

# Consuming altmetrics: observations and lessons

Scott Chamberlain<sup>1,\*</sup>

<sup>1</sup>*Biology Dept., Simon Fraser University, Burnaby, BC, Canada V5B 1E1*

Keywords: altmetrics; R; software; API

---

\* E-mail: [scott@ropensci.org](mailto:scott@ropensci.org)

## INTRODUCTION

Altmetrics, or article level metrics, measures the impact of individual articles, or objects, usually at the object level (the word 'altmetrics' was coined by Jason Priem et al. [1]). This is in stark contrast to the Journal Impact Factor, which is a proprietary summation of the impact of all articles in a journal (owned and calculated by Thomson Reuters). Altmetrics have many advantages over journal level metrics, including the following three: a) Openness: Altmetrics are largely based off of data that is open to anyone (though there are some that aren't, e.g., Web of Science, Scopus). If data source are open conclusions based on altmetrics can be verified by others, and tools can be built on top of altmetrics; b) Speed: Altmetrics are nearly real-time metrics of scholarly impact [1] - citations can take years to accrue, but mentions and discussion that can be searched on the web take hours or days; c) Diversity of sources: Altmetrics include far more than just citations, and provide metrics in a variety of domains, including discussion by the media (mentions in the news), discussion by the public (facebook likes, tweets), and importance to colleagues (citations).

There are many potential uses for altmetrics, including:

- *Research.* As altmetrics rise in use and popularity, research on altmetrics themselves will inevitably become a more common use case. Some recent papers have answered the questions: How do different altmetrics relate to one another [2, 3]? What is the role of Twitter in the lifecycle of a paper [4]? Can tweets predict citations [5]? These questions involve collecting altmetrics in bulk from altmetrics providers, and manipulating, visualizing, and analyzing the data. This use case often requires using scripting languages (e.g., Python, Ruby, R) to consume altmetrics on a locally. Consuming altmetrics from this perspective is somewhat different than the use case in which a user views altmetrics hosted elsewhere in the cloud. The "local" use case is the target use case on which this paper is concerned.
- *Credit.* Scholars already put altmetrics on their CV's. With the rise of altmetrics, this will become much more common, especially with initiatives like that of NSF that allows scholars to get credit for *products*, not just papers - and products like datasets and videos can not be measured by citations or Journal Impact Factors. This use case will involve scholars with a wide variety of technical skills; and will likely be made easy with tools from ImpactStory or other providers. Piwowar and Priem discuss this use case further [6].

- *Filtering.* Scholars could not possibly find relevant papers efficiently given that there are now thousands of scholarly journals. Altmetrics can be used to filter. For example, Twitter can lead you to papers that are of great interest to people whose opinion you respect.

In this paper I discuss altmetrics from the perspective of developing and using scripting interfaces for altmetrics. From this perspective, there are a number of considerations: where can you get altmetrics data; data consistency; data provenance; altmetrics in context; and technical barriers to use.

## ALTMETRICS DATA PROVIDERS

There are a number of publishers now presenting altmetrics alongside peer-reviewed articles on their websites and on PDFs (for examples, see Wiley-Blackwell [7], Nature [8], Public Library of Science [9], Frontiers [10], Biomed Central [11]). Most of these publishers do not provide public facing APIs (Application Programming Interface - a way for computers to talk to one another) for altmetrics data, but instead use Altmetric or ImpactStory to provide altmetrics data on their papers - one exception is PLoS, which collects their own altmetrics, with an open API to use their altmetrics data. At the time of writing there are four major entities that aggregate and provide altmetrics data: PLoS, ImpactStory, and Altmetric, and Plum Analytics (see Table 1 for details). There are a few other smaller scale altmetrics providers, such as CitedIn ([12]) and ScienceCard ([13]), but are relatively small in scope and breadth. PLoS and ImpactStory have open APIs, while Altmetrics API limits API requests by hour and day, and by paid vs. non-paid accounts. PLoS provides data in JSON (JavaScript Object Notation) and XML (Extensible Metadata Language), ImpactStory in JSON only, and Altmetric in JSON and JSONP. PLoS provides much more granular data than the others, with daily, monthly and yearly totals; ImpactStory provides only total values; and Altmetric provides total values, plus incremental summaries of their proprietary Altmetric score. PLoS is a publisher, while the mission of the other two is to collect and provide altmetrics data. PLoS and ImpactStory are non-profit, while Altmetric is for-profit.

