ISQ

TOPIC

ALTMETRICS

NISO
How the information world
CONNECTS

IP

William
Gunn

Social Signals Reflect Academic Impact:

What It Means When a Scholar Adds a Paper to Mendeley

WILLIAM GUNN

"The notion that the impact factor can encapsulate the value of everything a scholar produces is a bit simplistic." Todd Carpenter, Executive Director, National Information Standards Organization[1]

The academic social network Mendeley[2] has emerged as one of the most interesting sources of altmetrics. With a community of 2.4 million academics who have uploaded over 420 million documents across every discipline from life science to math to the arts and humanities, Mendeley is making it possible for academics, institutions, and funding organizations to really see the true picture of the impact of their research, not just on their field, but on all the stakeholders in research.

Altmetrics

Altmetrics, or "alternative metrics," are so called to distinguish them from bibliometrics, the traditional, decades-old system of counting citations and academic journal publications and also from webometrics, the measurement of webpage rank or influence by analyzing links between pages on the web.[3] There are a number of new kinds of data that are being collected about scholarly works, such as article pageviews, document saves or bookmarks, PDF downloads, tags, likes or shares on social networks, saves to reference managers, forks and patches of experimental code, and comments or posts on blogs, each reflecting a different dimension of influence.[4] These various metrics, collectively called altmetrics,[5] have been the subject of extensive study over the past few years[6] and show modest correlation to traditional citation-based metrics, but also reveal new types of impact: impact on the non-publishing consumers of research and also impact of non-journal forms of academic output such as code, datasets, or simply individual bits of data or figures too small for a traditional publication.

Social signals

Examining item usage to determine impact is a very old practice.[7] Libraries and publishers have been collecting and using usage-based metrics for a long time in the form of COUNTER reports,[8] ILL requests, and similar indicators, so altmetrics aren't novel in the application of usage metrics to the assessment of academic impact, but rather seek to add new types of usage, new objects of use, and to do this at web scale, rather than locally to one institution.[9] One of the more interesting forms of usage is what's reflected in scholars' use of social networks to discover and share academic material. This usage comes in many forms, some heavy and content-rich, such as blog posts or Wikipedia links, some plentiful yet content-poor. On the plentiful side, Twitter has emerged as an important source of scholarly signals.[10],[11] While this is convenient—because many scholars use Twitter, tweets are public, and they can easily be gathered and analyzed—the limited context available with a tweet provides an indication that the article cited may have been read, but little more. On the other hand, blog posts and Wikipedia references provide a very strong signal that a work is useful to scholars, but the relative amount of the literature which appears in a blog post is fairly small, limiting its systematic use. The happy middle ground is occupied by social bookmarking tools and academic reference managers. These tools have broad enough adoption by scholars to have reasonably good coverage of the literature, and the presence of a document in a reference manager is a much clearer signal that the article is influencing research. Mendeley is one of those tools and it provides plenty of context via metadata capture and user profiles, opening up the possibility of filtering the social signals according to the needs of the entity examining its impact. It is important to note that differences in how the various communities use the available tools modify how impact is reflected by the tool and, in addition, the newness of many of these tools biases them to more recent literature. This article will discuss Mendeley as a source of altmetrics and what types of impact are reflected in the data available from the platform.

What data does Mendeley collect?

Mendeley is a reference management tool for researchers to organize, share, and discover research. It has broad adoption across disciplines with the largest numbers of researchers currently in life sciences, chemistry, math, and computer science, but also with representation from the social sciences and non-journal based humanities disciplines as well. Accordingly, the research catalog has the best coverage in the sciences, often having greater than 90% of recent issues of many journals. The greater representation of the sciences

in Mendeley is thought to be primarily a reflection of its PDF-centric workflow and the journal article-centric communication in the sciences.

Researchers use Mendeley to store research papers and other publications along with the metadata about those publications, to share those papers or collections of papers with colleagues, and to discover new material based on what others are reading. The activity on Mendeley, therefore, provides many signals that reflect different types of impact, and there have been numerous studies comparing how many people have an item in their Mendeley library with citations, Impact Factor,[10] F1000,[13] article downloads, and social bookmarking.

