation [21,22], and indicating sequence features, proteins, chemicals, biological processes, molecular functions, and cellular locations [23].

Corpora provide considerable value to the community, but their development and maintenance requires a significant investment of time and resources [24]. As it is the case with other biomedical datasets and resources [25–27], a substantial fraction of existing biomedical corpora remain underused and are hard to access. This is partly due to the very nature of biomedical corpora. Most corpora are generated by dedicated research groups in response to specific research needs. Even though they may later be shared with the community via initiatives like BioCreative [28], their narrow focus means that most corpora will be of relatively little interest to the wider biomedical community. Their existence may not be known outside of the research focus community, and they will generally lack unified formats and mechanisms of access. Being linked to a specific research need also entails that funding for their development is likely to expire, making corpora harder to find and prone to obsolescence, as the databases and ontologies they depend on continue to evolve.

The issues surrounding the lack of use, obsolescence and difficult accessibility of scientific resources led to the development of the FAIR principles in 2016 [29]. The FAIR principles provide guidelines for the development of a data and systems infrastructure that enables data and knowledge to be “findable”, “accessible”, “interoperable” and “reusable” by both humans and machines. The FAIR guidelines hence encourage long-term data stewardship that enables discovery and innovation through the reuse of data [29–31]. Under these principles, critical infrastructure has been developed to unify and efficiently crosslink large biomedical resources, such as the sequence databases hosted by the National Center for Biotechnological Information and the European Bioinformatics Institute [32,33]. Infrastructure has also been created to integrate and consolidate the development of biomedical ontologies using unified standards and best-practices and accessible repositories [34–36].

The great advances in standards and infrastructure prompted by the adoption of FAIR practices in biomedical resources have not fully percolated to corpora. Even though some standards and repositories for annotation have been developed [37,38], there is no universally accepted central repository for corpora, and issues like continued development, community adoption and obsolescence remediation remain to be addressed. Here we propose and explore an innovative approach to address these issues, through the distributed embedding of biomedical corpora within biomedical ontologies. We demonstrate the feasibility of this approach on a manually-curated corpus of evidence-based statements in scientific manuscripts, using the Evidence and Conclusion Ontology (ECO) as the primary ontological framework [39]. We showcase the advantages of this approach in generating fully accessible, evolvable corpora that can be easily and extensively reused and expanded by the community and that escape obsolescence by evolving organically with the ontology. We discuss how this approach can be easily extended to generate a new generation of cross-ontology corpora that

seamlessly builds up an all-encompassing knowledge-base of textual annotations of biomedical concepts.

2. Methods

2.1. Annotation of evidence

Training and guidelines for annotation are documented in the Zenodo ECO-CollecTF reference project repository. Briefly, curators annotate only individual or consecutive sentences with self-contained evidence-based assertions in the Results (or Results and Discussion) section of selected manuscripts. Curators create an annotation for text that maps to a native evidence term in ECO if that evidence is used in that sentence or a contiguous one to make an assertion about a biological entity (**Figure 1**). We consider five primary categories of biological entities that reference broad ontology terms from the Gene Ontology (biological process - GO:0008150, molecular function - GO:0003674, cellular component - GO:0005575) [40], the Sequence Ontology (SO:0000110) [41] and the Ontology of Microbial Phenotypes (OMP:0000000) [42], as well as a broadly defined taxonomy/phylogeny category to capture statements about evolutionary relationships. In addition, annotation attributes were used to capture the curator's confidence in the text-ECO term mapping (high/medium/low), the strength of the assertion statement (high/medium/low), whether the annotation was for consecutive sentences, or whether the assertion was negative. Articles were selected from CollecTF [43], a database of bacterial transcriptional regulation. Annotations were made using BRAT [44].

