Title

Comparing the Use of Research Resource Identifiers and Natural Language Processing for Citation of Databases, Software, and Other Digital Artifacts

Journal

COMPUTING IN SCIENCE & ENGINEERING, 22(2)

ISSN

1521-9615

Authors

Hsu, Chun-Nan
Bandrowski, Anita E
Gillespie, Thomas H
et al.

Publication Date

2020

DOI

10.1109/MCSE.2019.2952838

Peer reviewed

Comparing the Use of Research Resource Identifiers and Natural Language Processing for Citation of Databases, Software and Other Digital Artifacts

Chun-Nan Hsu[1], Anita Bandrowski[1,2], Thomas H Gillespie[1], Jon Udell[3], Ko-Wei Lin[1], Ibrahim Burak Ozyurt[1], Jeffrey S. Grethe[1], Maryann E. Martone[1,2]

[1]Department of Neurosciences and Center for Research in Biological Systems, University of California, San Diego, La Jolla, CA 92093-0608
[2]SciCrunch, Inc. San Diego, CA
[3]Hypothes.is, San Francisco, CA

Corresponding author: Maryann E. Martone, mmartone@ucsd.edu

Abstract

The Research Resource Identifier was introduced in 2014 to better identify biomedical research resources and track their use across the literature, including key digital resources like databases and software. Authors include an RRID after the first mention of any resource used. Here we provide an overview of RRIDs and analyze their use for digital resource identification. We quantitatively compare the output of our RRID curation workflow with the outputs of automated text mining systems used to identify resource mentions in text. The results show that authors follow RRID reporting guidelines well, and that our Natural Language Processing (NLP) based text mining was able to identify nearly all of the resources identified by RRIDs as well as thousands more. Finally, we demonstrate how RRIDs and text mining can complement each other to provide a scalable solution to digital resource citation.

Introduction

Research Resource Identifiers (RRIDs) are globally unique resolvable identifiers assigned to key research resources in the biomedical domain. RRIDs were introduced in 2014 to solve two fundamental problems in the biomedical literature: 1) The inability to identify what research resource was used in a given study; 2) the inability to track the use of resources across studies.

By research resource, we mean the key tools and reagents used by researchers in their experiments that are known to be sources of variation across experiments. Examples of resources that can be identified using the RRID system include biological resources such as antibodies, cell lines, plasmids, and organisms, but also digital tools such as databases, software for statistics and analysis, and other digital resources used in the research workflow.

For digital artifacts, the SciCrunch Registry supplies RRIDs. The SciCrunch Registry allows simple registration and classification of digital resources of all types, including databases, community portals, software tools, standards, and platforms, including commercial tools.

RRIDs differ both in their granularity and in the types of digital artifacts they identify from proposed recommendations for data and software citation (e.g., Data Citation Principles [1] Software Citation Principles [2]). The existing data citation systems are meant to point to a specific dataset with a persistent identifier, most commonly a DOI. In contrast, resources identified by RRIDs, including software tools, represent community resources developed and maintained by teams over many years.

The history of the RRID project is provided in detail in [3]. The project arose primarily out of the Neuroscience Information Framework (NIF) [4], [5] and its sister project, the NIDDK Information Network (dkNET; [6]). There are a wide variety of existing conventions for referencing a digital repository or its contents in the literature, e.g., URLs, reference to an article that describes the resource or free text. Because of this, a very simple question such as "How many people have used this resource?" cannot be answered without resorting to extensive manual labor and/or advanced Natural Language Processing (NLP) [7], [8]. To address this problem, NIF worked through FORCE11, a cross-disciplinary organization dedicated to transforming scholarly communication, to launch the Resource Identification Initiative (RII) [3], [9] to create a single unified standard for identifying and tracking the use of research resources in the scientific literature.

The Resource Identification Initiative working group at FORCE11 designed a syntax for RRID mentions:

"RRID:<prefix><Identifier>",

where <prefix> indicates the source registry and <identifier> is an accession number assigned by an independent registry that oversees a particular type of resource. For example, "RRID:SCR_003070" is a syntactically valid RRID for the software tool ImageJ, and "SCR" is the prefix of the SciCrunch Registry. Supplementary Table S1 provides a list of these registries, the resource categories that they cover, and the prefix for each (also available at http://tiny.cc/0a1y7y).

