# Consuming altmetrics: some observations and lessons

Scott Chamberlain[1,*]

[1] *Biology Dept., Simon Fraser University, Burnaby, BC, Canada V5B 1E1*

Keywords: altmetrics; R; sotware; API

* E-mail: scott@ropensci.org

## ABSTRACT

XXXX


## INTRODUCTION

Altmetrics, or article level metrics, measure the impact of individual articles, or objects, usually at the object level, or the author level. This is in stark contrast to the 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 quantifying more than just citations, and provide metrics on a variety of impacts (e.g., discussed by the media (mentions in the news), discussed by the public (facebook likes, tweets), and importance to colleagues (citations)).

Altmetrics can be consumed in a variety of contexts: as static text, images, or graphs alongside a pdf or website, as a javascript widget in a website and more. A use case that will, and should, be increasingly common is using scripting languages (e.g., Python, Ruby, R) to consume altmetrics on a computer locally (rather than in a browser), often for purposes of research on altmetrics themselves. Consuming altmetrics from this perspective is somewhat different than the typical use case where a user looks at altmetrics in a web browser.

This paper takes a look at the perspective of developing and using scripting interfaces to altmetrics. From this perspective, there are a number of considerations: data stanardization and consistency; API speed; authentication; and data provenance. First, I'll go over the current altmetrics data providers.

## ALTMETRICS DATA PROVIDERS

There are many publishers that are now presenting altmetrics alongside their papers on their websites. However, these publishers do not yet provide public facing APIs (Application Programming Interface) at the time of writing. There are four major entities that aggreagate and provide altmetrics data: PLoS, ImpactStory, and Altmetric, and Plum Analytics (see Table 1 for details). Plum Analytics does not have an open public facing interface or API, so will not be discussed further. There are a few other smaller scale altmetrics providers, such as CiteIn (http://citedin.org/). PLoS and ImpactStory have open APIs, while

Altmetrics API limits API requests by hour and day, and by payed vs. non-payed accounts. PLoS provides data in JSON (JavaScript Object Notation) and XML (Extensible Metadata Language), ImpactStory in JSON only, and Altmetric in JSON and JSONP. In terms of granularity, 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 three providers overlap in some sources of altmetrics they gather, but not all. The fact that there is some complementarity in sources opens the possibility that metrics can be combined from across the different providers. For those that are compelementary, this should be relatively easy. 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., Twitter).

Note that in discussing the three providers, we are only aware of the details of each provider that are open to the public. For example, Altmetric provides some services to paying customers, which we don't cover here.

TABLE I. Details on the current altmetrics providers

| Variable | PLoS | ImpactStory | Altmetric | Plum Analytics |
|---|---|---|---|---|
| Open API? | Yes | Yes | Limited[d] | No |
| Data format | JSON,XML | JSON | JSON,JSONP | Unknown |
| Granularity[b] | D,M,Y | T | I | Unknown |
| API Authentication | None | API key | API key | Unknown |
| Business type | Publisher | Altmetrics provider | Altmetrics provider | Altmetrics provider |
| Business model | Non-profit | Non-profit | For-profit | For-profit |
| Rate limiting | Not enforced | Not enforced[c] | 1 call/sec.[d] | Unknown |
| Products covered | Articles | Many[e] | Articles | Many[f] |

[a] Payed accounts with perks
[b] D: day; M: month; Y: year; T: total; I: incremental summaries
[c] Note: They recommend delaying a few seconds between requests
[d] Also hourly and daily limits enforced; using API key increases limits
[e] articles, code, software, presentations, datasets
[f] articles, code, software, presentations, datasets, books, theses, etc. (see http://www.plumanalytics.com/metrics.html for a full list)

## DATA STANDARDIZATION AND CONSISTENCY

Now that there are multiple providers for altmetrics data, data consistency is something to keep in mind. For example, PLoS, ImpactStory and Altmetric do collect altmetrics from some of the same data sources. Are the numbers they present to users the same for the same paper, or are they different due to different collection dates or methods of collection? Each of the three providers of course has the right to collect metrics as needed for their purposes,

I retrieved a set of 500 more or less random DOIs for full articles from PLoS journals - this way all three providers would have data on the papers. I collected metrics from each of the three providers for each of the 500 DOIs. TALK ABOUT RESULTS... A subset of 20 of the 500 DOIs are shown in Figure 1, showing the value of each altmetric from each of the altmetrics providers for each of the 20 DOIs. 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.