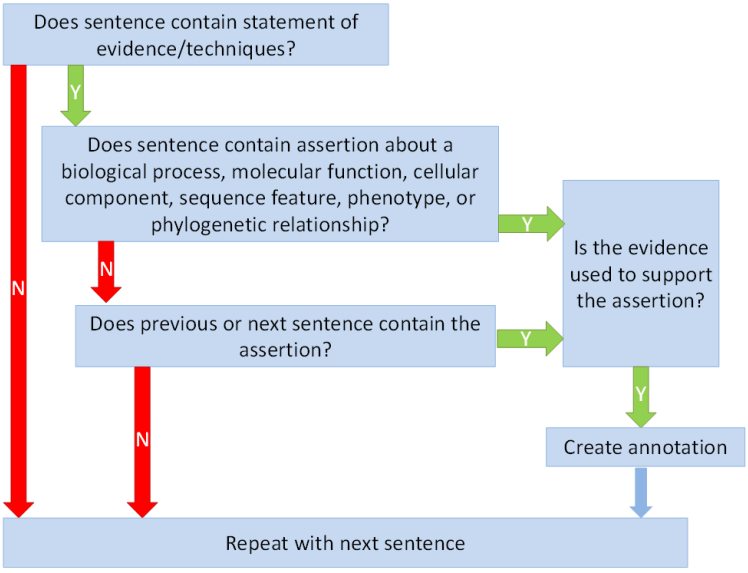

Figure 1. Schematic representation of the annotation workflow for the ECO-CollecTF corpus.

2.2. Manuscript parsing and computation of inter-annotator agreement

A set of 84 articles was downloaded from PubMed Central in XML format, and parsed to extract the result sections, known to contain the largest concentration of evidence passages [45], using custom Python scripts. HTML tags were removed and non-ASCII characters mapped to ASCII. The Natural Language Toolkit [46] (NLTK) was used to break the text into sentences. Inter-annotator agreement was calculated for each pair of curators using Cohen's *K* [47]. The IAA of the corpus was computed as the average of these pairwise scores.

2.3. Ontology embedding

The corpus is released as an ECO OBO file containing annotations. Annotations are conceived as self-contained, independent units associated to ECO terms in the ontology through an "annotation" property, formatted for convenience in JSON. This property contains several primary attributes: source, curator, relationship, manuscript, sentence and annotation. Relationships encapsulate one or more annotations on an individual sentence from an individual curator and can span multiple ontologies. Sentences are identified relative to their order in the reference manuscript, and provided also in parsed form as text. The annotation markup is defined by offsets within the sentence, and contains also the ancillary attributes described above (text-ECO term mapping confidence, assertion strength, consecutive sentences, negative assertion and entity category). The source is identified by the ECO-CollecTF project DOI, the manuscript by its PubMed identifier (PMID), the relationship type by a Relations Ontology (RO) identifier and entity categories by the respective ontology term identifier [48]. For curators, annotations and relationships, we follow the recommendation for the Dublin Core™ Metadata Term element identifier [49] and use the Version 4 UUID namespace as a Uniform Resource Name (URN) [50]. Curators are linked to manuscripts via a dynamic table referenced in the project source. To generate the OBO file, BRAT ".ann" files containing the annotations were processed using Python scripts, converted into the JSON annotation structure and assigned to the OBO annotation fields for the ECO terms used in the annotation. All text parsing and format conversion scripts are available in the project reference.

3. Results and discussion

3.1. The ECO-CollecTF corpus

The ECO-CollecTF corpus is the result of a two-year curation effort involving the ECO and CollecTF teams and six undergraduate curators. A primary element of this effort was the definition of the annotation scope for evidence in scientific manuscripts. Evidential techniques are often mentioned in scientific text but techniques, by themselves, do not constitute evidence. The ECO structure reflects this by defining two sub-ontologies: evidence and assertion, and deriving terms as cross-products of these two fundamental elements. The annotation reflects this framework by focusing on sentences that contain both a statement of evidence and an assertion based on this evidence. Because such statements are usually split into two declarative sentences, we annotated also pairs of consecutive sentences following this schema. After training on

the annotation methods and approach, curator teams were assigned scientific manuscripts focusing on bacterial transcriptional regulation, and the curation process was supervised via regularly scheduled team meetings to discuss annotation issues. This effort resulted in the annotation of 84 unique documents, yielding 5162 annotations. The average inter-annotator agreement (IAA), as measured by Cohen's *K*, was 0.69, comparable to reported K scores in annotation tasks of similar complexity [51,52].

3.2. Ontology embedding of the ECO-CollecTF corpus

Corpora present unique problems for the adoption of FAIR standards. These difficulties originate primarily from corpora's targeted, specific and research-driven nature, and are confounded by the availability of few standards, corpora's dependence on evolving external resources and the temporality of funding. These factors result in limited visibility, reproducibility and accessibility, as well as severely constrained reusability due to obsolescence. To address these issues, here we reevaluated the paradigm of corpus storage, moving away from the conventional storage of corpora as independent entities and towards their embedding within ontologies.