RRIDs are supplied by authors at the time of submission, review, or after acceptance of the manuscript. Over 120 journals now request RRIDs to be included as part of their instructions to authors (e.g., journals published by the Cell Press, *eLife*, *the Journal of Neuroscience*, *Endocrinology*, to name a few). In 2019, RRIDs were incorporated into the journal article tagging suite (JATS, ANSI/NISO Z39.96-2019), an XML standard used by the US National Library of Medicine and many publishers to mark up different parts of a scientific paper. JATS 1.2, released in May, 2019 (https://jats.nlm.nih.gov/publishing/tag-library/1.2/), includes advice for how to typeset RRIDs, see https://jats.nlm.nih.gov/publishing/tag-library/1.2/element/resource-id.html).

Table 1 shows the most recent statistics of the total number of RRIDs identified and curated by the RRID curation team and the number of digital resource RRIDs from the SciCrunch Registry ("SCR") as of June 4, 2019, as well as the number of

journals and the number of articles where the authors used RRIDs to cite their use of research resources. SCR RRIDs constitute about 18% of all RRIDs. The raw number includes also missing RRIDs supplied by curators, so the actual number supplied by authors is approximately 24,000. For digital resources, the number of RRIDs continues to grow every week at a rate of roughly 5-10 submissions per week.

	Count	Percentage
RRIDs from the curated database	192700	
SCR RRIDs from the curated database	34558	17.93%
Unique SCR RRIDs	2518	
Journals containing RRIDs	869	
Journals containing SCR RRIDs	466	53.62%
Papers containing RRIDs	13676	
Papers containing SCR RRIDs	7408	54.17%

Table 1: Statistics on the use of RRIDs as of June 4, 2019. Journals containing RRIDs: The count of journals found to have at least a paper containing at least a single RRID. Journals containing SCR RRIDs: The count of journals found to have at least a single paper containing at least a single SCR RRID and the percentage of these journals over the count of "Journals containing RRIDs." Papers containing RRIDs: the count of papers found to contain at least a single RRID. Papers containing SCR RRIDs: the count of papers found to contain at least a single SCR RRID and the percentage of these papers over the count of "Papers containing RRIDs."

Previously, we have seen the impact of RRIDs in improving identifiability of research resources [3] where papers that use RRIDs show 95% identifiability of resources used compared to ~50% without [10]. In this paper, we give an updated description of the overall RRID system and assess its effectiveness as an unambiguous indicator of resource usage by comparing the usage record of RRIDs that we collected through our RRID curation system with usage records gleaned from an NLP text mining system that identifies mentions of resources in the text of published articles. We will focus our analysis only on those RRIDs that point to digital resources.

Materials and Methods

Overview of RRID system and workflow

RRIDs for digital resources and services, e.g., core facilities, are issued by the SciCrunch Registry. Authors search for RRIDs through the Resource Identification Portal

http://scicrunch.org/resources

or one of the allied portals, e.g., dkNET or NIF, which also expose RRIDs. If authors are unable to find an RRID, they may submit the resource to the Registry through the Resource Identification Portal, NIF or dkNET. An SCR accession number is immediately issued, but the database is actively curated by a team at UCSD.

As with many long-lived registries, the types of accession numbers issued by the Registry at different points in its lifespan changed. The Registry maintains mappings between the various identifiers arising from the previous versions.

We maintain a resolution service, the SciCrunch Resolver, of the form:

https://scicrunch.org/resolver/RRID:SCR_XXXXXX.

Two versions exist: a human readable version, e.g.,

https://scicrunch.org/resolver/RRID:SCR_003070

and a machine-readable version, e.g.,

https://scicrunch.org/resolver/RRID:SCR_003070.xml.

The first is useful for authors and readers and allows the viewing of aggregated data on resolver pages maintained by the SciCrunch platform, or redirects to original records in the core registries. The second only returns metadata maintained by SciCrunch.

RRIDs are also resolvable through Identifiers.org, Name-to-Thing (N2T). They are also available in Cross Ref's Event data

(https://www.crossref.org/services/event-data/),

which maintains relationships between DOIs and other digital artifacts.