The four providers overlap in some sources of altmetrics they gather, but not all (see Appendix Table A1). The fact that there is some complementarity in sources opens the possibility that different metrics can be combined from across the different providers to get more a complete set of altmetrics. For those that are complementary, this should be relatively easy, and we don't have to worry about data consistency. However,

when they share data sources, data may not be consistent between providers for the same data source (see *Data standardization and consistency* below).

One of the important aspects of altmetrics is that most of the data collected by altmetrics aggregators like ImpactStory is that they aren't creating the data themselves, but rather are collecting the data from other sources that have their own licences. Thus, data licenses for PLoS, ImpactStory, and Altmetric are generally restricted to match those of the original data provider (e.g., some data providers do not let anyone to cache data).

Note that in discussing the four providers, we are only aware of the details of each provider that are open to the public. For example, Altmetric and PlumAnalytics provide some or all of their services to paying customers, which I don't cover here.

Variable	PLoS	ImpactStory	Altmetric	Plum Analytics
Open API?	Yes	Yes	Limited <sup>d</sup>	No
Data format	JSON,XML	JSON	JSON,JSONP	JSON
Granularity <sup>b</sup>	D,M,Y	T	I	T
API Authentication	None	API key	API key	API key (contact first)
Business type	Publisher	Altmetrics provider	Altmetrics provider	Altmetrics provider
Non-profit	Yes	Yes	No	No
Income based on charging	For publications	unknown	Publishers	Institutions
Rate limiting	Not enforced	Not enforced <sup>c</sup>	1 call/sec. <sup>d</sup>	Unknown
Products covered	Articles	Many <sup>e</sup>	Articles	Many <sup>f</sup>
Software clients	R <sup>g</sup>	R,Javascript <sup>h</sup>	R,Python <sup>i</sup>	Unknown

<sup>a</sup> Payed accounts with perks

<sup>b</sup> D: day; M: month; Y: year; T: total; I: incremental summaries

<sup>c</sup> Note: They recommend delaying a few seconds between requests

<sup>d</sup> Also hourly and daily limits enforced; using API key increases limits

<sup>e</sup> articles, code, software, presentations, datasets

<sup>f</sup> articles, code, software, presentations, datasets, books, theses, etc. (see <http://www.plumanalytics.com/metrics.html> for a full list)

<sup>g</sup> <https://github.com/ropensci/alm>

<sup>h</sup> R (<https://github.com/ropensci/rimpactstory>), Javascript (<https://github.com/highwire/opensource-js-ImpactStory>)

<sup>i</sup> R (<https://github.com/ropensci/rAltmetric>), Python (<https://github.com/lnielsen-cern/python-altmetric>)

TABLE I: Details on the four largest altmetrics providers.

## CONSISTENCY

Now that there are multiple providers for altmetrics data, data consistency is an important consideration. For example, PLoS, ImpactStory , Altmetric, and PlumAnalytics collect altmetrics from some of the same

data sources. Are the numbers they present to users consistent for the same paper, or are they different due to different collection dates, data sources, or methods of collection? Each of the providers that give aggregate altmetrics can collect and present altmetrics as needed for their target audience. However, as altmetrics consumers and researchers, we should have a clear understanding of the potential pitfalls when using altmetrics data for any purpose, especially research where data quality and consistency is essential.

I used a set of 565 articles (using their DOIs to search by) from PLoS journals only - this way all four providers would have data on the articles. I collected metrics from each of the four providers for each of the 565 DOIs. Note that I excluded data from PlumAnalytics for Citeulike as it was not provided (but they do collect it; pers. comm. Andrea Michalek). In addition, Facebook data was excluded from PlumAnalytics results because it was unclear how to equate their data with the data from the other providers. For each DOI I removed any missing values, and calculated the maximum difference between values (i.e., providers) and plotted the distribution of these maximum difference values for seven altmetrics that were shared among the providers (Fig. 1). Figure 1 shows that, at least with respect to absolute numbers, PMC citations are very different among providers, while PLoS views (html + pdf views; relevant only to PLoS ALM, ImpactStory, and PlumAnalytics) are somewhat less variable among providers. The remaining metrics were not very different among providers, with most values at zero, or no difference at all.