Since corpora reference ontology objects, the most logical approach to embedding a corpus in an ontology is to store annotations within the ontology objects they reference. This, in turn, requires that annotations be self-consistent and independent, but incorporate enough metadata to effectively reconstruct the corpus from the ontology. To accomplish these, we took an annotation-centric approach for embedding the ECO-CollecTF corpus in ECO (**Figure 2**).

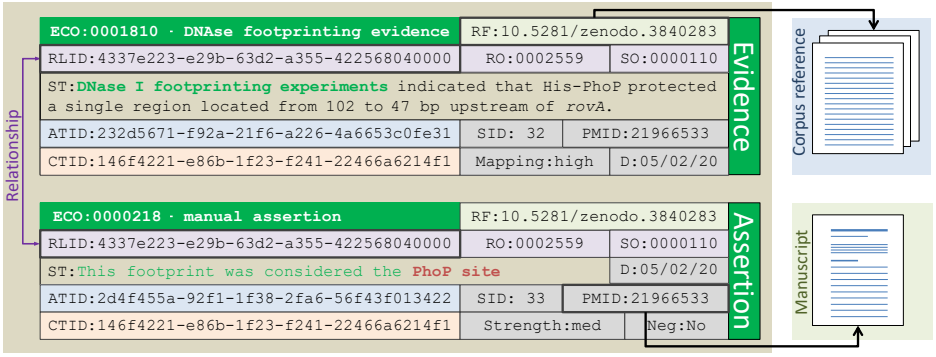

Figure 2. Embedding of annotations in ECO terms. The two annotations, linked by their relation identifier, detail the text mapping to the ECO evidence term, as well as the sentence mapping to the ECO assertion term.

To encapsulate annotations in the ontology, we define a new, JSON-formatted 'annotation' property for ECO terms (green boxes in **Figure 2**). The annotation property contains several unique identifiers. The sentence (SID) and PubMed (PMID) identifiers link the annotation to a specific sentence in a published article, which is marked up with offsets and also provided in parsed form. Each annotation and its curator have their own UUID identifiers (ATID and CTID). The annotation also contains qualifier fields of interest to the curation effort, such as the ECO mapping confidence and the object category, referenced by an external ontology identifier (e.g.

SO:0000110). Functional ECO terms are the cross-product of evidence and assertion terms. The ECO-CollecTF corpus mirrors this reference framework by embedding independent annotations to evidence and assertion terms. The relationship between both terms is instantiated in a relationship array by a UUID identifier shared by both annotations. This array also holds the relation type, defined by a Relation Ontology (RO) identifier [48], as well as any relation-specific properties. The corpus, as an entity, is defined by an external digital object identifier (DOI). This designates a repository that contains all the training documentation and annotation guidelines, as well as the necessary code to parse source articles, compute corpus statistics and embed BRAT annotations into an ECO OBO file. To ensure that the entire approach is entirely reproducible, and that other contributors can expand corpus with new annotations, the repository contains a link to a dynamic curator table that maps curator UUIDs with article PMIDs.

The embedding of the ECO-CollecTF within ECO pioneered here represents a marked paradigm shift in corpus storage, with far-reaching implications for corpus generation and maintenance. This shift derives from a reassessment of the concept of corpus. A corpus is, by definition, a mapping between referenceable objects and segments of text referring to said objects. Published text is, by nature, immutable, whereas reference frameworks, such as ontologies, are continuously evolving entities. This makes ontologies a logical target for corpus storage. This transition, together with the shift towards an annotation-centric model adopted here, addresses many of the issues hindering the adoption of FAIR principles in corpora. The embedding of the corpus within the ontology directly addresses findability by leveraging a well-established, community-supported resource. Moreover, offloading the corpus onto the ontology also promotes long-term support and community involvement. Most significantly, corpus embedding directly addresses the fundamental challenge of obsolescence. Embedded in the ontology, the corpus can evolve organically with the ontology. Annotations can be transferred, if appropriate, to related terms when a term is obsoleted, and the format can be updated following ontology interoperability initiatives.