SciBot RRID Curation Pipeline

Approximately 120 journals ask authors for or require RRIDs and 67 actively engage typesetters who ensure the RRIDs syntax is correct. For many journals, compliance is voluntary. These journals provide RRID instructions to authors but do not typeset the RRID. A team at UCSD actively monitors the published literature for RRIDs and maintains a curated dataset of RRIDs, visible through the resolver service and as a service to each resource provider.

To assist in human curation, we developed SciBot (RRID:SCR_016250), a semi-automated curation tool that streamlines the process of validating RRIDs in published papers using the Hypothes.is (RRID:SCR_000430) web annotation platform. The pipeline is described in more detail in [11]. For example, PMID: 31112613 describes the use of a bioinformatics software tool "Jellyfish, RRID:SCR_005491" in the text. This mention of the use of the software tool "RRID:SCR_05491" is automatically annotated by SciBot, which then calls the resolver API to retrieve metadata and related papers about the resource as an annotation for a human

curator to review. Once the curator confirms the mention, a record of the mention of this RRID in the paper will be saved.

We have been using Hypothes.is and the SciBot workflow since February 2016 for RRID curation. Annotations are exported from Hypothes.is into an annotation database maintained locally. Data are submitted to Cross Ref's Event database (RRID:SCR_016281); data for individual papers and RRIDs can be found via the Event API example.

Curators find papers with RRIDs by searching Google Scholar (RRID:SCR_008878) and PubMed (RRID:SCR_004846) for "RRID:". They install the SciBot bookmarklet in their web browsers and activate it to annotate the HTML version of the article. Importantly, Hypothes.is allows users to attach tags to an annotation. For the purpose of curating RRIDs, we developed a set of tags to guide and manage a team of curators. Definitions of these tags are given in Supplementary Table S2 (also available at http://tiny.cc/0a1y7y).

RDW Text Mining Pipeline

To extract mentions of digital research resources independent of the RRID system, we utilized RDW (RRID:SCR_012862) [8], a text analysis tool suite that uses Named Entity Recognition (NER) to extract resource entities from longer text documents, focusing on digital resources. A word like "ImageJ" is relatively unambiguous, but many other software tool names, such as "David," are ambiguous. Therefore, the RDW system recognizes tool names in sentence context using a Conditional Random Field model that enables recognition of tool names beyond those provided in training data. RDW extracts tool mentions from the methods section of each paper. In this way, RDW searches for papers that are much more likely to have *used* a tool as opposed to papers that simply *discuss* a tool. The RDW pipeline also extracts URLs and RRIDs referring to a resource, but in this case, RDW uses simple pattern matching and the SciBot regular expression for URLs and RRIDs, respectively, across the full text, including footnotes.

The text corpus that RDW searched contained 2,341,133 articles from the open access subset of PubMed Central and 738,910 articles extracted from 79 Elsevier journals through Elsevier's text and data mining API service, 72,493 from 70 journals from Springer-Nature's API, and 151,784 from 29 Wiley journals that were provided directly from Wiley as a part of a collaboration agreement. RDW recognized mentions of digital resource names, RRIDs or URLs from a total of 701,110 articles.

The RDW text mining tool's accuracy has been rigorously evaluated as reported in [8]. A new estimation of its correctness rates using an independently collected corpus of thousands of annotated resource mentions showed that given 90%/10% train/test split, RDW yielded a precision of 94.1% and recall of 85.8% and F1 of 89.8% for resource named entity recognition.

Comparison of Datasets Acquired via Text Mining vs RRIDs

To assess the impact of the RRID we created three datasets labeled SciBot-Curator, RDW and RRID-by-RDW respectively. Table 2 compares the composition of the contents of the sources where the three datasets were acquired. Table 3 shows the statistics of the size of the resulting datasets, which are available for download from Zenodo (RRID:SCR_004129) at https://doi.org/10.5281/zenodo.3241632.

SciBot-Curator dataset

The SciBot-Curator dataset contains the curation records of digital resources collected through the curation pipeline by our curators. This dataset was retrieved from our curation record repository database on May 16, 2019 for this comparative study. 26,748 total RRID mentions for digital resources and services were found in 7,268 articles from 462 journals. Note that these numbers also include the missing RRIDs supplied by curators which were used to compare the performance of human curators and RDW.