Mendeley can return quite a lot of aggregated, anonymous, data about the usage of a publication found in its catalog. Figure 1 shows an example of the data returned from a document details call to Mendeley. Note that some documents which have only been uploaded by one researcher may not be available via the API due to the content quality filter that suppresses results for these documents. Using Scopus[14] data as a "ground truth" dataset to enrich the consensus metadata provided by researchers using Mendeley, we will be able to tune our content quality filters more finely and will be able to remove the requirement for a document to have been uploaded more than once in order to have a canonical representation, catalog page, and API availability.

Discussion of a few of the items returned by such a details call and what they, individually and in the aggregate, can tell us about scholarly activity is in order.

Keywords

Keywords are user-generated content that provides an indication of what the author thinks are significant concepts or relationships in the paper. Mendeley currently only returns the author-supplied keywords in response to a request for the public details for a paper. Any tags that an individual user has added can only be retrieved by permission of the user through a separate user-specific call for the document details.

Identifiers

Identifiers are the other names by which the document is known. These may be a PubMed ID (PMID), an arXiv ID, a DOI, an ISBN, or an ISSN. Included elsewhere in the document details data is a UUID (universally unique identifier) for the document, an article page URL, and the "page slug", which is the bit of the URL that uniquely identifies the catalog page for the document. These

Figure 1: Example of the data returned from a document details call to Mendeley

```
{"abstract":"Diabetic complication is comprised of [truncated]",
"keywords":[
        Interleukin 18",
        "diabetic nephropathy",
        "high sensitive CRP",
        "proinflammatory cytokine",
        "oxidative stress",
        "adipokine"],
"website":"http:\/\/www.ncbi.nlm.nih.gov\/pubmed\/20186552",
"identifiers":{
        "pmid":"20186552",
        "issn":"14325233",
        "doi":"10.1007\/s00592-010-0178-4"},
"stats":{
        "readers":7,
        "discipline":[
                {"id":3,"name":"Biological Sciences","value":71},
                {"id":19,"name":"Medicine","value":29}],
        "country":[
                {"name":"United States","value":29},
                {"name":"Brazil","value":29},
                {"name":"United Kingdom","value":14}],
        "status":[
                {"name":"Doctoral Student","value":29},
                {"name":"Student (Master)","value":14},
                {"name":"Post Doc","value":14}]},
"issue":"2",
"pages":"111-7",
"publication_outlet":"Acta Diabetologica",
"type":"Journal Article",
"url":"interleukin-18-contributes-more-closely-progression-diabetic-nephropathy-other-diabetic-
complication",
"uuid":"8af2c880-cd0d-11df-922b-0024e8453de6",
"authors":[
        {"forename":"Takayuki","surname":"Fujita"},
        {"forename":"Norikazu","surname":"Ogihara"},
        {"forename":"Yumi","surname":"Kamura"},
        {"forename":"Atsushi","surname":"Satomura"},
        {"forename":"Yoshinobu","surname":"Fuke"},
        {"forename":"Chie","surname":"Shimizu"},
        {"forename":"Yuki","surname":"Wada"},
        {"forename":"Koichi","surname":"Matsumoto"}
],
"title":"Interleukin-18 contributes more closely to the progression of diabetic nephropathy
than other diabetic complications.",
"volume":"49",
"year":2010,
"categories":[338,43],
"oa_journal":false,
"mendeley_url":"http:\/\/api.mendeley.com\/research\/interleukin-18-contributes-more-closely-
progression-diabetic-nephropathy-other-diabetic-complication\/"}
```

identifiers are useful for querying other databases about documents found at Mendeley to find out what data the other database may have, or as a shorthand way of making subsequent calls to the Mendeley API for a given document. Mendeley can also return a PMC (PubMed Central) ID (which is different from a PubMed ID) and an OAI (Open Archives Initiative) ID, if available (not shown above).

Stats

The Stats array contains several data structures which contain descriptive information about the document.