The adoption of an annotation-centric framework enhances reproducibility and reusability. In contrast with conventionally-released corpora, individual annotations in the ECO-CollecTF corpus are directly accessible to users, enabling sub-setting and customization. Users may, for instance, reassess inter-annotator agreement using metrics of their choice, subset the corpus to include only sentences with multiple annotator agreement, or extract only annotations involving SO category objects. Importantly, users can freely add annotations to the corpus, as long as they abide by the published annotation guidelines. Lastly, the integration of the corpus in the ontology enhances ontology development, open communication and community building, by bringing together focused research groups with ontology development teams. This interaction has beneficial effects on ontology development, prompting the fine-tuning of term and relationship definitions as curators identify inconsistencies. It also generates a library of examples of use, now currently available in ECO (v2019-10-16), that facilitates ontology use through textual illustration of ontology concepts and assists curators in making informed decisions about their applicability in different contexts.

3.3. Towards multi-ontology embedded corpora

An implicit advantage of the annotation-centric corpus-ontology embedding illustrated by the ECO-CollecTF corpus is that it is directly extensible to multi-ontology corpora (**Figure 3**). Annotation in different sentences, and to different terms in multiple ontologies, can be formalized through the embedding of annotations in the respective ontologies, with multiple relation identifiers in the relation array structuring the interrelationships between ontology terms.

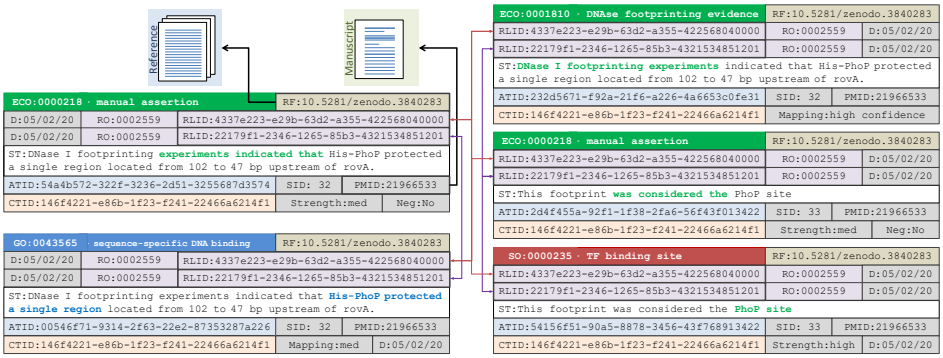

Figure 3. Illustration of a multi-ontology corpus annotation involving two sentences and mappings to terms from ECO, GO and SO.

Multi-ontology embedded corpora build on the advantages outlined here for the ECO-CollecTF corpus. Offloading annotations to the ontology enables the corpus to evolve with the ontologies harboring it, delaying obsolescence and expanding accessibility, since the corpus becomes accessible through any of its enabling ontologies. Interoperability mechanisms and protocols are already in place for ontologies, resulting in a direct enhancement of corpus interoperability for embedded corpora. The ability of users to subset and customize also increases with multi-ontology embedded corpora, since users can slice corpora using ontologies, focusing only on specific referential elements. Most importantly, the transition of corpora from standalone entities to ontology embeddings fosters the development of an open-ended, mutually-beneficial collaborative environment for the ontological and research communities, promoting alignment of both corpora and ontologies, and fostering community involvement.

Many further steps are required to complete the vision of ontology-embedded corpora advocated in this work. An important step will be the development of a centralized source text resource, derived from existing repositories like PubMed, where text elements (such as sentences) have been pre-parsed and uniquely identified. Such a resource will greatly facilitate the deployment of embedded corpora, making annotations slimmer and preventing congruence errors in annotated text. Even though many of the issues limiting FAIR adoption in corpora could in theory be addressed by a centralized corpus repository, addressing obsolescence would require a significant, continued investment in coordination with ontology developers and other stakeholders. By distributing the corpus across ontologies, corpus embedding provides an accessible, rapidly implementable and sustainable solution that provides added value to

participating ontologies and has the potential to bring together a diverse set of scientific disciplines.

4. Availability

The ECO-CollecTF reference and the ECO OBO file are available as Zenodo repositories (ECO-CollecTF reference – DOI:10.5281/zenodo.3840283, ECO OBO file - DOI:10.5281/zenodo.3843501). The project is available through GitHub (https://github.com/ErillLab/OntoCorp).

