Mendeley: Enabling and understanding scientific collaboration

William Gunn

Head of Academic Outreach, Mendeley, Menlo Park, CA, USA
Tel.: +1 646 755 9862; E-mail: william.gunn@mendeley.com

Abstract. This short paper provides an update on Mendeley's approach to scientific collaboration. Mendeley is Desktop software for organizing and collaborating. It provides a website for discovery of research and expertise, and it is a programmatic interface that provides data and functionality for many other services. Based on its collection of data on readings by its members it provides a way to understand scientific impact that compliments citation analysis and impact factors.

Keywords: Scientific collaboration, altmetrics, scientific workflow, crowdsourcing, online communities, impact factor

Mendeley was founded in November 2007 by three German PhD students and is based in London. In 2013 the small startup was bought by Elsevier. Today Mendeley is Desktop software for organizing and collaborating, a website for discovery of research and expertise, and a programmatic interface that provides data and functionality for many other services.

Mendeley's approach to facilitating collaboration has been to provide a tool that serves an individual need, and then extend that functionality to incorporate social features. Whereas some approaches have built separate sites where researchers are expected to go, we wanted to fit into the researcher's daily workflow by supporting activities they're already doing, and not by adding a new one. Collaboration isn't a separate activity from research; it's a part of many activities one engages in the course of research. To accomplish this, we've set up instrumentation at several points along the research process and are recording data about these activities. That approach is what gives us such rich data, and has allowed us to capture it at a level not previously possible.

A basic tool for Mendeley facilitates building and sharing of your research reference document collection which can be used for research, writing, or discussions with colleagues. Documents are added to your collection as PDFs, via the web importer, or via the research catalog. The distinguishing feature here is the connection to the scholarly cloud. For example, you have folders, but you can also have groups – "shared folders" – which work like the regular ones, except this folder is mirrored not only between your laptop and desktop, for example, but also between the client applications of everyone who is a member of that folder. Once documents are added, they can be searched in full text and annotated and tagged in various ways. When those annotations are done on a paper within a group, everyone gets the annotations as well.

For the researcher, Mendeley software takes some of the misery out of the writing process. Like other software, we integrate with your word processor, but we extend a social feature into that as well, where you can share a document with another Mendeley users, and they'll instantly have all the references you cited with no need to synchronize libraries. You also have the option of putting everything you cite into a group that your collaborators are in, as well to facilitate access to these citations by the group.

We have a catalog of research on the web. The catalog can be searched like other catalogs, but we leverage the power of our network here as well, because we feed the number of times a document has been added to our researchers' libraries, across the entire platform, into the search algorithm. The search doesn't just match strictly on keywords, but on the relevance of that keyword to the field as evidenced by article use. Individual document pages show related documents, which helps people find papers that might not have matched their exact search query but are still relevant to them. The catalog itself consists of the metadata of all the documents added by all of our researchers (where they haven't indicated that it's an unpublished work because some documents can be added as private). This has allowed us to build a catalog with 90% coverage of published academic literature in just a few years, and to keep it updated. (This is a rough estimate – the coverage is greater in some scientific disciplines and lower in the areas of the humanities, but due to the dynamic nature of our catalog, it's hard to come up with precise numbers.) Because we didn't license expensive databases to build this catalog, we can make it freely accessible via the web to both people and other applications. We can enrich our catalog with data no one else has because we know how many researchers have added a given document to their Mendeley library, what tags they've added, which groups they've added it to, and even what page they're on and what passages people have highlighted. This then enables us to actively recommend documents that a researcher doesn't have in their library, but that we can predict would be of interest to them. In a new feature in the developer's preview release, we extract figures and data tables from the documents to make the data easier to access and re-use.

From the building blocks noted above it is useful to categorize the two main types of data we collect. These data are them used to enable better understanding the researcher's thought processes and then their connections to the research community:

- Metadata about PDFs added – this goes to our research catalog, along with the discipline, educational status, and country of the researcher. We show popular tags applied to the paper (we exclude rare tags that only mean something to the researcher), and in the future we may add "frequently highlighted passages" from papers.
- Metadata about the researcher – name, discipline, academic status, and anything else they want to share, such as affiliation, country, papers they've authored, presentations they've given, special skills and expertise, etc.

