Altmetric: enriching scholarly content with article level discussion & metrics

Euan Adie and William Roe

euan@altmetric.com

Altmetric LLP

The Macmillan Building

4 Crinan St.

London, N1 9XW, UK

Scholarly content is increasingly being discussed, shared and bookmarked online by researchers. Altmetric is a start-up that focuses on tracking, collecting and measuring this activity on behalf of publishers and here we describe our approach and general philosophy. Over the past year we've seen sharing and discussion activity around approximately 750k articles. The average number of articles shared each day grows by 5 - 10% a month. We look at examples of how people are interacting with papers online and at how publishers can collect and present the resulting data to deliver real value to their authors and readers.

Introduction

Scholars are increasingly visible on the web and social media [1]. While the majority of their online activities may not be directly related to their research they are nevertheless discussing, sharing and bookmarking scholarly articles online in large numbers.

We know this because our job at Altmetric is to track the attention paid to papers online. Founded in January 2011 and with investment from Digital Science we're a London based start-up that identifies, tracks and collects article level metrics on behalf of publishers.

Article level metrics are quantitative or qualitative indicators of the impact that a single article has had. Examples of the former would be a count of the number of times the article has been downloaded, or shared on Twitter. Examples of the latter would be media coverage or a blog post from somebody well respected in the field.

Tracking the conversations around papers

Encouraging audiences to engage with articles online isn't anything new for many publishers. The Public Library of Science (PLoS), BioMed Central, Cell Press and Nature Publishing Group have all tried encouraging users to leave comments on papers with varying degrees of success but the response from users has generally been poor, with only a small fraction of papers ever receiving notable attention [2].

A larger proportion of papers are discussed in some depth on academic blogs and a larger still proportion shared on social networks like Twitter, Facebook and Google+. Scholars seem to feel more comfortable sharing or discussing content in more informal environments tied to their personal identity and where they know to some extent who their audience are… or indeed that they have an audience at all.

This presents publishers with a problem and an opportunity.

The problem is that discussion of the article is spread across many different blogs and social media sites making it hard for casual readers to discover and for authors and editors to respond to or otherwise engage with.

The opportunity is for publishers to deliver real value to their audiences by aggregating the relevant data, curating it and presenting it in context. Authors keen to track the responses to their work get a service they find useful and readers have one more reason to visit a publisher's site over repositories like PubMed Central.

What Altmetric does

Altmetric's mission is to make article level metrics easy. It collects download statistics, reference manager counts, links from mainstream media sources and social media shares & discussions, enriches the resulting data with demographics and profile information then makes everything available through APIs and for analysis through a web app called the Explorer.

Uses of the Altmetric data include adding links out to relevant discussions from article pages, powering analytics pages for authors and adding social data to search indexes to power “most shared” or “most popular” lists.

We deal primarily with scholarly articles, which we define very broadly as discrete pieces of content that have been assigned at least one recognized scholarly identifier, some examples being Digital Object Identifiers (DOIs), PubMed record IDs (PMIDs) and Handles. Preprints and working papers submitted to institutional repositories or subject specific databases usually have at least one recognized identifier assigned to them and are covered by our definition.

Since August 2011 Altmetric has seen approximately 2.8 million unique articles that have at least one indicator available. The majority of these were simply saved to an online reference manager but a significant fraction – 745k, around 25% of the total - were also shared or discussed online. Twitter and Facebook are currently the services on which we see the most activity (see Table 1).

Service	**Jun 14th to 21st**	**Sep 6th to 13th**	**Change**
Twitter	37,858	43,542	+13%
Facebook	1,216	3,181	+62%
Reddit	156	154	-1%
Pinterest	40	133	+69%

Blogs	471	430	-10%

Table 1: volumes of activity on selected online services in two one week periods - June 14th to June 21st 2012 and September 6th to September 13th - that involved at least one link to a scholarly article, as tracked by Altmetric.

The number of unique articles we see shared or discussed each day is growing by 5-10% a month.