5. Acknowledgements

The authors would like to express their gratitude to Marcus C. Chibucos and James Hu for insightful discussions and technical assistance. This work was supported by the National Science Foundation, Division of Biological Infrastructure [1458400] and the National Institutes of Health [R01GM089636, U41HG008735], and by a management commission from Plan TL (Plan de Impulso de las Tecnologías del Lenguaje) of the Spanish Ministerio de Asuntos Económicos y Transformación Digital to BSC-CNS.

6. Author contributions

Conceptualization: ETH, SMG, JBM, SN, MG, IE; funding acquisition: MG, MK; methodology: ETH, SMG, JBM, SN, MG, IE; project administration: MG, IE; resources: RJ, AG-A, MK; software: ETH; supervision: ETH, MG, IE; data curation: ETH, SMG, AM, KL, AS, EK, MS; visualization: ETH, IE; writing – original draft: ETH, IE; writing – review & editing: ETH, AM, EK, JBM, SN, AG-A, MK, MG, IE.

References

[1] Marx V. The big challenges of big data. Nature 2013;498:255–60. https://doi.org/10.1038/498255a.

[2] Reshetova P, Smilde AK, van Kampen AH, Westerhuis JA. Use of prior knowledge for the analysis of high-throughput transcriptomics and metabolomics data. BMC Syst Biol 2014;8:S2. https://doi.org/10.1186/1752-0509-8-S2-S2.

[3] Hirschman L, Burns GAPC, Krallinger M, Arighi C, Cohen KB, Valencia A, et al. Text mining for the biocuration workflow. Database 2012;2012:bas020–bas020. https://doi.org/10.1093/database/bas020.

[4] Verspoor K, Cohen KB, Lanfranchi A, Warner C, Johnson HL, Roeder C, et al. A corpus of full-text journal articles is a robust evaluation tool for revealing differences in performance of biomedical natural language processing tools. BMC Bioinformatics 2012;13:207.

[5] Kwon D, Kim S, Wei C-H, Leaman R, Lu Z. ezTag: tagging biomedical concepts via interactive learning. Nucleic Acids Res 2018;46:W523–9. https://doi.org/10.1093/nar/gky428.
[6] Kim J-D, Ohta T, Tsujii J. Corpus annotation for mining biomedical events from literature. BMC Bioinformatics 2008;9:10. https://doi.org/10.1186/1471-2105-9-10.
[7] Islamaj Dogan R, Chatr-aryamontri A, Kim S, Wei C-H, Peng Y, Comeau D, et al. BioCreative VI Precision Medicine Track: creating a training corpus for mining protein-protein interactions affected by mutations. BioNLP 2017, Vancouver, Canada,: Association for Computational Linguistics; 2017, p. 171–5. https://doi.org/10.18653/v1/W17-2321.
[8] Wei C-H, Leaman R, Lu Z. Beyond accuracy: creating interoperable and scalable text-mining web services. Bioinformatics 2016;32:1907–10. https://doi.org/10.1093/bioinformatics/btv760.
[9] Kim J-D, Wang Y, Fujiwara T, Okuda S, Callahan TJ, Cohen KB. Open Agile text mining for bioinformatics: the PubAnnotation ecosystem. Bioinformatics 2019;35:4372–80. https://doi.org/10.1093/bioinformatics/btz227.
[10] Wei C-H, Lee K, Leaman R, Lu Z. Biomedical Mention Disambiguation using a Deep Learning Approach. Proc. 10th ACM Int. Conf. Bioinforma. Comput. Biol. Health Inform. - BCB 19, Niagara Falls, NY, USA: ACM Press; 2019, p. 307–13. https://doi.org/10.1145/3307339.3342162.
[11] Chen Q, Du J, Kim S, Wilbur WJ, Lu Z. Deep learning with sentence embeddings pre-trained on biomedical corpora improves the performance of finding similar sentences in electronic medical records n.d.:15.
[12] Zhang Y, Chen Q, Yang Z, Lin H, Lu Z. BioWordVec, improving biomedical word embeddings with subword information and MeSH. Sci Data 2019;6:52. https://doi.org/10.1038/s41597-019-0055-0.
[13] Wang Y, Kim J-D, Sætre R, Pyysalo S, Ohta T, Tsujii J. Improving the Inter-Corpora Compatibility for Protein Annotations. J Bioinform Comput Biol 2010;08:901–16. https://doi.org/10.1142/S0219720010004999.
[14] Doğan RI, Leaman R, Lu Z. NCBI disease corpus: A resource for disease name recognition and concept normalization. J Biomed Inform 2014;47:1–10. https://doi.org/10.1016/j.jbi.2013.12.006.
[15] Rebholz-Schuhmann D, Kirsch H, Arregui M, Gaudan S, Riethoven M, Stoehr P. EBIMed--text crunching to gather facts for proteins from Medline. Bioinformatics 2007;23:e237–44. https://doi.org/10.1093/bioinformatics/btl302.
[16] Gerner M, Nenadic G, Bergman CM. LINNAEUS: a species name identification system for biomedical literature. BMC Bioinformatics 2010;11:85.
[17] Pafilis E, Frankild SP, Fanini L, Faulwetter S, Pavloudi C, Vasileiadou A, et al. The SPECIES and ORGANISMS Resources for Fast and Accurate Identification of Taxonomic Names in Text. PLoS ONE 2013;8:e65390. https://doi.org/10.1371/journal.pone.0065390.
[18] Ohta T, Pyysalo S, Rak R, Rowley A, Chun H-W, Jung S-J, et al. Overview of the Pathway Curation (PC) task of BioNLP Shared Task 2013. BMC Bioinformatics 2015;16:S2.
[19] Hicks A, Hogan W, Pepine C, Boire N, Herring C, Seppala S. Introducing Hypertension FACTS: Vital Sign Ontology Annotations in the Florida Annotated Corpus for Translational Science. Thirty-First Int. Flairs Conf., AAAI Press; 2018, p. 7.