Each record was converted to contain pairs of the standard forms of PMID (PubMed ID) and RRID to facilitate comparison. Many records from the RRID curation dataset only have a DOI instead of a PMID. They were converted if possible and were discarded if no PMID was available for the paper, because the records from the text mining dataset use PMID. A total of 310 articles and 1,328 records were discarded from the original total of 26,749 records. Another 2 records were removed from the analysis because their RRID accession number did not conform with the SCR prefix. We then removed duplicate records to obtain a dataset of 25,224 records of distinct triples of PMID, RRID and curator tags.

Among these triples, we have 23,745 distinct pairs of PMID and RRID (Table 3). This number is smaller than the number of distinct triples (25,224) because an RRID may be mentioned in a paper (PMID) multiple times and each mention may have a different curation tag.

RDW dataset

The RDW dataset contains the records of resources mentioned in papers that were identified by RDW. This dataset primarily comprises resource name mentions extracted by NER, but also contains resources identified by URL matching and RRIDs. We retrieved the data on May 6, 2019 by issuing queries to an Elasticsearch endpoint that supports the Research Information Network in dkNET populated by the Foundry scalable data integration system [12]. From these articles, 1,599,963 records of digital resource and PMID pairs were identified.

RRID-by-RDW

We extracted a dataset comprising the records of the RRID mentions identified by RDW using the SciBot regular expressions (regex) to match the pattern "RRID:<accession#>" appearing in a published article. This dataset, dubbed "RRID-by-RDW," provides an unprecedented direct comparison of the use of an ID system with the text mining/NLP approach to

identify mentions by authors. The raw data contains 64,549 records. However, among them 11,377 do not have a PMID and only 7,094 are "SCR" records. We removed 24 records that cannot be mapped to any PMID to obtain a total of 7,070 records.

While access to the full texts of the biomedical corpus would benefit our ability to monitor and analyze resource usage across biomedicine, our ability to do so is still extremely limited. In our analysis, the curated dataset had many more RRIDs, papers and journals represented than the RRID-by-RDW data set (Table 3), because the curators have access to closed access papers through our institutional subscriptions, while the RDW must rely primarily on the open access subset of PubMed Central for text mining. Also, mapping across the 3 identifier systems: PMIDs, PMCIDs and DOIs, in biomedicine is still quite difficult.

Measuring Differences

We then used the SciBot-Curator dataset as the ground truth of whether a resource (RRID) was used in a published study. From the ground truth we could evaluate the correctness rates of RDW overall and RDW by RRID pattern matching. More specifically, if a PMID and RRID pair record appears in the RDW dataset and in the SciBot-Curator dataset, then we counted it as true positive (TP), unless the SciBot-Curator's curator tag says otherwise (see below). Similarly, we can define false positive (FP), false negative (FN) and true negative (TN) as follows.

- TP := if a record of PMID and RRID pair in RDW matches at least a record in the SciBot-Curator dataset with identical PMID and RRID.
- FP := if a record of PMID and RRID pair in RDW matches one or more records in the SciBot-Curator dataset with identical PMID, implying they both considered that paper, but matches ***no*** record in the SciBot-Curator dataset with an identical RRID among those with an identical PMID.
- FN := if a record of PMID and RRID pair in SciBot-Curator matches one or more records in the RDW dataset with an identical PMID, implying they both considered that paper, but matches ***no*** record in RDW with an identical RRID among those with an identical PMID.
- TN := Not considered. A well-defined true negative depends on curator tags.

These quantities were calculated by importing the datasets into a relational database and querying the number of records that share their PMIDs and RRIDs.

Most records in the SciBot-Curator dataset do not have any tag, meaning that human curators had no problem with the RRID mentions identified by SciBot. When human curators did annotate a SciBot identified RRID mention with tags, we compared records with tags with those in RDW and calculated correctness rates by the following definitions:

- TP := if a record in RDW matches a record in the SciBot-Curator dataset with a tag that is ***not*** "RRIDCUR:Incorrect" or "RRIDCUR:InsufficientMetaData."
- FP := if a record in RDW matches a record in the SciBot-Curator dataset with a tag that is either "RRIDCUR:Incorrect" or "RRIDCUR:InsufficientMetaData."
- FN := if a record of PMID and RRID pair in SciBot-Curator with a tag that is ***not*** "RRIDCUR:Incorrect" or "RRIDCUR:InsufficientMetaData," matches one or more records in the RDW dataset with an identical PMID, but matches no record in RDW with an identical RRID.
- TN := if a record of PMID and RRID pair in SciBot-Curator with a tag that is "RRIDCUR:Incorrect" or "RRIDCUR:InsufficientMetaData," matches one or more records in the RDW dataset with an identical PMID, but matches no record in RDW with an identical RRID.