What are some possible reasons why similar metrics differ across providers? Numbers could differ for a number of reasons. First, data could be collected from different middle-men. For example, Twitter data is notorious for not being persistent. Thus, you either have to query the Twitter firehose constantly and store data, or go through a company like Topsy to collect tweets. Whereas ImpactStory collects tweets from Topsy, PLoS ALM and Altmetric collect tweets with an unknown method. Different intermediary data sources could lead to different data. Second, data could be collected at different dates. This could easily result in different data even if data are collected from the same source. This is especially obvious as ImpactStory collects some metrics via the PLoS ALM API, so those metrics that ImpactStory has from the PLoS ALM API should be the same as those that PLoS has. Fortunately, date is supplied in the returned data by three of the providers (PLoS ALM, ImpactStory, and Altmetric). Thus, I examined whether or not date could explain differences in metrics from the different providers. Figure 2 shows that there are definitely some large differences in values that could be due to differences in the date the data was collected, but this is not always the case (i.e., there

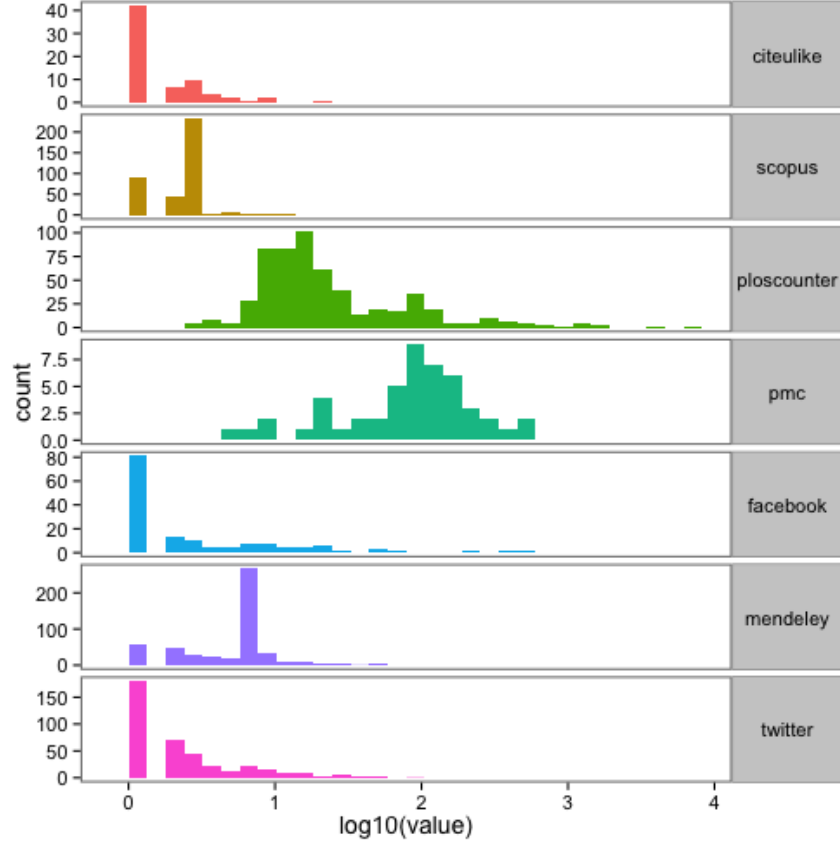


FIG. 1. Distribution of absolute differences in least and greatest value of each of seven different altmetrics on a set of 565 DOIs from Altmetric, ImpactStory, and PLoS ALM. Values were log10 transformed to improve visual comprehension.

are a lot of large difference values with very small difference in dates).

That was a rough overview of hundreds of DOIs. What do the differences among providers look like in more detail? I used a set of 20 DOIs from the 565 above to show the value of each altmetric from each of the altmetrics providers for each of the 20 DOIs (see Fig. 3). Note that in some cases there is very close overlap in values for the same altmetric on the same DOI across providers, but in some cases the values are very different.

An example giving what results look like may be instructive. Here is an example of calling the API of each the four providers to combine data from different sources, for the DOI *10.1371/journal.pbio.1001118* [?] (Table A2). There are many metrics that have exactly the same values among providers, though there are differences, which could be explained by the difference in the date data was collected. For example, PLoS ALM gives 3845 for combined PLoS views, while ImpactStory gives 3746 views. This is undoubtedly