New metrics using new data

The data Altmetric collects can also be used as the basis for new, quantitative indicators of impact. The field of alt-metrics – from which our name was derived - is “the creation and study of new metrics based on the Social Web for analyzing, and informing scholarship” [3].

The “alt“ in alt-metrics stands for “alternative”. Alt-metrics presents an alternative to the current practice of relying only on citation counts and journal impact factors for the quantitative analysis of impact by introducing new complementary approaches and sources of data.

Alt-metrics is data driven, enabled by access to a high volume of data that is quick to accrue - in contrast to citations, which are slow to collect and often few in number.

The promise of alt-metrics is two-fold:

- Stakeholders can use alt-metrics data to come up with new, relevant metrics better suited to situations not well served by citation counts;
- Alt-metrics data may also act as a leading indicator for existing, traditional metrics like citation counts [1, 4].

Specific examples of the first might include metrics for funders interested in the wider impact of the work they support or for publishers interested in measuring reach amongst audiences who don't traditionally cite, like patient advocates or many practitioners.

In general at Altmetric we're more interested in supplying the underlying data than creating new metrics of our own. The exception is the Altmetric score, which we created to serve as a general measure of the quality and quantity of attention that a scholarly article has received online.

The quantity part is straightforward: the more people discussing or sharing an article online the higher the score. The quality part is more complex and has a subjective component. A detailed explanation can be found online at altmetric.com.

Philosophy & approach

Early on in the company's development we made a few key decisions that helped shape the services that Altmetric now offers.

The first was that we'd always collect as much data as possible about as many scholarly articles as we could, including articles from publishers who weren't yet our customers.

This costs more up-front but gives us a number of advantages. It means we can put the numbers we supply in context (the amount of data we hold allows us to determine whether a particular level of attention is above or below what could reasonably be expected) and allows users to perform competitive analysis, benchmarking the levels of attention their content is getting against that of other journals or publishers.

The second was that we'd optimize for speed and scalability. Again this cost us more up front but means publishers have the flexibility to pull in Altmetric data however they want: in the client's browser, in bulk or even while pages are being rendered on the publisher's server, while ensuring that end-users see snappy response times.

Finally, we focus on data sources that can be manually audited by our users. If Altmetric says that an article has been tweeted about five times then users should be able to get the relevant five links, Twitter usernames and timestamps to go check for themselves.

Technical problems faced

Altmetric collects data by licensing data streams or connecting to the Application Programming Interfaces (APIs) – these are sets of programming instructions that allow you to automatically retrieve data from a server - of major social media sites and online reference managers and by crawling manually curated lists of blogs, newspapers and magazines.

