



STI 2018 Leiden

*23rd International Conference on Science and Technology Indicators
"Science, Technology and Innovation Indicators in Transition"*

STI 2018 Conference Proceedings

Proceedings of the 23rd International Conference on Science and Technology Indicators

All papers published in this conference proceedings have been peer reviewed through a peer review process administered by the proceedings Editors. Reviews were conducted by expert referees to the professional and scientific standards expected of a conference proceedings.

Chair of the Conference

Paul Wouters

Scientific Editors

Rodrigo Costas
Thomas Franssen
Alfredo Yegros-Yegros

Layout

Andrea Reyes Elizondo
Suze van der Luijt-Jansen

The articles of this collection can be accessed at <https://hdl.handle.net/1887/64521>

ISBN: 978-90-9031204-0

© of the text: the authors

© 2018 Centre for Science and Technology Studies (CWTS), Leiden University, The Netherlands



This ARTICLE is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Factors contributing to altmetrics' complexity: a problem statement

Stacy Konkiel* and Euan Adie*

[*stacy@altmetric.com](mailto:stacy@altmetric.com); euan@altmetric.com

Altmetric, 1 Mark Square, London, EC2A 4EG (United Kingdom)

Introduction

In this paper, we lay out a problem statement for the challenges that altmetrics providers face in making the data they aggregate as transparent as the community desires. We use Altmetric as a case study for exploring altmetrics' complexity and the related issues that providers face in communicating that complexity to the diverse stakeholders they serve. We then pose questions for consideration by the larger scientometrics community, with an aim towards achieving a consensus solution (or set of solutions).

Problem statement: It is difficult to communicate altmetrics' data complexity to even the most savvy data consumers

We are concerned that the necessary complexity of altmetrics aggregators--dealing as we do with heterogeneous data from diverse sources, each with their own formats, privacy policies, and proprietary restrictions, along with aggregators' own technical constraints and curation policies--can create an unintentional "black boxing" of altmetrics data.

Reasons for altmetrics' complexity

Numerous studies have already identified ways in which complexity affects the understandability and perceived reliability of altmetrics (Liu, Jean & Adie, Euan, 2013; Mukherjee, Subotić, & Chaubey, 2018; Gumpenberger, Glänzel, & Gorraiz, 2016; Zahedi, 2017; Zahedi, Fenner, & Costas, 2015).

Altmetrics' complexity stems from several sources.

The nature of altmetrics data

First, there is the nature of the various kinds of data that we deal with. This data includes:

- Raw counts (counting how often research is shared in a source that we track, which are often processed in order to account for gaming or to limit the appearance of retweets or syndicated news content);
- Summary indicators & metrics (such as the total metrics for a group of publications that appear in search results, or broken down by institutions, departments, or researchers, as well as indicators like the Altmetric Attention Score and percentiles);
- Demographic information, which may include secondary analysis (for example, Twitter users' demographic classification, or geolocations for mainstream media sources, derived from raw data);
- Qualitative data, including the full-text or snippets of mentions; and
- Publications' metadata (which is often compiled into a single record, even for research that may have several versions).

Each aggregator collects and manages these diverse data in different ways, making “apples-to-apples” comparisons difficult, especially in the event of hard to find documentation.

Data sources' restrictions

Complexity is also often driven by source data providers' restrictions. For example, altmetrics' ever-changing nature (Gumpenberger, Glänzel, & Gorraiz, 2016) is manifested in data deletion by Twitter users; Twitter's Terms of Service requires aggregators like Altmetric to remove tracked tweets from our archives. As altmetrics providers who often license data from third parties, we often have to make changes to our data based on providers' terms and conditions. In the future, privacy laws like the European Union's General Data Privacy Regulations will complicate things further. These restrictions are external factors, and we must do our best to adapt and respect them as quickly as possible upon implementation of a change.

Necessary responses to gaming

Altmetrics' attempted gaming by authors and journals can also encourage opacity, specifically of altmetrics vendors' data collection and processing techniques. We are constantly adapting to gaming practices (Gordon, Lin, Cave, & Dandrea, 2015). Keeping our anti-gaming mechanisms secret is one way of preventing their circumvention.

Technical constraints

Technical constraints can also contribute to unintentional “black boxing” for altmetrics. In an ideal world, altmetrics providers' documentation would always be up-to-date, our source code would be perfectly commented, and our support staff would be kept immediately in the loop on changes to how we collect and manage altmetrics data. We would also never have to deal with changes to data providers' platforms or terms of service, the obsolescence of hardware or software, or staffing shortages when hiring talented developers. But these technical constraints do exist and can affect our ability to keep our data as transparent as we desire, despite our best efforts to mitigate their effects.