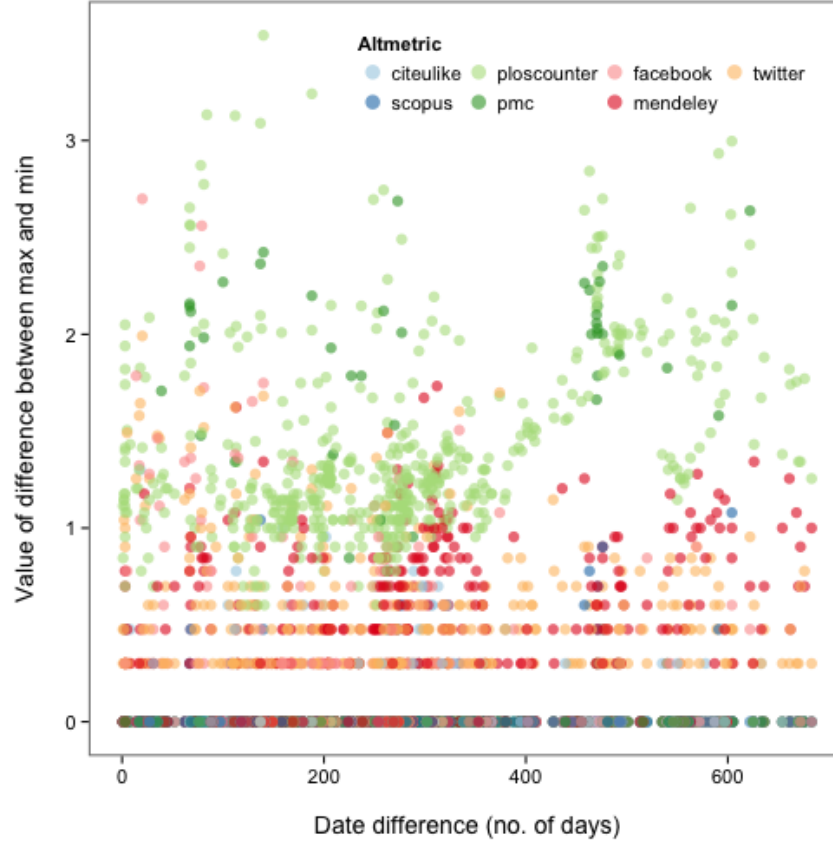


FIG. 2. Distribution of absolute differences in least and greatest value of each of seven different altmetrics on a set of 565 DOIs from Altmetric, ImpactStory, and PLoS ALM. Values were log10 transformed to improve visual comprehension.

explained by the fact that PLoS ALM data was last updated on May 24, 2013, while ImpactStory's data was last updated on May 18, 2013. There are some oddities, however. For example, Altmetric gives nine tweets, while ImpactStory only gives three tweets. ImpactStory's data was updated more recently (May 18, 2013) than that of Altmetric (July 28, 2012), which suggests something different about the way tweets among the two providers are collected as updated date alone can not explain the difference. In fact, Table A1 shows that ImpactStory collects tweets from Topsy (<http://topsy.com/>), while Altmetric collects them with an unknown method, which leads to different data.

The above findings on data consistency in fact suggest that altmetrics are inconsistent among providers of aggregate altmetrics. Casual users, and especially those conducting altmetrics research, should use caution when using altmetrics data.

