Assessing Value of Biomedical Digital Repositories*

Chun-Nan Hsu[1], Anita Bandrowski[2], Jeffrey S. Grethe[3], Maryann Martone[3]

[1]Division of Biomedical Informatics, Department of Medicine, University of California, San Diego, La Jolla, CA, USA.

[2]SciCrunch Inc. San Diego, CA, USA.

[3]Center for Research in Biological Structure, University of California, San Diego, La Jolla, CA, USA.

Corresponding Author:

Chun-Nan Hsu

Email address: chunnan@ucsd.edu

* This manuscript was originally prepared as a response to *NOT-OD-16-133 Request for Information (RFI): Metrics to Assess Value of Biomedical Digital Repositories* by the National Institute of Health. C.-N. Hsu presented the contents orally in the 2016 workshop of collaborative data projects, Taipei, Taiwan, December 8, 2016. http://odw.tw/2016/

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2688v1 | CC BY 4.0 Open Access | rec: 3 Jan 2017, publ: 3 Jan 2017

Abstract

Digital repositories bring direct impacts and influence to the research community and society but at the moment it is challenging to objectively measure their value. We distinguished the difference between impacts and influence and discussed measures and mentions as the basis of a quality metric of a digital repository. It is challenging to define a single perfect metric that covers all quality aspects. We argue that these challenges may potentially be overcome through the introduction of standard resource identification and data citation practices. We briefly summarized our research and experience in the Neuroscience Information Framework, the BD2K BioCaddie project on data citation, and the Resource Identification Initiative. We outline our accomplishments and challenges ahead. Full implementation of these standards will depend on cooperation from all stakeholders --- digital repositories, authors, publishers, and funding agencies, for which we have been gaining support with endorsements and resource investments.

Impact vs. Influence

Assessing the value of digital repositories shares many similar challenges to assessing the value of any scholarly work. One of them is whether to distinguish between direct impact and broad influence. By direct impact we refer to actual changes that the work brings to the field in terms of outcomes, practices, and methodologies. In biomedical sciences, these include, for example, new drugs, new models of molecular interactive pathways, new experimental methods, etc. By influences, we refer to how widely the work has been disseminated and viewed across a broad community so that a work can influence other work, either by inspiring new research ideas or preliminary testing of hypotheses. Impact and influence may be correlated but that is not always the case. A highly influential work may have a low impact and vice versa. A digital repository may have a high influence in that it is viewed many times, but low impact in that there is no evidence that the actual products are used to advance science. However, the products may be very useful for educational purposes. The converse is also true; a digital repository may not be well known across a wide swath of the community, but its products may be highly impactful in a smaller community. Understanding where each resource fits and therefore how to evaluate their success and perhaps improve both dimensions requires that it be possible to measure these in some objective and preferably automated or semi-automated way.

Measure vs. Mention

While traditional metrics of a scientific work are based on citations -- whether the work is *mentioned* in scientific publications, digital repositories allow *measures* through the count of access in different ways, URL connections, data transferring, etc. One may argue that measures of access more accurately reflect the value of a digital repository for without access, a digital repository is not used and cannot create values. However, as we discussed above, the value of a work may present as impact or influence. Usually, mention-based metrics, such as citations, reflect influence better, for a work can be mentioned only after it is known. However, citations can also reflect actual use of the resource within a published study. Currently, both are hard to track; this makes proper citation of data products in the literature extremely important.

Measure and mention are not correlated all the time for a digital repository (Huang et al. 2015; Huang 2016; Rose & Hsu 2016). Moreover, different measure-based metrics, for example, URL connection count, and FTP download count, size of data transferring, are not always correlated. This applies not only when comparing digital repositories but also when comparing content units within a digital repository. Results in (Huang 2016; Rose & Hsu 2016) show that ranking protein structures in RCSB PDB (Protein Data Bank), a data repository of protein structure data, by different measures of access give uncorrelated results. In the study, we ranked protein structures according to their frequencies of Web accesses (http views) and FTP accesses (file downloads). We found that the top 20 of the two resulting ranked lists share no protein structures. Moreover, the two frequencies are not correlated, in the sense that a protein structure that is highly accessed by Web browsers is not necessarily highly accessed by FTP, and vice versa.