❶ Readers
This is the number of Mendeley users who have a given document in their library. This number includes all copies of
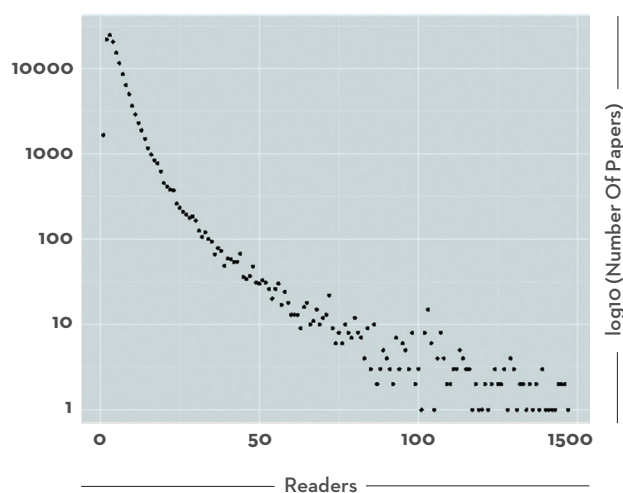
Figure 2: Distribution of document readership in Mendeley among a sample of 140K non-review papers published in 2012 and indexed by PubMed

a document, including citation-only entries, and is updated approximately daily. This value is perhaps one of the most interesting from an altmetrics point of view; more details about how this number is derived can be found in the section below on *Mendeley Readership*.

② Discipline

This is the breakdown of the disciplines of the readers, given as whole number percents of the total readership. The discipline name, ID, and percentage are given for the top three disciplines. This information can give a picture of the relative impact of a document on a specific field. For example, in the data in Figure 1, five of the readers come from Biological Sciences and two come from Medicine. Because the numbers add up to 100%, there are no other disciplines reading this document. At the moment, a reader may have only one discipline, which s/he selects at signup, and all reading of the user is attributed to that discipline. Mendeley plans to transition to a flexible tag-based system for discipline assignment in the future.

③ Country

This is a reporting of data about the geographic dispersal of readers, reported as percents. This data can be used to plot the impact of a work or set of works on a map at the country level. More granular readership information is coming, but due to privacy issues there are no current plans to report city-level data.

④ Status

This is similar to the reporting of data on the readership by academic discipline. Status is also selected by users at signup. One way to use this data is to determine if research is having more of an

impact on early-stage researchers relative to senior investigators, but there are classifications for non-research professions as well, which allows practitioner vs. researcher analyses.

Categories

Categories are given as numerical IDs and map onto the disciplines and sub-disciplines that Mendeley users assign themselves.

URL and UUID

These give the value of the unique identifier of the document in Mendeley, as well as the page slug for the article. So if you had a PMID and wanted to find the page on Mendeley for the article, you would first do a details call using the PMID, then append the value of the page slug to "http://www.mendeley.com/catalog/" to get the article page URL. The API also returns a slash-encoded version of the URL for the catalog page in the mendeley_url field, to allow developers to choose the mechanism for constructing links that works best for them.

Groups

If a document is present in a group on Mendeley, the information about what public groups it belongs to will also be returned. Only public groups will be shown in a request for document details using the public group method. If you want information about documents in private groups, you have to request permission via OAuth to access a user's private group information. Information about which groups

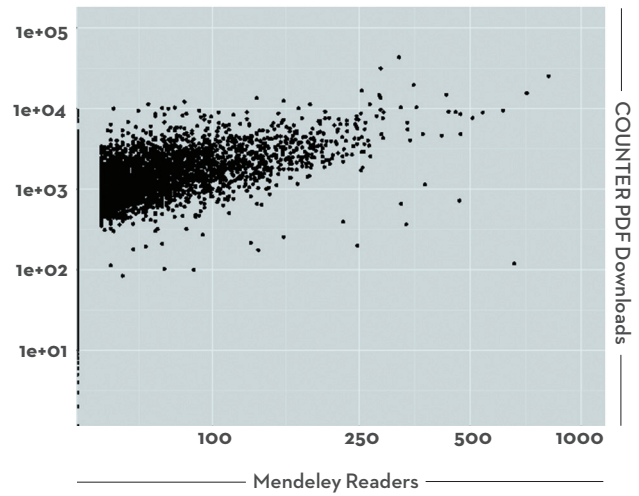Figure 3: F1000-reviewed papers often have higher readership