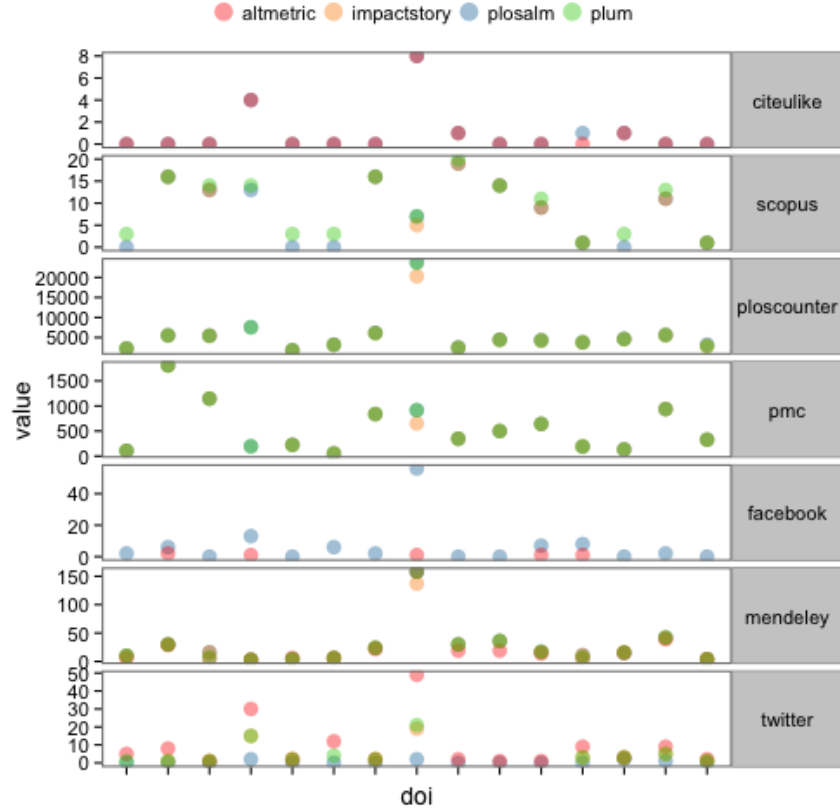


FIG. 3. A comparison of seven different altmetrics on a set of 20 DOIs from Altmetric, ImpactStory, and PLoS. This demonstrates how altmetrics can be very similar across providers for some DOIs, but very dissimilar for others.

provider	citeulike	scopus	ploscounter	pmc	facebook	mendeley	twitter	date_modified
PLoS ALM	1	1	3845	192	8	11	0	2013-05-24
Altmetric	0				1	5	9	2012-07-28
ImpactStory		1	3746	192		11	3	2013-05-18
PlumAnalytics		1	3746	192			3	

TABLE II: Example of combining results across three data providers on one DOI. Note that dates that data were last modified are the same for PLoS ALM and Altmetric, but different for ImpactStory. Missing values represent data that is not given by that provider.

### A crosswalk among providers

As discussed above, when similar data sources are collected by altmetrics providers, ideally, there would be a way to go between, for example, data from Twitter for PLoS, ImpactStory, Altmetric, and PlumAnalytics. Each of the four providers of course has the right to collect metrics as needed for their purposes, but as altmetrics consumers, we should be able to compare data from the same source across providers. In Appendix Table A1, I provide a table to crosswalk metrics for the same data source among providers.



## ALTMETRICS DATA PROVENANCE

Altmetrics data comes from somewhere - tweets from Twitter, citations from Web of Science or Scopus, bookmarks from Citeulike, etc. Provenance is concerned with the origin of an object, the ability to trace where an object comes from in case there is any need to check or validate data.

Why should we care about provenance in altmetrics? First, in general, altmetrics are based on completely digital data. This means that all use of, research on, hiring decisions based on, and conclusions drawn from, altmetrics data should theoretically be traceable back to the original production of that data. This is somewhat unusual - most research fields are based on data collected at some point that can not be traced - but this should be possible in altmetrics. Second, a specific example will demonstrate the power of data provenance in altmetrics. Imagine if a research paper makes controversial claims using altmetrics data on a set of objects (e.g., scholarly papers). An independent person could theoretically drill down into the data collected for that paper, and gain further insight, and potentially dispute or add to the latter mentioned paper.

Data for the same altmetrics resource could be calculated in different ways and collected at different times for the same object. The four providers already provide the date the metrics were updated. However, there is little information available, via their APIs at least, regarding how data were collected, and what, if any, calculations were done on the data before providing the data. The altmetrics community overall would benefit from transparency in how data are collected.

A good step in the right direction is that ImpactStory provides a field named *provenance\_url* with each metric data source. For example, for a recent paper [14], a GET call to the ImpactStory API returns many metrics, one of which is 10 bookmarks on Delicious. Importantly, they also return the field *provenance\_url*, in this case <http://www.delicious.com/url/9df9c6e819aa21a0e81ff8c6f4a52029>, which takes you directly to the human readable page on Delicious from where the data was collected. This is important for researchers as ideally all of our research is replicable. A nice bit about digital data such as altmetrics is that we can trace back final altmetrics from providers such as ImpactStory to their original source.

The PLoS ALM API provides something less obvious with respect to provenance, a field called *events\_url*, which for the same paper above [14] returns 82 bookmarks on Citeulike, and the human readable link to where the data was collected <http://www.citeulike.org/doi/10.1371/journal.pone.0000308>.

Plum Analytics does something interesting with respect to provenance - in addition to the canonical URL, they collect alias URLs for each object that they collect metrics on. For example, for the DOI *10.1371/journal.pone.0018657*, they collect many URLs that point to that paper. This makes sense as a digital product is inevitably going to end up living at more than one URL (the internet is a giant copying machine after all), so collecting URL aliases is a good step forward for altmetrics. It appears ImpactStory does this as well.

What is ideal with respect to data provenance? Is the link to where the original data was collected enough? Probably so, if no calculations were done on the original data before reaching users. However, some of the providers do give numbers which have been calculated. For example, ImpactStory puts some metrics into context by calculating a percentage relative to a reference set. Ideally, how this is done should be very clear, and verifiable.