[20] Ohta T, Pyysalo S, Tsujii J, Ananiadou S. Open-domain Anatomical Entity Mention Detection. Proc. Workshop Detect. Struct. Sch. Discourse, Association for Computational Linguistics; 2012, p. 27–36.
[21] Pyysalo S, Ginter F, Heimonen J, Björne J, Boberg J, Järvinen J, et al. BioInfer: a corpus for information extraction in the biomedical domain. BMC Bioinformatics 2007;8:50.
[22] Vincze V, Szarvas G, Farkas R, Móra G, Csirik J. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. BMC Bioinformatics 2008;9:S9. https://doi.org/10.1186/1471-2105-9-S11-S9.
[23] Bada M, Eckert M, Evans D, Garcia K, Shipley K, Sitnikov D, et al. Concept annotation in the CRAFT corpus. BMC Bioinformatics 2012;13:161.
[24] Cohen KB, Fox L, Ogren PV, Hunter L. Empirical data on corpus design and usage in biomedical natural language processing n.d.:5.
[25] Johnson HL, Baumgartner WA, Krallinger M, Cohen KB, Hunter L. Corpus Refactoring: a Feasibility Study. J Biomed Discov Collab 2007;2:4. https://doi.org/10.1186/1747-5333-2-4.
[26] Roche DG, Kruuk LEB, Lanfear R, Binning SA. Public Data Archiving in Ecology and Evolution: How Well Are We Doing? PLOS Biol 2015;13:e1002295. https://doi.org/10.1371/journal.pbio.1002295.
[27] Stodden V, Seiler J, Ma Z. An empirical analysis of journal policy effectiveness for computational reproducibility. Proc Natl Acad Sci 2018;115:2584–9. https://doi.org/10.1073/pnas.1708290115.
[28] Hirschman L, Yeh A, Blaschke C, Valencia A. Overview of BioCreAtIvE: critical assessment of information extraction for biology. BMC Bioinformatics 2005;6:S1. https://doi.org/10.1186/1471-2105-6-S1-S1.
[29] Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016;3:160018. https://doi.org/10.1038/sdata.2016.18.
[30] Wise J. Implementation and relevance of FAIR data principles in biopharmaceutical R&D. Drug Discov Today 2019;24:6.
[31] Lannom L, Koureas D, Hardisty AR. FAIR Data and Services in Biodiversity Science and Geoscience. Data Intell 2020;2:122–30. https://doi.org/10.1162/dint_a_00034.
[32] NCBI Resource Coordinators. Database Resources of the National Center for Biotechnology Information. Nucleic Acids Res 2017;45:D12–7. https://doi.org/10.1093/nar/gkw1071.
[33] UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res 2019;47:D506–15. https://doi.org/10.1093/nar/gky1049.
[34] Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 2007;25:1251–5. https://doi.org/10.1038/nbt1346.
[35] Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N, et al. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Res 2009;37:W170-173. https://doi.org/10.1093/nar/gkp440.
[36] Ong E, Xiang Z, Zhao B, Liu Y, Lin Y, Zheng J, et al. Ontobee: A linked ontology data server to support ontology term dereferencing, linkage, query and integration. Nucleic Acids Res 2017;45:D347–52. https://doi.org/10.1093/nar/gkw918.