Figure 4: Correlation of Mendeley readership and PLOS-published COUNTER-compliant PDF download data

a document is in serves a similar function as do tags, so group memberships can be considered publicly available tags for a document. In addition, tags added to papers in public groups are available through a request for the details of documents in the group. Another way in which public groups can be used for altmetrics is by crawling publicly available groups and the membership of those groups to look at researcher-level altmetrics. For example, a researcher may be a member of a large number of groups, an administrator of a group with a large number of members, or listed as an author on a paper widely shared among a clinical practice or nursing group. This sharing among practitioner groups is another way to pick up impact of a paper on the non-citing readership.

Mendeley readership

The number of readers of a document on Mendeley is one of the potentially most interesting numbers from an altmetrics point of view. This number reflects the number of Mendeley users who have the document in their library. On a lower level, this number is the size of a document cluster. The Mendeley catalog is generated by a clustering algorithm, which runs approximately daily across the entirety of the Mendeley catalog (currently 420 million documents, increasing about a half a million a day), and clusters duplicates of the same document into one canonical representation. The size of this cluster is the readership of the document it contains. Occasionally, when the catalog is regenerated, multiple clusters will be generated for the same

document. This happens primarily with documents that have been uploaded hundreds of times in various forms and with various modifications made to the metadata by users. If there is duplication, the number of clusters is usually around three to five, with readers distributed randomly among them. This cluster instability is the reason that numbers for a given document sometimes seem to go down; the remedy is to track and combine the various duplicates of the document until they all collapse into one. Once Mendeley builds a "ground truth" set of metadata into the catalog via Scopus, documents will be assigned to a permanent cluster, anchored to the canonical metadata, where available. This will eliminate the issue of cluster instability.

Mendeley readership compared to other metrics

The distribution of readers of a document in Mendeley is distributed in a similar manner to citations (Figure 2). A small fraction of 2012 papers in PubMed have the majority of the citations, and so also with Mendeley readership (though not necessarily the same papers). There is a relationship between Mendeley readers and other altmetrics as well. Mendeley readership and F1000 scores are roughly correlated (Figure 3), as are Mendeley readers and COUNTER-compliant downloads of papers published by PLOS[15] (Figure 4).

There are a few things to keep in mind when considering the meaning of Mendeley readership or any other altmetric. The first thing to remember is that Twitter has only been around since 2006 and Mendeley since 2008, so given that
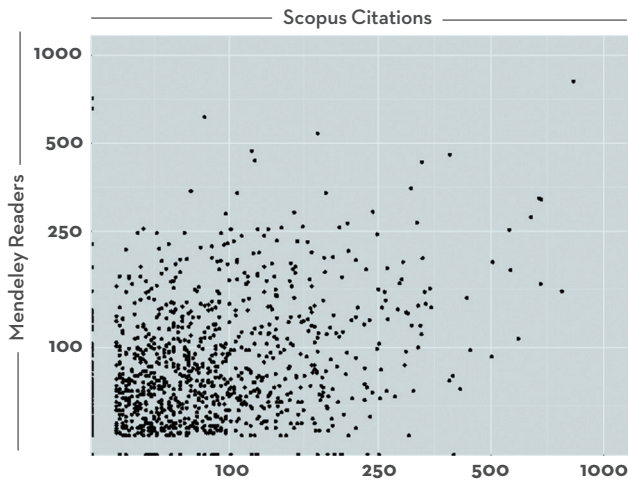
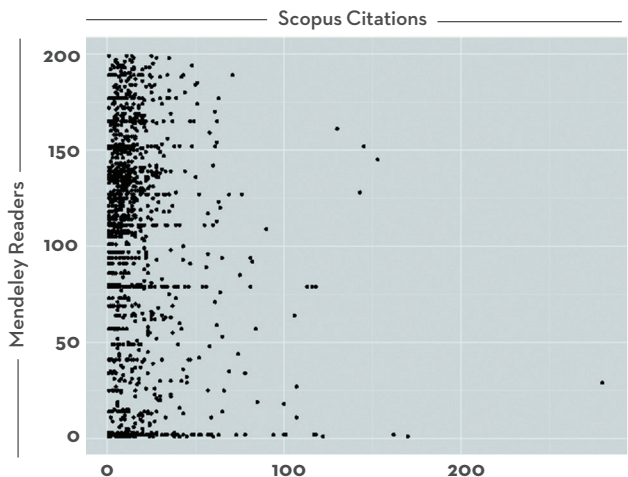Figure 5: Correlation between readers and citations within the same discipline