On a related note to provenance, historical data is often helpful in a research context, for example, to know the change in a metric, rather than the point value at a certain date. PLoS provides historical altmetrics data on some of their metrics (except in case of licenses, e.g., Web of Science, Scopus), while Altmetric provides only historical data on their Altmetric score (see below section *Putting altmetrics in context*), and ImpactStory does not provide historical data. Of course, historical data, especially as more and more products are tracked, may become expensive to store, so perhaps won't be emphasized by altmetrics providers. In addition, a technical barrier comes in to play in that pushing a lot of data via an API call can get very time consuming.

## PUTTING ALTMETRICS IN CONTEXT

Raw altmetrics data can be number of tweets, or number of html views on a publishers website. What do these numbers mean? How does the paper or dataset I care about compare to others? ImpactStory gives context to their scores by classifying scores along two dimensions: audience (scholars or public) and type of engagement (view, discuss, save, cite, recommend). Users can then determine whether a product (paper, dataset, etc.) was highly viewed, discussed, saved, cited, or recommended, and by scientists, or by the public. This abstracts away many details; however, users can drill down to the underlying data via their API and web interface. Altmetric uses a different approach. They provide context for only one metric, the

altmetric score. This is a single aggregate metric, the calculation of which is not known. They do provide context for the altmetric score, including how it compares to a) all articles in the same journal, b) all articles in the same journal published within three weeks of the article, c) all articles in the Almetric database, and d) all articles in the Almetric database published within three weeks of the article. Altmetric gives detailed context for some altmetrics, including Facebook, Twitter, and blogs (see e.g., [http://altmetric.com/details.php?citation\\_id=1270911](http://altmetric.com/details.php?citation_id=1270911)). PlumAnalytics do not combine altmetrics into a single score as does Almetric, but do bin similar types of altmetrics into captures, citations, social media, mentions, and usage (<http://www.plumanalytics.com/metrics.html>; though you can dive into the individual altmetrics).

One of the advantages of altmetrics is the fact that they measure many different things, important to different stakeholders (public, scholars, funders) [? ? ]. Thus, combining altmetrics into a single score defeats one of the advantages of altmetrics over the traditional measure of impact, journal impact factor, a single metric summarising data on citations. The single Altmetric score is at first appealing given its apparent simplicity. However, if altmetrics are to avoid the mistakes befallen on the Journal Impact Factor [15], we should strive for meaningful altmetrics important to different stakeholders, that retain their context (Tweets vs. citations).

A specific example highlights the importance of context. A recent paper of much interest titled *Glass shape influences consumption rate for alcoholic beverages* [16] on the date of writing has an Altmetric score of 316. This score is compared relative to the same journal (PLoS One), and all journals at different points in time. Other altmetrics are reported but are not given any context. ImpactStory reports no single score, gives raw altmetrics data and gives context. For example, ImpactStory reports that there are 149 tweets that mentioned the paper, and this number of tweets puts the paper 97th-100th percentile of all Web of Science indexed articles that year (2012). This context for tweets about an article is more informative than knowing that the paper has an Altmetric score of 316 - people should know the context for what audience tweets represent, and number of tweets relative to a reference set gives a bit of information on the impact of the paper relative to others. Of course not all journals are indexed by Web of Science, and the important reference set for one person (e.g., papers in journals in their specific field) may be different from another (e.g., papers for colleagues at their university or department).

There is still work to do with respect to context. Future work should consider further dimensions of context.

For example, perhaps users should be able to decide how to put their metrics into context - instead of getting raw values and values relative to a pre-chosen reference set, users could choose what reference they want to use for their specific purpose. In addition, but much harder to achieve, is sentiment, or the meaning of the mention. That is, was a tweet or citation about a paper mentioned in a negative or positive light?

## TECHNICAL BARRIERS TO USE

Some altmetrics users may only require basic uses of altmetrics, like including altmetrics on their CV's [6] to show the various impacts of their research. Some may want to go deeper, and perhaps collect altmetrics at finer time scales, or with more detailed data, than are given by altmetrics aggregate providers. What are the barriers to getting more detailed altmetrics data?

Diving deeper into altmetrics means considering whether one can access data, whether the data source is machine readable, and how easy the data is to retrieve and manipulate once retrieved.