We respect arguments in favor of making altmetrics software open source so that it may be examined by the community and better understood. On the other hand, it is important to also acknowledge the hidden challenges of open source development. Open source done well requires full and proper documentation and technical support at the point of origin, which is hard to guarantee for the reasons described above. Moreover, the Altmetric technical infrastructure is specialized, having dependencies (e.g. code libraries) that would render it unusable by others without a lot of work.

Curation-related decisions

Finally, there is the question of altmetrics providers' philosophy in how each chooses to track altmetrics, and the related nuance of how those decisions are applied to each data source. For example, Altmetric decided long ago that we would aim to only track and report mentions that are fully auditable--for example, where one can verify for themselves that a Facebook share happened by looking at the profile page where a paper has been posted. But there are data sources that for reasons of user privacy or intellectual property are not auditable (e.g. Mendeley and the Open Syllabus Project), which are nonetheless important to track and from trustworthy sources. It is difficult--though admittedly not impossible--to make these decisions completely transparent.

Examples of data complexity from the trenches

Here, we share some specific examples of data complexity from Altmetric's own experiences, to better illustrate the processes by which altmetrics aggregators attempt to address them.

Changes to the Altmetric Attention Score over time

The Altmetric Attention Score (AAS) was created as a basic indicator of online attention for research. It is calculated using a weighted approximation of the potential reach or dissemination of a piece of research, based on the sources, authors, and volume of mentionsⁱ that research receives.

The AAS's calculation has changed several times since it was introduced, often in response to attempts to game the score. For example, we recently encountered a journal article that briefly had the highest Score ever recorded, achieved seemingly overnight. Upon further examination of the article's attention data, we realized that while we can usually catch and control for automated gaming attempts (e.g. Twitter bots), we were not able to identify human-initiated tweets solicited as part of an organized campaign (in this case, a religious leader had published a paper and asked his followers to share it online, which they did en masse). In response to this and similar attempts at gaming the AAS, we decided to cap the number of social media posts that would count towards the Altmetric Attention Score for papers we suspected of gaming. But until now, we have not discussed this tactic publicly, to prevent circumvention of our anti-gaming measures.

The introduction of new sources to Altmetric also means changes to the AAS for research in our database. When new sources are added, we usually begin factoring in those sources to the AAS calculation for research we track, recalculating existing Scores for items in our database where necessary. This inevitably changes the Attention Scores for hundreds if not thousands of outputs in our database.

Data privacy policies and laws

Currently, most altmetrics aggregators get a portion of our data from a data broker, Gnip, that enforces data sources' user privacy policies in a fairly straightforward way. For example, Twitter requires that once a tweet is deleted or a user's profile is set to "private", that third parties that have indexed that tweet must also delete the tweet from their database and remove it - and any public record that it existed - from public view. For Altmetric, that means that if a tweet about a paper is deleted, Gnip includes it in a stream of deleted tweet IDs that we must check our database for. We then use that list to update our live database.

But Gnip's streamlined approach to notifying aggregators of data deletions exists alongside other, messier means of requesting data removal. For example, we have honored one-off requests from those who email us to ask that we remove tweets, blog posts, and peer reviews that they have authored. If those requests were to increase, they would quickly become difficult to manage. When GDPR rules take effect in May 2018, we expect that we may in fact face this scenario. If this happens, GDPR will be complex to deal with transparently and in the spirit of the law.

Both types of data deletion regulations are made more complicated by the fact that our data is shared with researchers who have the ability to download and store user data indefinitely, even on Open Access repositories like Figshare. Though some interpretations of GDPR indicate that research may be a protected categoryⁱⁱ, researchers and altmetrics aggregators

alike may in reality occupy a legal gray area when it comes to how end-users' data is stored and shared for academic studies. The Cambridge academic who recently sold Facebook profile data collected originally for research purposesⁱⁱⁱ hasn't done anybody any favors in this context: platforms are likely to take a harder line in the future.

Human-curated data

Though a vast majority of the data processing at Altmetric is automated, individuals do guide how that data is collected (and whether it is collected at all) by way of various curation and software development activities. For example, Altmetric staff make decisions like:

- Should this blog or news source be indexed?
- Should this news source be labeled “Tier 1” (meaning it is of high importance or prestige and thus worthy of being featured in the Altmetric Explorer “Highlights” dashboard^{iv})?
- What keywords should we look for in Twitter user profiles in order to classify users into cohorts like “scientist” or “practitioner”?
- Does an expected spike in attention for a paper constitute spam or is it simply organic, viral growth?

These kinds of decisions are made on a daily basis and are subject to change over time. It can be difficult to keep them well-documented for the public.