So, in addition to facilitating collaboration, Mendeley is involved in understanding that collaboration and in understanding what impacts scientific advancement. In our system, each document addition is a "read" and is stamped with some metadata describing the context of the read event. You can think of a read as similar to a citation in that it's an indicator of the academic impact a paper is having. We have to define impact more broadly here, because many people use Mendeley – reading and sharing papers – who aren't authors. This includes groups such as clinicians, patients, small biotech firms, etc. Since we're capturing more dimensions of impact than a citation does, there's not going to be a perfect correlation between the readership of a paper and the citations of that paper, but there are indications of general trends. Figure 1 shows the correlation between Mendeley readers and citations for a sampling of papers from Pubmed. While there's a broad correlation between impact and citation, there are also some places where the correlation doesn't exist. Some examples of this are papers about the research practice itself – these are widely read but not widely cited. There are also papers which are widely cited by not widely read! Often for techniques or citations supporting common assertions in the introduction of a paper, the citation will simply be re-used from a previous paper rather than going back to the original citation itself. If everyone was using a reference manager (and it doesn't have to be Mendeley!) to generate

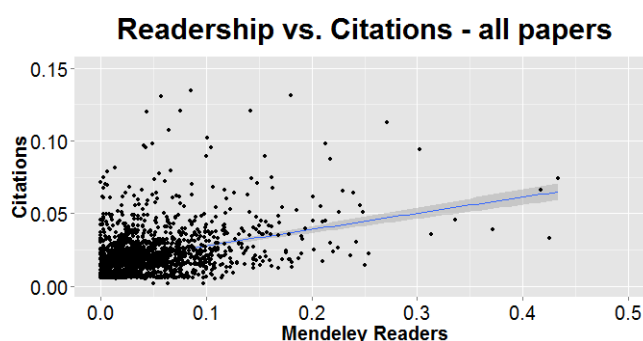**Readership vs. Citations - all papers**

Fig. 1. Among highly cited oncology papers in Pubmed, some are highly cited but not read (reviews, obligatory citations?), and some are highly read by not cited (citations-to-be?). Measurements were normalized so the highest measurement is 1, and axes were truncated for display purposes. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/ISU-140738.)

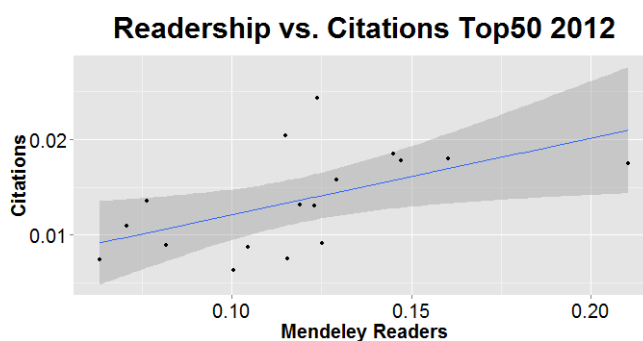**Readership vs. Citations Top50 2012**

Fig. 2. Among highly cited oncology papers from 2012, there exists a moderate correlation between readership and citations. Measurements were normalized so the highest measurement is 1, and axes were truncated for display purposes. Reviews were also excluded from this plot. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/ISU-140738.)

their citations, this would happen less. Figure 2 shows the correlation between Mendeley readers and citations for the 50 high-impact papers selected as part of the Reproducibility Initiative. This Initiative is a partnership between Mendeley and Science Exchange to replicate high impact research and develop best practices for the conduct of replication studies (http://reproducibilityinitiative.org). For high-impact papers which are reporting primary research results, the trend is pretty good.

So the primary way in which our data is related to citation is that it comes with a payload (who's doing the reading), it accrues faster, and it illuminates previously hidden impact. One can imagine this extending to capture things like the sentence preceding the citation (because we're integrated with the writing software), a user-supplied reason for the citation, or using the citation type ontology (CiTO) which allows construction of statements about citations which give character to the citation like extending "a cites b" to "a cites b in a specific manner or with a certain intent". It also allows facilitating awareness of the impact of non-paper research outputs such as datasets, code, performances, websites, or works other than narrative publications. Downstream citation analysis would then come with a wealth of information about where in the paper the citation referred, what kind of citation it is, and information about the makeup of the citing cohort.

This approach we just described comes under the evolving concept of altmetrics. We've been working on altmetrics before it was a term and have been working with the altmetrics community since the very

beginning. We're currently working with NISO to develop a draft standard for how altmetrics can be gathered, reported, and used in downstream applications.

So these are some of the concepts and approaches we have incorporated as we evolve Mendeley to support the changing nature of scientific collaboration. As we do so there are a number of issues that are on our plate to address. These include identify, privacy, attribution, gaming and filtration. We will continue to work with the scientific and knowledge communities as new concepts and tools are developed, a few of which we are hearing about at this workshop.