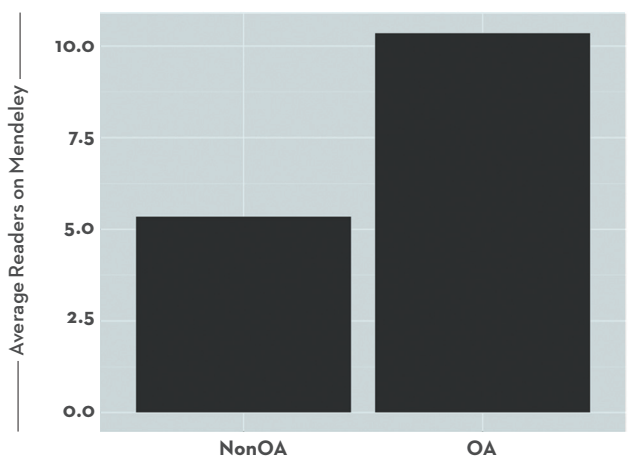Figure 6: Correlation between readers and citations across multiple disciplines

Figure 7: Mendeley readership of open access vs. non-OA articles

papers accrue most of their citations in the first two to five years after publication,[16],[17] it's reasonable to expect altmetrics to favor recent papers as well. Several ways to address this bias have been previously reported.[18],[19] It's also important to keep the different citation practices of various fields in mind when comparing quantitative metrics to citations. There's a within-field correlation between readers and citations of papers published by PLOS (Figure 5), but when looking at multidisciplinary non-open access publications such as *Cell, Nature,* and *Science,* the relationship appears much weaker (Figure 6). In addition, open access (OA) papers enjoy a significant readership advantage relative to non-OA papers (Figure 7).

Where do we go from here with altmetrics?

There's a growing interest in altmetrics from funders, institutions, researchers, and publishers. There are several commercial and non-profit companies that are operating in this space (ImpactStory,[20] Altmetric.com,[21] Plum Analytics,[22] PLOS,[23] and Mendeley's Institutional Edition[24]). In addition, many publishers such as Nature and Springer are beginning to report their own altmetrics. Clearly, now is the time to capitalize on the interest and attention to finally bring assessment of research out of the systems belonging to the print era and into a more modern, multifaceted system that takes advantage of the flexibility and scale of the web. Future extensions to altmetrics are expected to include more semantics about the inter-document links. For example, not just how many people cited a paper on Twitter, but who they were, or not just how many readers a paper has, but whether or not those reading a paper are highly read themselves. This discussion has focused on the journal article, but altmetrics providers such as ImpactStory are already tracking the impact of datasets and code along with more traditional academic outputs. The overall goal is to be able to relate this impact data to actual outcomes such as changed clinical practice, economic impacts, and policy implementations.

WILLIAM GUNN (william.gunn@mendeley.com) is Head of Academic Outreach at Mendeley.

Note: Data and code from this article are available upon request.

Future extensions to altmetrics are expected to include more semantics about the inter-document links. For example, not just how many people cited a paper on Twitter, but who they were, or not just how many readers a paper has, but whether or not those reading a paper are highly read themselves.