When a curator annotates a SciBot identified RRID mention as "Incorrect" or "InsufficientMetaData," it means that the RRID mention is either incorrect or impossible to verify due to insufficient metadata provided by the authors. Therefore, those RRID mentions are not legitimate and when RDW identifies those resources as mentioned in the paper, it incorrectly identifies the resource based on the RRID, and should therefore be deemed as a false positive (FP).

For example, in this snippet from PMID: 29540552, the authors specified an incorrect but well-formed RRID for ImageJ:

Fluorescence was visualized using a Leica TCS SP2 confocal microscope equipped with a 405 nm diode laser. The mean fluorescence intensity was quantitated using ImageJ software (RRID:SCR_001775).

The SciBot-Curator dataset contains a record:

["29540552", "RRID:SCR_003070", "RRIDCUR:Incorrect"]

Meanwhile, the RDW dataset contains this record:

["29540552", "RRID:SCR_003070"]

The record states that RDW recognized "ImageJ" in exactly the same paper and linked it to its RRID, but did not recognize that the author specified RRID mismatched. For our comparative study here, we counted this case as a false positive to highlight the difference between curation and text mining.

Other tags arise when authors' citation of RRID is not perfect, including "MetadataMismatch", "Duplicate", and "SyntaxError," but the use of the resource referred to by that RRID in the study was stated, and therefore, when RDW also identifies that resource, it should be considered to be correct and as a true positive (TP). "Unresolved" is used to trigger a discussion among curators and implies that the use of the resource was stated, but not picked up by the current version of the tool.

Results

Comparison of RRID and NLP

We prepared three datasets: 1) records from SciBot-Curator' curation results (SciBot-Curator); 2) RDW's text mining results (RDW), and 3) RRID pattern matching results (RRID-by-RDW), to compare matches and mismatches among the research resources identified by different approaches to assess their strengths and weaknesses. The matches and mismatches were quantified by using the SciBot-Curator dataset as the ground truth to evaluate whether RDW and RRID-by-RDW identified the same research resources as by the curation pipeline.

Table 4 shows the counts of TP, FP, FN and FN as defined in the Materials and Methods section, and Table 4 shows the comparison of RDW and RRID-by-RDW in terms of recall, precision, and F1-score against the SciBot-Curator dataset as the ground truth. The number shows that RRID-by-RDW matches SciBot-Curator closely and outperforms RDW NER by a significant margin. However, the number of resources identified in the RRID-by-RDW pipeline is considerably less than those identified by the curators, which may reflect differences in the corpora used for these 2 data sets (see Conclusions and Future Work).

Equipped with a matching algorithm by Machine Learning more flexible than simple pattern matching, RDW accomplished a higher recall than RRID-by-RDW by about 6% because it also detects resource names regardless of whether they have RRIDs. However, while formal evaluation of RDW's correctness rates against test benchmarks as reported in the publications are high, when compared to records in the SciBot-Curator dataset, the false positive by RDW is high (10+49+137+10271 in Figure 1), suggesting that RDW identified too many research resources in the same set of articles compared to SciBot-Curator' records.

	RDW	RRID-by-RDW
Recall	**0.9703**	0.9135
Precision	0.6537	**0.9775**
F1-score	0.7811	**0.9444**

Table 4: Comparison of the correctness rates of RDW vs. SciBot-Curator and RRID-by-RDW vs. SciBot-Curator.