- *Data access* Many altmetrics sources are accessible as the data providers have open, or at least partly open, APIs (e.g., Crossref, PLoS). Other data sources can provide problems. For example, you can only get tweets from Twitter for the past 30 days, after which point you have to pay for a service that caches historical Twitter data (e.g., Topsy). Others are totally inaccessible (e.g., Google Scholar citations).
- *Machine readable* Ideally, altmetrics are provided through an API. However, some metrics of interest may only be in PDFs, spreadsheets, or html, which can not be easily consumed and mashed up. For these metrics, the user should seek out aggregators such as those discussed in this paper to do the heavy lifting. Alternatively, technically savy researchers could write their own code, or leverage tools such as ScraperWiki <https://scraperwiki.com/>.
- *Ease of use* Fortunately, many libraries, or extensions, exist for a number of programming languages relevant to scholars (Python, R), which deal with interacting with altmetrics data (e.g., Figshare API libraries [17], Twitter API libraries [18]). See Table 1 for links to libraries for aggregate providers. These libraries take care of the data collection and transform data to user friendly objects, allowing users to do the real work of analysis and inference.

## CONCLUSION

XXXXXX

## ACKNOWLEDGMENTS

I thank Martin Fenner for inviting me to write a paper in this special issue on altmetrics, and for helpful feedback from Carl Boettiger and Karthik Ram on earlier versions of this manuscript.

## REFERENCES

- 
- [1] J Priem, D. Taraborelli, P. Groth, and C. Neylon, “Altmetrics: A manifesto,” (2012).
  - [2] Koon-Kiu Yan and Mark Gerstein, “The spread of scientific information: insights from the web usage statistics in plos article-level metrics,” *PloS one* **6**, e19917 (2011).
  - [3] Johan Bollen, Herbert Van de Sompel, Aric Hagberg, and Ryan Chute, “A principal component analysis of 39 scientific impact measures,” *PloS one* **4**, e6022 (2009).
  - [4] Emily S Darling, David Shiffman, Isabelle M. Côté, and Joshua A Drew, “The role of twitter in the life cycle of a scientific publication,” *PeerJ PrePrints* **1**, e16v1 (2013).
  - [5] Gunther Eysenbach, “Can tweets predict citations? metrics of social impact based on twitter and correlation with traditional metrics of scientific impact,” *Journal of Medical Internet Research* **13** (2011), 10.2196/jmir.2012.
  - [6] Heather Piwowar and Jason Priem, “The power of altmetrics on a cv,” *Bulletin of the American Society for Information Science and Technology* **39**, 10–13 (2013).
  - [7] “Wiley altmetrics,” <http://onlinelibrary.wiley.com/doi/10.1111/2041-210X.12069/abstract>.
  - [8] “Nature magazine altmetrics,” <http://www.nature.com/nature/journal/v497/n7450/nature12188/metrics>.
  - [9] “Public library of science (plos) altmetrics,” <http://www.plosone.org/article/metrics/info:doi/10.1371/journal.pone.0064133>.
  - [10] “Frontiers publishing altmetrics,” [http://www.frontiersin.org/Journal/AbstractImpact.aspx?s=1187&name=Statistical\\_Genetics\\_and\\_Methodology&ART\\_DOI=10.3389/fgene.2013.00086&type=1](http://www.frontiersin.org/Journal/AbstractImpact.aspx?s=1187&name=Statistical_Genetics_and_Methodology&ART_DOI=10.3389/fgene.2013.00086&type=1).
  - [11] “Biomedcentral altmetrics,” [http://www.altmetric.com/details.php?citation\\_id=704519](http://www.altmetric.com/details.php?citation_id=704519).

- [12] “Citedin,” <http://citedin.org/>.
- [13] “Sciencecard,” <http://sciencecard.org/>.
- [14] Heather A Piwowar, Roger S Day, and Douglas B Fridsma, “Sharing detailed research data is associated with increased citation rate,” *PLoS One* **2**, e308 (2007).
- [15] AND The PLoS Medicine Editors, “The impact factor game,” *PLoS Med* **3**, e291 (2006).
- [16] Angela S. Attwood, Nicholas E. Scott-Samuel, George Stothart, Marcus R. Munafò, and Jerson Laks, “Glass shape influences consumption rate for alcoholic beverages,” *Plos One* **7** (2012), [10.1371/journal.pone.0043007](https://doi.org/10.1371/journal.pone.0043007).
- [17] “Figshare api libraries,”.
- [18] “Twitter api libraries,” <https://dev.twitter.com/docs/twitter-libraries>.