Meanwhile, in addition to citations in publications, mention-based metrics may include citations in press reports, blogs, social media, and other forms of publications, currently measure by services such as Altmetrics (Altmetrics 2016). These may not be correlated either, and may better reflect the influence of a work than its impact. Citations may be in different forms, including directly mentioning various names of a digital repository, citing the publications describing a digital repository or mentioning the URL links to a digital repository. For example, an author may cite RCSB PDB by its various publications, URL links to its portal Web page (with different versions throughout the years after it went online), PDB IDs or URL links of protein structures.

Authors not only cite RCSB PDB in different forms, the annual growth rates of the counts of these different citations forms are not correlated, for either data repository as a whole (Huang et al. 2015), or for protein structures (Huang 2016; Rose & Hsu 2016). Authors most frequently chose to cite publications, because usually that is how repositories instruct authors to do in a "how to cite us" page. However, URL link mentions are growing rapidly. Though the PDB ID is designed as a unique ID to mention specifically to a protein structure in PDB, the ID itself is not globally unique without a prefix, and may coincide with a wide variety of entities (Rose & Hsu 2016). PDB IDs are always 4 characters in length. The first character is a numeral in the range 1-9, while the last three characters can be either numerals (in the range 0-9) or letters. Examples of other IDs and/or entities matching this format include "1USD" as currency, "2NO3" as a chemical compound, and "1E10" as a floating-point number; while "1USD", "2NO3" and "1E10" are all legitimate PDB IDs.

Table 1 shows all the issued PDB IDs presented in full-text format articles. The statistics was obtained from publications containing mentions of PDB ID from the PubMed Central (PMC), where we obtained 1,015,179 articles in NXML format, and 1,093,980 articles in plain text format as of August 2015. Removing duplicate PMC IDs yielded a total of 1,015,233 articles. Table 2 compares the top 10 PDB protein structures by the frequency of PDB ID mentions and the top 10 ordered by the frequency that their original publications were cited in the references by subsequent articles in the PubMed. The two lists share only two PDB protein structures (2RH1 and 2A79), suggesting that high PDB ID mentions and high publication citations are not necessarily correlated (Huang 2016).

Standardization of Mentions and Use

Currently, one of the most difficult problems facing assessments of digital repositories is the lack of formal systems of citation that allow measures of influence and direct impact to be calculated using modern information technology. As documented by (Huang et al. 2015), the current means of referencing a digital repository or its content in the literature or any other work involve a range of styles including URLs, reference to a particular article describing the resource, accession numbers and free text. Because of this, a very simple question like: how many people have documented use of this resource cannot be answered without resorting to extensive manual labor or advanced natural language processing (NLP) (Rose & Hsu 2016; Ozyurt et al. 2016).

Through the Neuroscience Information Framework and the Data Citation Working groups at FORCE11, we've successfully worked to change this by developing and promoting standards for both resource use and data citation, with a focus on the literature.

Perspectives from the Neuroscience Information Framework

The Neuroscience Information Framework (NIF) has been cataloging and tracking the digital research resource landscape for over 8 years. We maintain a large database that tracks how a resource has evolved over the years, including whether it is no longer in service. Currently, a relatively small number of resources (229 as of Oct 17, 2016 (11)) are completely out of service; many more, however, grow stale over time. Over time, we have developed some criteria for determining whether a resource is vibrant and growing or moribund: 1) when was the last time a web page was updated on the site; 2) when was the last time data were added; 3) Do the data represent a significant fraction of data available in a community or a very limited amount? 4) When a resource is down, does anyone complain? We call the latter the "squawk factor".