Quality Control

The statistics for the RRID curator tags from the SciBot-curator dataset shows how authors reported RRID of digital resources in their publications (Table 5, first two columns). "Missing" is the top issue, where authors did not report an RRID for the resource that they used, constituting 41% of all RRID mentions identified by curators through the SciBot-assisted curation pipeline. Our initial investigation suggests that most of them are from journals that only ask for RRIDs for a subset of resource categories in their instructions to authors, e.g., organisms and antibodies, but not digital resources. Following "Missing" are "Duplicate" and "Unrecognized." Both constitute less than 10%. The numbers of dubious resources, i.e., "InsufficientMetadata" and "Incorrect" tags, are small. The use of the"Validated" tag is also low but we have noted that curators tend to use tags when there is a problem rather than when everything is correct. Overall, the results suggest that authors follow the journal instructions for RRID reporting well, formating the RRID mostly correctly (0.15% has a syntax issue) and fitting the specification. Unresolved RRIDs are rare (0.1%).

Table 5 also shows how well RDW and RRID-by-RDW match the records of the SciBot-Curator dataset in the presence of various curator tags. Note that mentions tagged by "Incorrect" and "InsufficientMetaData" are deemed not legitimate. If RDW or RRID-by-RDW identify those cases as legitimate, they will be counted as false positives. Thus, the lower the numbers and ratios the better for these tags. Otherwise, it is desirable that the numbers and ratios are high. From the table, RDW was able to recover a large proportion of "Missing" and "Unresolved" mentions but missed more than half of "SyntaxError" RRID mentions. In that case, even RRID-by-RDW did not recover many. The ratios of matched "Incorrect" and "InsufficientMetaData" by RDW are adequate but with room for improvement.

Conclusions and Future Work

Use of RRIDs for digital resources

This study represents the first in depth analysis of patterns of RRID usage for digital resources across a large number of papers. RRIDs are supplied by authors. The number of problematic RRIDs, including those tagged as "Unrecognized," "Unresolved," "Misplaced," and "SyntaxError," is very small, representing less than 3% of the total (Table 5), consistent with our earlier analysis of a much smaller sample in the pilot study [9]. Both of these suggest that authors are able to comply with the instruction and that they are careful when assigning RRIDs to their resources. As more and more tools are developed to support the use of RRIDs, we expect these errors to diminish. For example, *eLife* currently uses a version of SciBot to assess and verify RRIDs supplied by authors (*eLife* Blog http://tiny.cc/su1y7y).

Comparison of RRIDs vs. NLP

We sought to answer whether the use of the RRID system presented any advantage over the use of modern NLP based methods for accurate assessment of resource use in the literature.

We analyzed the dataset generated by our RDW pipeline that uses machine learning and NLP to detect resource mentions in the biomedical literature. The results show that RDW was able to identify nearly all of the resources identified by RRIDs as well as thousands more. However, comparison to the curated data set showed that it tagged too many resources that were not considered as resource mentions by human curators. Many factors may contribute to these large numbers of false positives, including errors made by RDW and resources detected by either curators or SciBot. Nevertheless, the results point to promising directions of using these two tools together to improve the curation process by assisting curators in identifying resources that are missing RRIDs.

The results by RRID-by-RDW illustrate the advantage of the use of an ID system such as RRID to identify mentions in the publications. Because RRIDs were designed to be uniform across publishers, the results here show that with access to the full text of an article, pulling out statistics of resource mentions based on RRIDs can be performed accurately with relatively simple text mining. NLP for NER is very computationally intensive. In contrast, when RRIDs are present, resource mentions can be extracted with much simpler regular expressions, making the system tractable for the millions of articles published in biomedicine every year. We do note, however, that malformed RRIDs require additional effort to detect.

Outlook

The use of RRIDs has grown steadily for identification of research resources in biomedicine and has expanded to include additional types of resources, e.g., plasmids. In 2019, RRIDs were incorporated into the journal article tagging suite (JATS), signaling that the academic publishing community has accepted RRIDs as a standard method for tagging research resources.

An intriguing question is whether RRIDs can be employed outside of biomedicine. We hope our experiences with introducing and using the RRID will help other disciplines replicate its success and build upon it, while acknowledging that each domain likely presents unique challenges.

References

[1] M. Fenner et al., “A data citation roadmap for scholarly data repositories,” Sci Data, vol. 6, no. 1, p. 28, Apr. 2019.

[2] A. Smith, D. Katz, and K. Niemeyer, “Software citation principles. PeerJ Computer Science 2: e86.” 2016.