REFERENCES

1. Careless, James. "Altmetrics 101: A Primer." *Information Today*, February 2013, 30 (2). www.infotoday.com/it/feb13/Careless--Altmetrics-101-A-Primer.shtml

2. Mendeley [website]. www.mendeley.com

3. Wang, Xianwen, Zhi Wang, and Shenmeng Xu. "Tracing scientist's research trends realtimely." *Scientometrics*, 2013, 95 (2): 717-729. doi: dx.doi.org/10.1007/s11192-012-0884-5 Preprint available at: arxiv.org/abs/1208.1349

4. Priem, Jason, Heather A. Piwowar, and Bradley M. Hemminger. "Altmetrics in the wild: Using social media to explore scholarly impact." Presented at: *ACM Web Science Conference 2012 Workshop*, Evanston, IL, June 21, 2012. altmetrics.org/altmetrics12/priem/

5. Priem, Jason, Dario Taraborelli, Paul Groth, and Cameron Neylon. *Altmetrics: a manifesto*. Version 1.01. September 28, 2011. altmetrics.org/manifesto/

6. Priem, Jason, Paul Groth, and Dario Taraborelli,. "The altmetrics collection." *PloS ONE*, November 1, 2012, 7 (11): e48753. doi: dx.doi.org/10.1371/journal.pone.0048753

7. Bollen, Johan, Herbert Van de Sompel, Joan A. Smith, and Rick Luce. "Toward alternative metrics of journal impact: A comparison of download and citation data." *Information Processing & Management*, December 2005, 41 (6): 1419–1440. doi: dx.doi.org/10.1016/j.ipm.2005.03.024

8. COUNTER [website]. www.projectcounter.org

9. Piwowar, Heather. "Introduction: Altmetrics: What, why and where?" *Bulletin of the American Society for Information Science and Technology,* April/May 2013, 39 (4): 8–9. www.asis.org/Bulletin/Apr-13/AprMay13_Piwowar.html

10. Priem, Jason, and Kaitlin Light Costello. "How and why scholars cite on Twitter." *Proceedings of the 73rd American Society for Information Science and Technology Annual Meeting*, November/December 2010, 47 (1): 1–4. doi: dx.doi.org/10.1002/meet.14504701201 Preprint at: www.asis.org/asist2010/proceedings/proceedings/ASIST_AM10/submissions/201_Final_Submission.pdf

11. Shuai, Xin, Alberto Pepe, and Johan Bollen. "How the Scientific Community Reacts to Newly Submitted Preprints: Article Downloads, Twitter Mentions, and Citations." *PLoS ONE*, November 1, 2012, 7 (11): e47523. doi: dx.doi.org/10.1371/journal.pone.0047523

12. The Thomson Reuters Impact Factor [webpage]. wokinfo.com/essays/impact-factor/

13. Faculty of 1000 (F1000) [website]. f1000.com

14. Scopus [website]. www.info.sciverse.com/scopus

15. *Article-Level Metrics Information* [including discussion of COUNTER-compliant data]. PLOS One. www.plosone.org/static/almInfo

16. Tsay, Ming-Yueh. "Library journal use and citation half-life in medical science." *Journal of the American Society for Information Science*, December 1998, 49 (14): 1283–1292. doi: dx.doi.org/10.1002/(SICI)1097-4571(1998)49:14<1283::AID-ASI6>3.0.CO;2-I

17. Schloegl, Christian, and Juan Gorraiz. "Comparison of citation and usage indicators: the case of oncology journals." *Scientometrics*, March 2010, 82 (3): 567–580. doi: dx.doi.org/10.1007/s11192-010-0172-1

18. Thelwall, Mike, Stefanie Haustein, Vincent Larivière, and Cassidy R. Sugimoto. "Do Altmetrics Work? Twitter and Ten Other Social Web Services." *PLoS ONE*, May 28, 2013, 8 (5): e64841. doi: dx.doi.org/10.1371/journal.pone.0064841

19. Priem, Jason, Heather A. Piwowar, and Bradley M. Hemminger. "Altmetrics in the wild: An exploratory study of impact metrics based on social media." Presented at: *Metrics 2011: Symposium on Informetric and Scientometric Research*, New Orleans, LA, October 12, 2011. Slides: https://docs.google.com/present/edit?id=0ASyDkfrsAcUjZGRmZzc4N2NfMzU2Y2s1N3c4ZzY&hl=en_US

20. ImpactStory [website]. impactstory.org

21. Altmetric [website]. www.altmetric.com

22. Plum Analytics [website]. www.plumanalytics.com

23. PLOS [website]. www.plos.org/

24. Mendeley Institutional Edition [website]. www.mendeley.com/mendeley-institutional-edition/