The Resource Identification (#RRID) Initiative RRID (Bandrowski et al. 2016; https://scicrunch.org/resources) is designed to help researchers sufficiently cite the key resources used to produce the scientific findings reported in the biomedical literature. A diverse group of collaborators are involved in the project, including the Neuroscience Information Framework which launched and has been leading the initiative, the Oregon Health & Science University Library which contributed to the early pilot project, with the support of the National Institutes of Health and the International Neuroinformatics Coordinating Facility. Resources (e.g. antibodies, model organisms, cell lines and digital tools) reported in the biomedical literature often lack sufficient detail to enable reproducibility or reuse. For example, catalog numbers for antibody reagents are infrequently reported, and the version numbers for software programs used for data analysis are often omitted. The issue is similarly applied to other types of digital repositories.

The Resource Identification Initiative aims to enable resource transparency within the biomedical literature through promoting the use of unique Research Resource Identifiers (RRIDs). In addition to being unique, RRID's meet three key criteria, they are:

1. Machine readable.
2. Free to generate and access.
3. Consistent across publishers and journals.

RRID's depend on comprehensive resource registries which provide an authoritative source for each resource type. Each is covered by a different database, e.g., the Antibody Registry, the SciCrunch (NIF) Resource Registry. These databases were aggregated and made available through the Resource Identification Portal (https://scicrunch.org/resources), supporting NIH's new guidelines for Rigor and Transparency in biomedical publications. The portal aims to promote research resource identification, discovery, and reuse and offers a central location for obtaining and exploring RRIDs. The current number of digital tools, including databases and software projects, listed in the Registry is over 13K (Bandrowski et al. 2016). The number of antibodies is > 2M and model animals are in the hundreds of thousands.

The project has been running since 2014. Currently, over 1226 papers have appeared with RRID's from over 160 biomedical journals. Cell Press has just adopted the standard (http://www.cell.com/star-methods) and eLife and the Endocrine Society just announced that they will be strongly encouraging authors to use RRID's in their journals.

RRID's provide the means for users to unambiguously the resources used within a study in their publication. Authors are asked to insert RRID's for resources *used* in their studies after the first reference to the resource in the materials and methods. To ensure that RRID's are easily identified and extractable from the literature, authors are asked to prepend the namespace RRID: before using the database accession number. Thus, RRID's specifically target the use of resource resources as opposed to mentions in an introduction or discussion. A simple search through Google Scholar for an RRID will return papers that have used a particular resource, e.g., 6 articles have appeared to date that used the PDB (Google Scholar 2016).

RRID's also provide a convenient means for authors to access digital resources used in papers. Research resource providers can update the registry in the portal when there is a need to transfer the data and software to another repository, but the RRID will remain the same to ensure that readers can always locate the data and software through a centralized registry. This new approach solves data access, sharing, archiving, and preservation at the same time. In addition, it provides a standard citation format that can be easily extracted to show what resources were used in a particular published study - allowing for measurement of impact.

Since maintaining a correct reference of the RRID increases visibility and thus influence of a research resource, and will bring direct impact eventually, providers of research sources will be highly motivated to maintain its correctness, closing a healthy positive feedback loop to sustain the whole system.

Data Citation Implementation Pilot Project (https://www.force11.org/group/dcip). RRID's address the citation of digital repositories and associated tools at a high level; however, we also need a system to cite individual data sets that may include only a subset of data in a repository or be assembled from multiple data sources. Precisely referring to which subset of data is retrieved and used can be a computationally intractable problem, which leads to some pessimistic views with regard to data citation (Buneman et al. 2016).

We argue that the ultimate purpose of data citation is not only to identify precisely a data subset for facilitating reproducibility, but also to ensure that both the individuals contributing data and the repositories housing them receive proper credit and attribution, as specified in the Joint Declaration of Data Citation Principles (JDDCP, Data citation 2014). The JDDCP has been endorsed by 253 individual scientists and 114 organizations, representing different sectors of stakeholders, including data centers/data repositories, educational institutions, funding agencies/organizations, libraries, publishers, registries/social networks/research networks, societies/associations/consortiums, and technology providers.