How Altmetric attempts to address data complexity (the story so far)

Over the years, Altmetric has attempted to address the above issues in data complexity in a number of ways:

1. We keep our Knowledge Base and Support portal up-to-date and as well-documented as possible.
2. We recently hired an Engagement Manager, who trains customers using our data and communicates the complexities of our products.
3. We create collateral and educational resources to communicate best practices and data caveats to those beyond our customer base.
4. We regularly work with scientometrics researchers to communicate our data's complexity, answer questions, and improve our products based on the feedback we receive. As part of this effort, we recently launched a formal Altmetric Researcher Data Access Program, to better connect with researchers and offer trainings for using our data.

While we are proud of the work we have done to date to demystify our data, we are also cognizant of the fact that there is likely more that we and other altmetrics aggregators could do, too.

Questions for consideration

With the above issues surrounding altmetrics data complexity in mind, we now offer to the STI community two questions for consideration and discussion at STI 2018.

1. *How can altmetrics providers realistically serve diverse stakeholders at the same time, with respect to communicating altmetrics' complexity?*

Altmetrics aggregators serve many stakeholder groups simultaneously: end users who prefer simple counts over nuanced numbers and interpretations; researchers who want full complexity and transparency around attention data for their research, for their own edification; specialist communities like scientometrics researchers and other app providers who repurpose our data; and all those in-between.

How can we improve our data and product documentation and training practices to meet the needs of these diverse user communities in a scalable and manageable way? In our view, this is one of the best ways to combat the unintentional “black boxing” of altmetrics data.

2. *How can altmetrics providers inoculate themselves and the community against data users who do not do due diligence?*

We encounter a handful of instances each year of those who misinterpret or misuse our data, despite our best efforts to educate the community. How can Altmetric (and other altmetrics aggregators who regularly encounter similar issues) make our documentation more apparent and available, to prevent this from happening in the future? Is it perhaps more of an issue of visibility within the community (e.g. participating at conferences)? Or is there another piece of the puzzle that we are missing? We think this is a crucial second way of addressing the misinformation about and misuse of altmetrics.

Discussion

In response to our questions, there will inevitably calls for altmetrics aggregators to adhere to the NISO Altmetrics Data Quality Code of Conduct (NISO, 2016), which suggests that by simply introducing greater transparency, accounting for replicability, and ensuring data accuracy, that high-quality data that meets the community’s needs will be produced. But a demand for these changes does not account for the challenges of data complexity, necessary obfuscation in order to prevent gaming, and the divergent interests and metrics-related literacy of the stakeholders that aggregators serve. We wish to invite the community to think through these challenges with us, building upon existing standards and best practices in order to find solutions to the necessary complexity of altmetrics.

References

- Gordon, G., Lin, J., Cave, R., & Dandrea, R. (2015). The Question of Data Integrity in Article-Level Metrics. *PLoS Biology*, 13(8), e1002161. <https://doi.org/10.1371/journal.pbio.1002161>
- Gumpenberger, C., Glänzel, W., & Gorraiz, J. (2016). The ecstasy and the agony of the altmetric score. *Scientometrics*, 108(2), 977–982. <https://doi.org/10.1007/s11192-016-1991-5>
- Liu, Jean, & Adie, Euan. (2013). Five challenges in altmetrics: A toolmaker’s perspective. *Bulletin of the American Society for Information Science and Technology*, 39(4), 31–34. <https://doi.org/10.1002/bult.2013.1720390410>
- Mukherjee, B., Subotić, S., & Chaubey, A. K. (2018). And now for something completely different: the congruence of the Altmetric Attention Score’s structure between different article groups. *Scientometrics*, 114(1), 253–275. <https://doi.org/10.1007/s11192-017-2559-8>

NISO. (2016). *Outputs of the NISO Alternative Assessment Metrics Project*. Bethesda, MD, USA: National Information Standards Organization (NISO). Retrieved from <https://www.niso.org/publications/rp-25-2016-altmetrics>

Zahedi, Z. (2017). What explains the imbalance use of social media across different countries? A cross country analysis of presence of Twitter users tweeting scholarly publications. In *4th Altmetrics Conference*. Toronto, Canada: figshare. <https://doi.org/10.6084/m9.figshare.5454475.v2>

Zahedi, Z., Fenner, M., & Costas, R. (2015, March 12). How consistent are altmetrics providers? Study of 1000 PLOS ONE publications using the PLOS ALM, Mendeley and Altmetric.com APIs. <https://doi.org/10.6084/m9.figshare.1041821.v2>

ⁱ <https://help.altmetric.com/support/solutions/articles/6000060969-how-is-the-altmetric-attention-score-calculated->

ⁱⁱ <https://iapp.org/news/a/how-gdpr-changes-the-rules-for-research/>

ⁱⁱⁱ <http://money.cnn.com/2018/03/20/technology/aleksandr-kogan-interview/index.html>

^{iv} <https://help.altmetric.com/support/solutions/articles/6000146655-analyse-your-results>