## APPENDIX A. CROSSWALK AMONG PROVIDERS.

The following Table A1 provides a crosswalk between altmetrics data collected by the three data providers. Note that these variables relate to one another across providers, but the data may be collected differently, and so for example, altmetrics collected for Twitter may differ between PLoS, ImpactStory and Altmetric. Where data sources are shared among at least two providers, I used only those fields that would give the same data if data were collected on the same date and all other things being equal. For example, PLoS ALM’s field *pubmed* is equivalent to ImpactStory’s *pubmed:pmc\_citations* field.

Footnotes from Table A1.

- a These are the exact names for each data source in the PLoS ALM API. For example: <http://alm.plos.org/api/v3/articles?ids=10.1371/journal.pone.0018657&source=twitter>.
- b You can not request a specific source from the ImpactStory API, so these are the names of the fields in the returned json. For example, see the json from this call: <http://api.impactstory.org/v1/item/doi/10.1371/journal.pone.0018657?key=YOURAPIKEY>.
- c You can not request a specific source from the Altmetric API, so these are the names of the fields in the returned json. For example, see the json from this call: <http://api.altmetric.com/v1/doi/10.1371/journal.pbio.0018657?key=YOURAPIKEY>.
- d Some of these names are the exact names returned in an API call; others are not.

Data source	PLoS <sup>a</sup>	ImpactStory <sup>b</sup>	Altmetric <sup>c</sup>	PlumAnalytics <sup>d</sup>
Biod	biod	No	No	
Connotea	connotea	No	No	No
Bloglines	bloglines	No	No	No
Nature blogs	nature	No	No	No
Postgenomic	postgenomic	No	No	No
Researchblogging	researchblogging	No	No	ResearchBlogging
WebOfScience citations	webofscience	No	No	No
Dryad	No	dryad:total_downloads package_views	No	Views, downloads
Figshare	No	figshare:views shares downloads	No	Recommendations, downloads, views
Github	No	github:forks stars	No	Collaborators, downloads, followers, forks, watches, gists
PLoS Search	No	plossearch:mentions	No	No
Slideshare	No	slideshare:favorites views comments downloads	No	Downloads, favorites, comments
Google+	No	No	cited by gplus count	No. +1's
MSM	No	No	cited by msm count	No
News articles	No	No	Yes	Yes
Reddit	No	No	cited by rdts count	Comments, upvotes-downvotes
Citeulike	citeulike	citeulike:bookmarks	No	Citeulike
Crossref	crossref	plosalm:crossref <sup>e</sup>	No	No
PLoS ALM	counter(pdf_views + html_views) <sup>f</sup>	plosalm(html_views, pdf_views)	No	Views of abstract, figures, full text, html, pdf, supporting data
PMC	pmc	plosalm:pmc_full-text + pmc-pdf <sup>g</sup>	No	No
PubMed	pubmed	pubmed:pmc_citations <sup>h</sup>	No	PubMed
Scienceseeker	scienceseeker	scienceseeker:blog_posts <sup>i</sup>	No	ScienceSeeker
Scopus citations	scopus	plosalm:scopus <sup>i</sup>	No	Scopus
Wikipedia	wikipedia	wikipedia:mentions	No	Wikipedia
Delicious	No	delicious:bookmarks	cited by delicious count	Delicious
Facebook	facebook_shares	facebook:shares <sup>j</sup>	cited by fbwalls count	Facebook clicks, comments, likes
Mendeley readers	mendeley shares	mendeley readers <sup>k</sup>	mendeley readers	Mendeley readers, groups
Twitter	twitter	topsy:tweets <sup>l</sup>	cited by tweeters count	Topsy tweets

TABLE III: Data sources used in taxize, tasks available, and links to them

<sup>e</sup> Collected from the PLoS ALM API.

<sup>f</sup> PLoS ALM also provides xml\_views.

- g Collected from the PLoS ALM API. Other PMC data fields collected from PLoS ALM (pmc\_abstract, pmc\_supp-data, pmc\_figure, pmc\_unique-ip) and from PubMed (suppdata\_views, figure\_views, unique\_ip\_views, pdf\_downloads, abstract\_views, fulltext\_views).
- h Should be equivalent to plosalm:pubmed\_central. ImpactStory also collects pubmed:pmc\_citations\_reviews f1000 pmc\_citations\_editorials.
- i Collected from the PLoS ALM API. Scopus citations also collected from Scopus itself, in the field scopus:citations.
- j ImpactStory also collects Facebook clicks, comments, and likes.
- k ImpactStory also collects Mendeley readers by discipline, number of groups that have added the article, percent of readers by country, and percent of readers by career\_stage.
- l ImpactStory also collects the number of influential\_tweets from Topsy.