Based on the eight principles given in JDDCP, FORCE 11 and other groups have been working on developing practical standards to implement data citations. One of these is the Data Citation Implementation Pilot Project (DCIP) as part of the NIH BD2K bioCADDIE (http://biocaddie.org) project that we have been working on. The primary goal is to provide basic coordination between publishers, repositories and identifier / metadata services for early adopters of data citation according to the JDDCP. To meet this goal we will provide authoritative guidance and group consultation on data citation implementation to help establish one or more benchmark implementations of data citation based on the JDDCP and Starr et al 2015 (Starr et al. 2015), its cross-domain implementation guidance.

The key ideas here include working with data repositories on best practices that repositories can follow to support data citation with the support of community metadata standards, the use of persistent identifiers (e.g., DOI's), and machine-readable landing pages, which provide essential information on the content and accessibility of data within the data repository. A landing page allows for an access point that is independent from any multiple encodings of the data that may be available (Starr et al. 2015), and thus avoids the complicated computational problem of citing arbitrary subsets of data precisely, as described in (Buneman 2016). A landing page can also provide information on access controls required by licensing or privacy considerations. In addition, user requested landing pages can be minted for custom data aggregations as well.

We are often asked how RRID's differ from the referencing of a specific data sets as proposed by the JDDCP. The issue is one of granularity. RRID's are meant to identify the parent entity like the PDB, while additional identifiers may be used to identify the specific data set used. This more granular data citation may comprise a subset of a data repository or a supraset across repositories. The RRID essentially functions as an ORCID to identify the organizational entities involved, e.g., the data repository, while the DOI points to a specific and unique data set.

Towards Reliable and Accurate Metrics

Though counting frequencies of standardized RRID mentions and data citations might not be the single perfect metric of the value of a digital repository, wide adaptation of these standards will definitely lead to a more reliable and comparable metric than the status quo and open up development of more sophisticated metrics like the h-index (Hirsch 2005) and pagerank (Page 1999) derived from raw frequencies of literature citations.

It may also be possible to request authors to explicitly distinguish why they chose to mention a digital repository -- whether they actually used the data or service to obtain their results, or they are merely related. Even without explicit citation mechanisms, it may be possible to make the distinction to some extent from the context where the mentions appear (e.g. in the methods section it may suggest that the data was actually used), and therefore distinguishing whether the data or service lead to direct impact (a mention definitely indicates influence of the resource in some way already). Similarly, it would be possible to distinguish whether the mention carries positive or negative sentiment of the resource. The key is that the standards bring unambiguous and persistent references to digital repositories.

Funding Statement

The research reported here is supported in part by Grant U24AI117966 National Institutes of Health Big Data to Knowledge (BD2K) Initiative, U24DK097771 National Institute of Diabetes and Digestive and Kidney Diseases, and U24DA039832 National Institute on Drug Abuse.

References

Altmetrics. What are altmetrics?https://www.altmetric.com/about-altmetrics/what-are-altmetrics/ (Accessed 27 December 2016).

Bandrowski, A., et al., *The Resource Identification Initiative: A cultural shift in publishing.* Journal of Comparative Neurology, 2016. **524**(1): p. 8-22.

Buneman P, Davidson S, Frew J. Why data citation is a computational problem. Communications of the ACM. 2016 Aug 24;59(9):50-7.

Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014 https://www.force11.org/group/joint-declaration-data-citation-principles-final (Accessed 14 November 2016).

Google Scholar listing of articles published with RRID for the Protein Data Bank. https://scholar.google.com/scholar?q=RRID%3ASCR_012820+OR+RRID%3Anif-0000-00135&hl=en&as_sdt=0%2C5&oq=RRID%3ASCR_012820+. (Accessed 17 October 2016).

Hirsch JE. An index to quantify an individual's scientific research output. Proceedings of the National academy of Sciences of the United States of America. 2005 Nov 15:16569-72.