[37] Kim J-D, Wang Y. PubAnnotation - a persistent and sharable corpus and annotation repository 2012:4.
[38] Comeau DC, Islamaj Dogan R, Ciccarese P, Cohen KB, Krallinger M, Leitner F, et al. BioC: a minimalist approach to interoperability for biomedical text processing. Database 2013;2013:bat064–bat064. https://doi.org/10.1093/database/bat064.
[39] Giglio M, Tauber R, Nadendla S, Munro J, Olley D, Ball S, et al. ECO, the Evidence & Conclusion Ontology: community standard for evidence information. Nucleic Acids Res 2019;47:D1186–94. https://doi.org/10.1093/nar/gky1036.
[40] Gene Ontology Consortium. Gene Ontology Consortium: going forward. Nucleic Acids Res 2015;43:D1049-1056. https://doi.org/10.1093/nar/gku1179.
[41] Eilbeck K, Lewis SE, Mungall CJ, Yandell M, Stein L, Durbin R, et al. The Sequence Ontology: a tool for the unification of genome annotations. Genome Biol 2005;6:R44. https://doi.org/10.1186/gb-2005-6-5-r44.
[42] Chibucos MC, Zweifel AE, Herrera JC, Meza W, Eslamfam S, Uetz P, et al. An ontology for microbial phenotypes. BMC Microbiol 2014;14:294. https://doi.org/10.1186/s12866-014-0294-3.
[43] Kılıç S, White ER, Sagitova DM, Cornish JP, Erill I. CollecTF: a database of experimentally validated transcription factor-binding sites in Bacteria. Nucleic Acids Res 2014;42:D156–60. https://doi.org/10.1093/nar/gkt1123.
[44] Stenetorp P, Pyysalo S, Topic G, Ananiadou S, Aizawa A. Normalisation with the BRAT rapid annotation tool. Proc. 5th Int. Symp. Semantic Min. Biomed., Zurich, Switzerland: 2012, p. 4.
[45] Islamaj Doğan R, Kim S, Chatr-aryamontri A, Chang CS, Oughtred R, Rust J, et al. The BioC-BioGRID corpus: full text articles annotated for curation of protein–protein and genetic interactions. Database 2017;2017. https://doi.org/10.1093/database/baw147.
[46] Bird S, Loper E, Klein E. Natural Language Processing with Python. O'Reilly Media, Inc.; 2009.
[47] Cohen J. A Coefficient of Agreement for Nominal Scales. Educ Psychol Meas 1960;20:37–46. https://doi.org/10.1177/001316446002000104.
[48] Smith B, Ceusters W, Klagges B, Köhler J, Kumar A, Lomax J, et al. Relations in biomedical ontologies. Genome Biol 2005;6:R46. https://doi.org/10.1186/gb-2005-6-5-r46.
[49] Dublin Core (TM) Metadata Initiative. Dublin Core (TM) Metadata Terms 2020.
[50] Internet Engineering Task Force. RFC 8141 Uniform Resource Names (URNs) 2017.
[51] Jimeno A, Jimenez-Ruiz E, Lee V, Gaudan S, Berlanga R, Rebholz-Schuhmann D. Assessment of disease named entity recognition on a corpus of annotated sentences. BMC Bioinformatics 2008;9:S3. https://doi.org/10.1186/1471-2105-9-S3-S3.
[52] Herrero-Zazo M, Segura-Bedmar I, Martínez P, Declerck T. The DDI corpus: An annotated corpus with pharmacological substances and drug–drug interactions. J Biomed Inform 2013;46:914–20. https://doi.org/10.1016/j.jbi.2013.07.011.