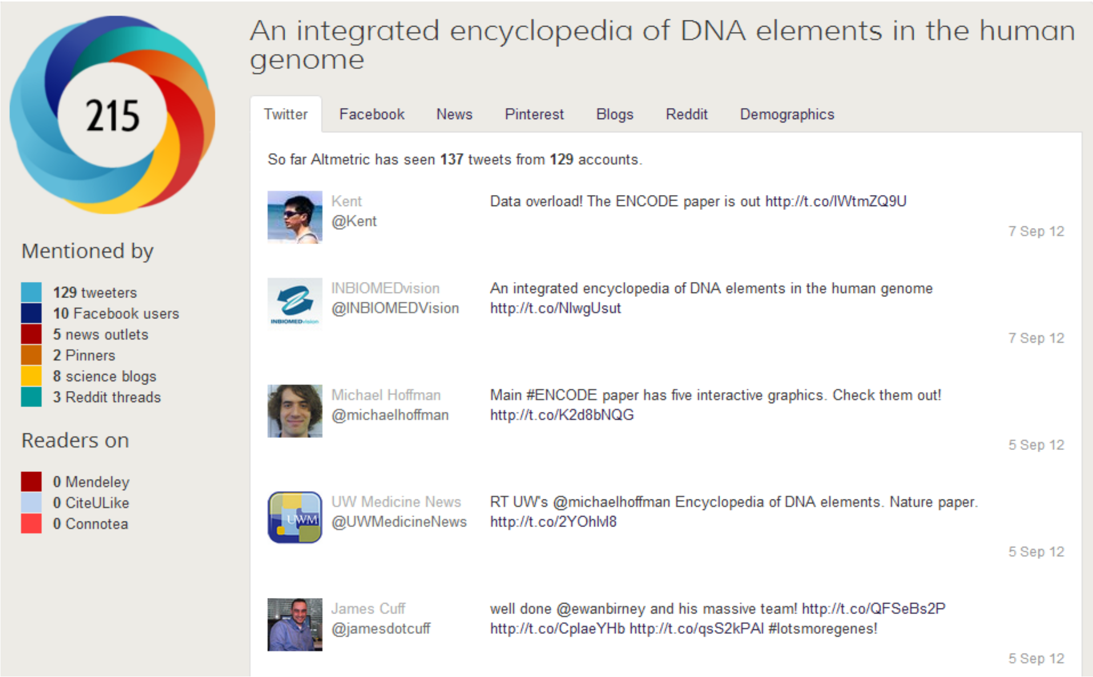
We call the individual tweets, blog posts, newspaper stories etc. that we collect *posts*. Each post is checked for links to scholarly articles, books or datasets. The posts that contain these links are called *mentions*.

Since Altmetric is about more than just simple counts we need to retain as much metadata about each mention as possible. Where we can we enrich that metadata with things like associated user profiles and social network information. The various sources have different data formats and schemas but they all have to come together in the Altmetric database and this constitutes a considerable data integration problem.

Furthermore the different data sources each have usage policies that we are required to adhere to. In order to ensure compliance with each site's terms and conditions, we need to trace and tag all data with its source, keeping track of which of our servers got data from where.

An important final stage of our data processing pipeline is article disambiguation. More often than not versions of the same article can be found in many different places as a preprint, on the publisher's site and in repositories like PubMed Central. One of Altmetric's key strengths is that it is able to pull these different versions together automatically wherever possible.

Example data



An integrated encyclopedia of DNA elements in the human genome

Mentioned by

🟦 **129** tweeters
🟦 **10** Facebook users
🟥 **5** news outlets
🟧 **2** Pinners
🟨 **8** science blogs
🟩 **3** Reddit threads

Readers on

🟥 **0** Mendeley
🟦 **0** CiteULike
🟥 **0** Connotea

So far Altmetric has seen **137** tweets from **129** accounts.

Kent @Kent	Data overload! The ENCODE paper is out http://t.co/lWtmZQ9U	7 Sep 12
INBIOMEDvision @INBIOMEDVision	An integrated encyclopedia of DNA elements in the human genome http://t.co/NlwgUsut	7 Sep 12
Michael Hoffman @michaelhoffman	Main #ENCODE paper has five interactive graphics. Check them out! http://t.co/K2d8bNQG	5 Sep 12
UW Medicine News @UWMedicineNews	RT UW's @michaelhoffman Encyclopedia of DNA elements. Nature paper. http://t.co/2YOhlvl8	5 Sep 12
James Cuff @jamesdotcuff	well done @ewanbirney and his massive team! http://t.co/QFSeBs2P http://t.co/CplaeYHb http://t.co/qsS2kPAI #lotsmoregenes!	5 Sep 12

Figure 1: screenshot of the Altmetric details page for the primary ENCODE paper taken the day it was published. The number in the circle in the top left is the Altmetric score.

The ENCODE consortium's primary paper, published in Nature in September 2012, is a good example of an article potentially enriched by collecting and presenting the discussions around it.