[3] A. E. Bandrowski and M. E. Martone, “RRIDs: A Simple Step toward Improving Reproducibility through Rigor and Transparency of Experimental Methods,” Neuron, vol. 90, no. 3, pp. 434–436, May 2016.

[4] D. Gardner et al., “The neuroscience information framework: a data and knowledge environment for neuroscience,” Neuroinformatics, vol. 6, no. 3, pp. 149–160, Sep. 2008.

[5] J. Cachat et al., “A survey of the neuroscience resource landscape: perspectives from the neuroscience information framework,” Int. Rev. Neurobiol., vol. 103, pp. 39–68, 2012.

[6] P. L. Whetzel, J. S. Grethe, D. E. Banks, and M. E. Martone, “The NIDDK Information Network: A Community Portal for Finding Data, Materials, and Tools for Researchers Studying Diabetes, Digestive, and Kidney Diseases,” PLoS One, vol. 10, no. 9, p. e0136206, Sep. 2015.

[7] Y.-H. Huang, P. W. Rose, and C.-N. Hsu, “Citing a Data Repository: A Case Study of the Protein Data Bank,” PLoS One, vol. 10, no. 8, p. e0136631, Aug. 2015.

[8] I. B. Ozyurt, J. S. Grethe, M. E. Martone, and A. E. Bandrowski, “Resource Disambiguator for the Web: Extracting Biomedical Resources and Their Citations from the Scientific Literature,” PLoS One, vol. 11, no. 1, p. e0146300, Jan. 2016.

[9] A. Bandrowski et al., “The Resource Identification Initiative: A cultural shift in publishing,” F1000Res., vol. 4, p. 134, May 2015.

[10] N. A. Vasilevsky et al., “On the reproducibility of science: unique identification of research resources in the biomedical literature,” PeerJ, vol. 1, p. e148, Sep. 2013.

[11] Z. Babic et al., “Incidences of problematic cell lines are lower in papers that use RRIDs to identify cell lines,” Elife, vol. 8, Jan. 2019.

[12] I. B. Ozyurt and J. S. Grethe, “Foundry: a message-oriented, horizontally scalable ETL system for scientific data integration and enhancement,” Database , vol. 2018, Jan. 2018.

Acknowledgements

This work was supported by NIH grant U24DK097771 supporting the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) Information Network (dkNET, https://dknet.org) and NIH’s National Institute of Drug Abuse award U24DA039832 supporting the Neuroscience Information Framework (http://neuinfo.org).

Conflicts of Interest

Anita Bandrowski, Maryann Martone and Jeffrey Grethe have an equity interest in SciCrunch, Inc., a company that develops services and tools based on RRIDs that may potentially benefit from the research results. Drs. Bandrowski and Martone are employed by the company. The terms of this arrangement have been reviewed and approved by the University of California, San Diego in accordance with its conflict of interest policies.

Dataset	Source	File type	Paper ID	Category	Identifier	Method	Curated?
SciBot-Curator	Google scholar, PubMed search, etc.	HTML, PDF	DOI	All: antibody, cell line, organism, digital resource (SCR) etc.	RRID (e.g., "RRID: SCR_003070")	SciBot RRID curation pipeline	Yes
RDW	PubMed Central (PMC), Elsevier, Springer-Nature API, Wiley	XML	PMID	Digital resource (SCR)	Text name (e.g., "ImageJ") and URL	NER text mining	No
RRID-by-RDW	PubMed Central (PMC), Elsevier, Springer-Nature API, Wiley	XML	PMID	Antibody and Digital resource	RRID	SciBot regex	No

Table 2: Comparison of the content composition of the sources where the three datasets were created for this study.

Dataset	Total mentions (triples of PMID and RRID and tags/context)	Distinct pairs of PMID and RRID	Distinct PMID	Distinct RRID
SciBot-Curator	25,224	23,745	6,866	2,344
RDW	1,599,963	1,599,963	701,110	9,047
RRID-by-RDW	7,070	6,994	1,747	1,429

Table 3: Statistics of the final filtered datasets ready for comparison. We note that due to the data processing steps prepared for the study, the numbers of SciBot-Curator shown here are different from the June 4, 2019 dataset shown in Table 1, which presents the most recent raw data of the use of all RRIDs and SCR RRIDs for an overview.