Huang YH, Rose PW, Hsu CN. Citing a data repository: A case study of the protein data bank. PloS One. 2015 Aug 28;10(8):e0136631. http://dx.doi.org/10.1371/journal.pone.0136631

Huang YH, A study of data citation. PhD Dissertation. Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan. 2015. http://www.airitilibrary.com/Publication/alDetailedMesh1?DocID=U0001-2601201621083800 (Accessed 26 December 2016)

List of resources no longer in service according to the Neuroscience Information Framework. https://neuinfo.org/Resources/search?q=%2A&l=&facet[]=Availability:THIS%20RESOURCE%20IS%20NO%20LONGER%20IN%20SERVICE&sort=asc&column=resource_name&sort=asc (Accessed 17 Oct 2016).

Ozyurt IB, Grethe JS, Martone ME, Bandrowski AE. Resource disambiguator for the web: extracting biomedical resources and their citations from the scientific literature. PloS one. 2016 Jan 5;11(1):e0146300.

Page L, Brin S, Motwani R, Winograd T. The PageRank citation ranking: bringing order to the web. Technical Report. Stanford InfoLab. 1999. http://ilpubs.stanford.edu:8090/422/ (Accessed 14 November 2016).

Rose PW and Hsu CN. bioCADDIE pilot project 3.2 Development of Citation and Data Access Metrics applied to RCSB Protein Data Bank and related Resources. 2015. https://biocaddie.org/group/pilot-project/pilot-project-3-2-development-citation-and-data-access-metrics-applied-rcsb (Accessed 14 November 2016).

Starr J, Castro E, Crosas M, Dumontier M, Downs RR, Duerr R, Haak LL, Haendel M, Herman I, Hodson S, Hourclé J. Achieving human and machine accessibility of cited data in scholarly publications. PeerJ Computer Science. 2015 May 27;1:e1.

Tables.

Table 1. Different types of mentions of issued PDB IDs identified in PMC. The statistics of mentions may include false positives due to errors by the text mining software for the last two types.

Identifier	Example	Machine readable	Mentions(*)	%
PDB ID	PDB ID: **1STP**	yes	14,888	4.8
PDB DOI	http://dx.doi.org/10.2210/pdb**1stp**/pdb	yes	155	0.05
External link tag	<ext-link … ext-link-type=”pdb” xlink:href=”**1STP**”>	yes	32,108	10
PDB file name	**1stp**.pdb	yes	895	0.03
PDB URL	http://www.rcsb.org/… /structureId=**1stp**	yes, but URL may change	657	0.2
Non-standard PDB ID	PDB code: **1STP**, PDB reference **1STP**, PDB accession number **1STP**, Many variations...	yes/no	22,081	7.1
PDB in context	“We employed the following **PDB** coordinates: glycogen phosphorylase, **1gpy** …”	yes/no with text mining	16,726	5.4
Free text	“We first placed S2 bound to human PI3KC; (**3ene**) into the reference coordinates …”	yes/no with text mining	221,287	72

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2688v1 | CC BY 4.0 Open Access | rec: 3 Jan 2017, publ: 3 Jan 2017

Table 2: Top 10 highly cited protein structures (top) and top 10 highly mentioned protein structures in PDB. "Year" shows when the PDB ID was issued.

Citation Rank	PDB ID	Year	# of Citations	# of Mentions	Mention Rank
1	1AOI	1997	1527	31	37
2	1BL8	1998	1234	35	24
3	1F88	2000	957	44	16
4	1GC1	1998	852	26	57
5	1RV1	2004	747	11	488
6	1FFK	2000	746	31	34
7	**2RH1**	2007	682	124	1
8	1YSG	2005	650	6	1984
9	**2A79**	2005	635	49	10
10	1AIK	1997	561	12	403
Mention Rank	PDB ID	Year	# of Mentions	# of Citations	Citation Rank
1	**2RH1**	2007	124	682	7
2	1UBQ	1987	96	222	142
3	1KX5	2002	69	272	87
4	2R9R	2007	65	433	20
5	3EML	2008	65	408	24
6	1U19	2004	64	227	134
7	1K4C	2001	59	454	18
8	2VT4	2008	55	356	38
9	2B4C	2005	55	289	71
10	**2A79**	2005	49	635	9

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2688v1 | CC BY 4.0 Open Access | rec: 3 Jan 2017, publ: 3 Jan 2017