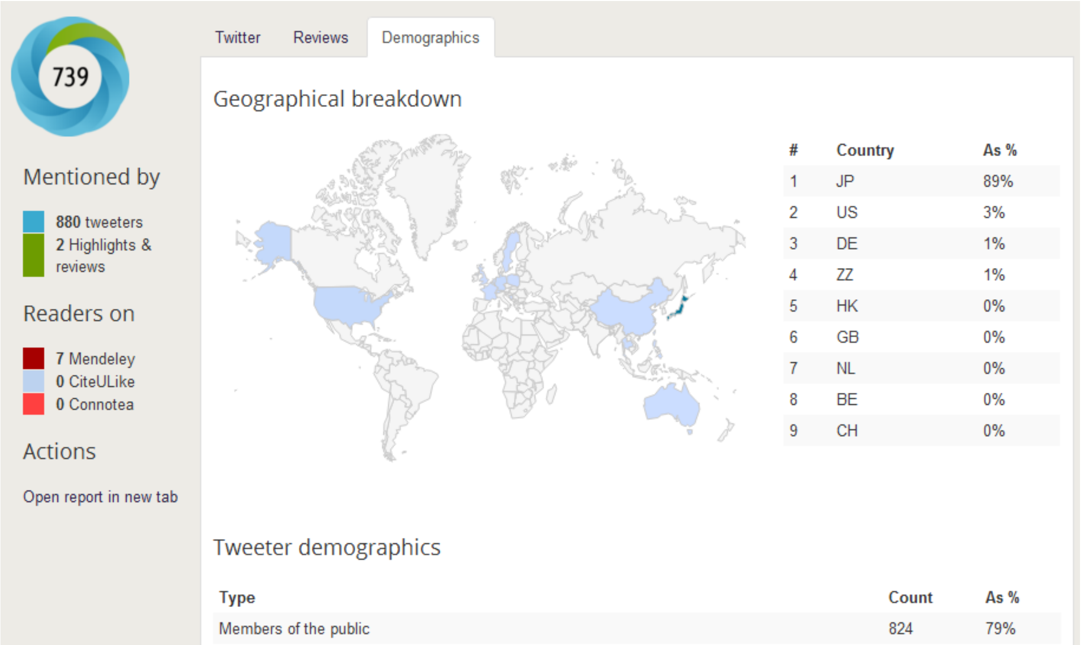
In the days after its publication Altmetric had picked up:

- A lengthy Q&A thread on Reddit with Michael Hoffman, a lead analyst on one section of the project
- Mainstream news coverage in The Guardian, Die Zeit, Science and Nature News that explained the research to the public
- Posts from a dozen science blogs. Early posts were generally enthusiastic; later ones concern themselves more with the hype the project got from the media

Altmetric allows customers like Nature Publishing Group to quickly find this kind of user generated content and promote or link to it as they see fit.

Another article that is interesting from an analytics perspective is Hess et al.'s "Gain of chromosome band 7q11 in papillary thyroid carcinomas of young patients is associated with exposure to low-dose irradiation", published in the Proceedings of the National Academy of Sciences (PNAS).

The Altmetric data reveals that it has been receiving an unusually high level of attention online since August 2011. A look at the associated demographics data explains why: 89% of the sharing activity is happening in Japan, where there are understandable concerns about the long term health effects of ionizing radiation after the reactor meltdowns at the Fukishima Daiichi nuclear power plant.



Mentioned by

🟦 880 tweeters
🟩 2 Highlights & reviews

Readers on

🟥 7 Mendeley
🟦 0 CiteULike
🟥 0 Connotea

Actions

Open report in new tab

Geographical breakdown

#	Country	As %
1	JP	89%
2	US	3%
3	DE	1%
4	ZZ	1%
5	HK	0%
6	GB	0%
7	NL	0%
8	BE	0%
9	CH	0%

Tweeter demographics

Type	Count	As %
Members of the public	824	79%

Figure 2: demographics for Hess et al.'s paper on thyroid cancers associated with low-dose irradiation. 89% of sharing activity happened in Japan and 79% of all tweets seen were by people who hadn't ever shared or discussed a scholarly article on Twitter before.