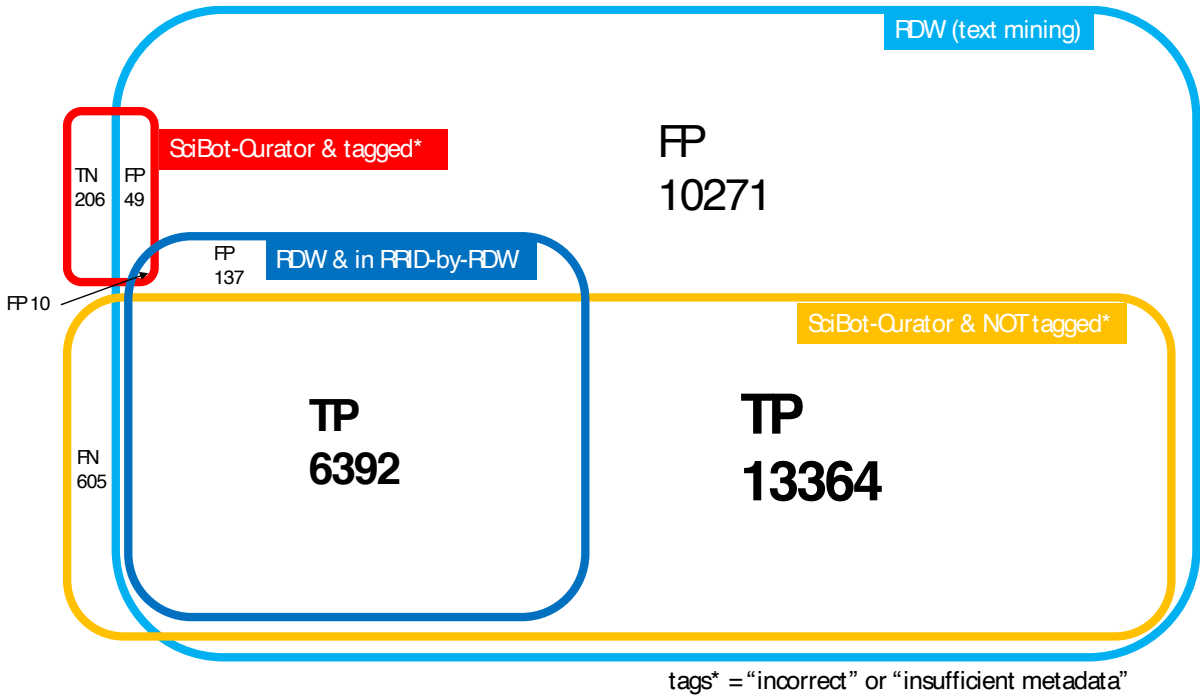

Figure 1: Match and mismatch counts of PMID-RRID pairs among the datasets from the SciBot-Curator, RDW, and RDW & in RRID-by-RDW. The Venn diagram illustrates how the counts were determined. Since the pairs in the SciBot-Curator without the "incorrect" or "insufficientMetaData" tags were considered the ground truth, pairs by RDW or RDW-by-RRID match the ground truth are TP (true positives), otherwise they are FP (false positives). The pairs tagged as "incorrect" or "insufficientMetaData" are negatives. The diagram shows that RDW captures nearly all the ground truth but also many false positives.

Tag	SciBot-Curator Count	RDW Count/Total	RRID-by-RDW Matched	Matched/Count	Matched		
Missing	9,801	0.4128	8,709	0.8886	61		
Duplicate	972	0.0409	781	0.8035	414		
Unresolved	25	0.0011	22	0.8800	0		
SyntaxError	36	0.0015	16	0.4444	5		
Unrecognized	621	0.0262	515	0.8293	265		
Misplaced	2	0.0001	0	0.0000	0		
Validated	39	0.0016	25	0.6410	4		
Incorrect	102	0.0043	24	0.2353	10		
InsufficientMetaData	337	0.0142	35	0.1039	0		
Total (distinct)	23,745						

Table 5: Statistics of the use of Hypothesis curation tags for RRID mentions in the SciBot- curators dataset and the number of matches by RDW and RRID-by-RDW under the presence of curator tags for RRID mentions. Note that "Total (distinct)" is the total number of distinct PMID-RRID pairs but each PMID-RRID pair may have zero, one, or two differnt tags.