### 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, and Altmetric. Each of the three 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.

## AUTHENTICATION

There are a variety of possible authentication methods, some of which include: a) no authentication, b) username and password pair, c) API key, and d) OAuth (including OAuth1 and OAuth2) (Table 1). These different options make sense in different use cases. The first, no authentication, used by PLoS, makes sense when an API first comes out and testers are needed to get feedback on the API. A benefit of an API with no authentication is the barrier to entry is lower. That is, if you don't have to ask a user to register to get an API key they are more likely to use the API. The second and third options, username/password pair and

API key are relatively similar; API keys are used by both ImpactStory and Altmetric (Table 1). The last option, OAuth, is not used by any of the altmetrics providers. This authentication method is however used by many API providers. From the viewpoint of a consumer in a desktop scripting language, OAuth can be painful. What works better for scripting languages are the first three options.

## DATA PROVENANCE

Data for the same altmetrics resource could be calculated in different ways and collected at different times for the same object. The three 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 for-profit providers, Altmetric and Plum Analytics, especially have no obligation to share these, but 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 [1], 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 readble page on Delicous 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 [1] 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.

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 some percentage relative to a reference set. Ideally, how this is done should be very clear, and replicable.

## PUTTING ALTMETRICS IN CONTEXT, OR NORMALIZATION

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 has 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. No context is given though for individual altmetrics (e.g., tweets).

Future work should consider further dimensions of context. For example, XXXX

## RELAVANCE OF ALTMETRICS

The internet has facilitated the existence of altmetrics, as measures of impact are all around us, and many are now machine readable. However, which altmetrics are relavant? More importantly, which altmetrics are relavant to the community you care about? XXXX. [NOT SURE IT MAKES SENSE TO KEEP THIS SECTION???]

## CONCLUSION

XXXXXX

## ACKNOWLEDGMENTS

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

## I. REFERENCES

[1] 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).

## APPENDIX A. CROSSWALK TABLE 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.

Here is an example of calling the API of each the three providers to combine data from different sources.

TABLE II. Data sources used in taxize, tasks available, and links to them

| Data source | PLoS[a] | ImpactStory[b] | Altmetric[c] |
|---|---|---|---|
| Biod | biod | No | No |
| Connotea | connotea | No | No |
| General blogs | bloglines | No | No |
| Nature blogs | nature | No | No |
| Postgenomic | postgenomic | No | No |
| Researchblogging | researchblogging | No | No |
| WebOfScience citations | webofscience | No | No |
| Dryad | No | dryad:total_downloads package_views | No |
| Figshare | No | figshare:views shares downloads | No |
| Github | No | github:forks stars | No |
| PLoS Search | No | plossearch:mentions | No |
| Slideshare | No | slideshare:favorites views comments downloads | No |
| Google+ | No | No | cited_by_gplus_count |
| MSM | No | No | cited_by_msm_count |
| News articles | No | No | Yes |
| Reddit | No | No | cited_by_rdts_count |
| Citeulike | citeulike | citeulike:bookmarks | No |
| Crossref | crossref | plosalm:crossref[d] | No |
| PLoS ALM | counter(pdf_views html_views xml_views) | plosalm(html_views, pdf_views) | No |
| PMC | pmc | pmc(suppdata_views figure_views unique_ip_views pdf_downloads abstract_views fulltext_views); (plosalm:pmc_abstract pmc_supp-data pmc_figure pmc_full-text pmc_pdf pmc_unique-ip)[d] | No |
| PubMed | pubmed | pubmed:pmc_citations_reviews f1000 pmc_citations_editorials pmc_citations (plosalm:pubmed_central)[d] | No |
| Scienceseeker | scienceseeker | scienceseeker:blog_posts | No |
| Scopus citations | scopus | scopus:citations; (plosalm:scopus)[d] | No |
| Wikipedia | wikipedia | wikipedia:mentions | No |
| Delicious | No | delicious:bookmarks | cited_by_delicious_count |
| Facebook | facebook | facebook:shares clicks comments likes | cited_by_fbwalls_count |
| Mendeley | mendeley | mendeley:discipline readers groups country career_stage | mendeley |
| Twitter | twitter | topsy:influential_tweets tweets | cited_by_tweeters_count |

[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] Collected from the PLoS ALM API.