Working with Altmetric

We originally conceived of Altmetric as purely a 'data as a service' business, serving relevant data to product or platform development teams within publishers, and this is still our core focus.

We do it through a paid API. Given a DOI or other identifier the API returns all of the data Altmetric holds about that article in a 'ready to use' format: we've already fetched titles and snippets for mentions, enriched the data with demographics like those mentioned in the previous section and supplied some context so you can tell if the article is getting more or less attention than might be expected. Our customers then deal with storing and presenting the data to their readers and authors.

As described earlier our alt-metrics data is collected *a priori* rather than on demand then cached in memory by our servers, which means the API can be very fast - important as some customers want to call Altmetric each time a page is requested by a reader.

Getting up and running in minutes: embedded badges

While our early customers were keen on the API model it turned out that many publishers were more interested in turn-key solutions. Working with some early adopters we developed embeddable badges. Just as anybody can take a YouTube video and display it on their own site a publisher can take an Altmetric badge and add it to an article page with two lines of code: any readers clicking on the badge are taken to a page hosted by Altmetric but carrying the journal's branding where they can see details of the relevant blog posts, news stories, tweets and so on.

The embeddable badges are a quick, low friction way of adding article level metrics to a journal. Because of the way we license content it's also cheaper for our customers than consuming the data as a service.

The Altmetric Explorer

Early on in the company's development we created an internal tool called the Explorer to help us curate and manage the data. Built entirely on the Altmetric API it turned out to be a good way to demo Altmetric to potential customers, making numbers more concrete and allowing us to show editors the discussions we'd already picked up about recently published articles.

We spent some time developing it further and now offer subscription based access. The Explorer helps us reach other audiences - while most Explorer users are editors or press officers at STM publishers we also have funders and a marketing company as customers.

Altmetric’s business model

All three of the products we offer (API access, embeddable badges and Explorer access) are sold on an annual subscription basis.

Our data licensing and aggregation costs scale with article volume so we tier our pricing accordingly. This means we can charge smaller publishers less and get the data we collect out to as many people as possible.

Our early customers found predicting article volumes difficult so we use the number of active journals published as a proxy which makes Altmetric’s pricing simpler to calculate and to budget for.

In some circumstances we apply price modifiers: for example, we'll cap the cost for newer or more niche journals on which we don't hold much data so that customers only pay for the value they see delivered. Equally if a journal sees far more or less traffic than normal our infrastructure costs will change accordingly and this is reflected in the product's price.

Product	**Description**
API	Data as a service. Instant access to all of the data we hold on a given set of papers. Can optionally be redistributed, resold.
Embeds	Very simple to add to article pages. Data is co-branded, all hosted by Altmetric.com. Cannot be redistributed or resold but is correspondingly cheaper.
Explorer	For analytics. Easy to use web front-end. Data for internal business use only

What the future holds

As time goes on and more data is accumulated we expect to be able to properly assess how useful the data Altmetric collects is a leading indicator of later success.

The growth rates of existing social media show no sign of letting up and innovative new communications platforms like Branch and Medium can only be a good thing for scientists. As the volume of data increases so does the potential signal – the challenge for Altmetric will be to help people identify the kinds of impact that matter to them, to supply them with the data they need and to help them filter out the noise.

References

1. Bar-Illan, J. 2012. Beyond citations: Scholars' visibility on the social Web. *arXiv*, arXiv:1205.5611. Available at http://arxiv.org/abs/1205.5611

2. Adie, E. 2009. Commenting on scientific articles (PLoS edition). Nascent blog. Available at http://blogs.nature.com/nascent/2009/02/commenting_on_scientific_artic.html

3. Priem, J. 2010. The alt-metrics manifesto. Available at http://altmetrics.org/manifesto/

4. Eysenbach, E. 2011. Can Tweets Predict Citations? Metrics of Social Impact Based on Twitter and Correlation with Traditional Metrics of Scientific Impact. J Med Internet Res 2011;13(4):e123 http://dx.doi.org/10.2196/jmir.2012