Submitted for presentation at the altmetrics15: 5 years in, what do we know? The 2015 Altmetrics Workshop, Amsterdam Science Park, 9 October 2015

Consistency among altmetrics data provider/aggregators: what are the challenges?

Zohreh Zahedi[1], Martin Fenner[2] & Rodrigo Costas[3]
[1&3] Centre for Science & Technology Studies (CWTS), Leiden University, Leiden, The Netherlands
[2]DataCite, Hannover, Germany

Introduction, Data and methodology

This research deals with investigating consistency of data across three altmetrics providers or aggregators: Altmetric.com, Mendeley and the Open Source software Lagotto (used by PLOS, CrossRef and others). The aim of this study is to explore if metrics for a same set of publications are consistent across them and if not, what are possible reasons that explain these differences. By consistency we mean having (reasonably) the same score for the same DOI per source across different altmetrics providers/aggregators. For a proper development of the altmetric research and practice, it is critical to understand any potential similarity or difference in metrics across different altmetric aggregators.
For this purpose, a random sample of 30,000 Crossref (15,000) and WoS (15,000) DOIs from 2013 has been considered. The data collection has been done at the same date/time on July 23 2015 starting at 2 PM CEST using the Mendeley REST API, Altmetric.com dump file and the Lagotto open source application. Similar sources and metrics across these 3 providers have been analyzed and compared (*Facebook, Twitter, Mendeley, CiteULike and Reddit*).

Results:
Coverage and intensity of DOIs across altmetrics providers/aggregators
Table 1 shows the overall statistics for all the 30,000 DOIs and average counts for the papers with at least one event (intensity) in the sample.

Table1. General statistics (sum and mean) of events for all the 30,000 DOIs across common sources

	P Mendeley	TMR	MR	P Facebook	TF	MF	P Twitter	TTw	MTw	P CiteULike	TCi	MCi	P Reddit	TRe	MRe
Mendeley.com	20,677 (69%)	212,292	10.2												
Lagotto	13,742 (46%)	143,362	10.6	894 (4%)	26,213	29.3	58 (.1%)	3438	59.2	756 (2.5%)	1100	1.4	109 (.3%)	359,585	3298.9
Altmetric.com	6209 (21%)	91,827	14.7	1622 (5%)	3773	2.3	6221 (21%)	35,738	5.7	593 (1.9%)	953	1.6	77 (.2%)	116	1.5

P:no of papers with at least one event per different source; TMR: total Mendeley readers, MR: mean Mendeley reader; TF: total Facebook counts, MF: mean Facebook count; TTw: total Twitter counts; MTw: mean Twitter counts; TCi: total CiteULike counts; MCi: mean CiteULike count; Tre: total Reddit; MRe: mean Reddit

According to Table 1, several discrepancies among these altmetrics data provider/aggregators in reporting metrics can be reported. Regarding the coverage of DOIs per provider, Mendeley has the highest coverage 20,677 (69%), Lagotto 20,364 (68%) and

Altmetric.com 6,946 (23%). As expected Mendeley provides the highest values of readership counts compared to Lagotto and Altmetric.com. Lagotto provides the highest number of Facebook counts, Reddit mentions and CiteULike counts. Altmetric.com provides the highest number of tweets. Regarding 'intensity' (average counts for the papers with at least one event) there are differences across the data providers in the common data sources (Tweets, Facebook, CiteULike, Reddit and Mendeley). Altmetric.com has a higher twitter coverage (21%) and Facebook coverage (5%) than Lagotto Twitter (0.1%) and Lagotto Facebook (4%). For CiteULike (2.5%) and Reddit (.3%) Lagotto has higher coverage than Altmetric.com (with CiteULike (1.9%) and Reddit (.2%)).

Differences in the metrics across sources for the overlapping DOIs

Regarding the overlapping papers, some differences in the metrics across the sources are also observed. For example, in case of Mendeley readerships, a small set of DOIs have higher Mendeley reader counts as reported by Lagotto (for 76 papers the differences is between 1 to 3 counts) and Altmetric.com (for 136 papers the differences is between 1 to 80 counts) than Mendeley itself. There are 3310 DOIs (with 1 to 60 count differences) where Mendeley scores is higher than in Altmetric.com. For 1 DOI, there is higher value (35 reader count) reported by Mendeley than Lagotto (zero reader count). Possible reason for the different values may be due to the delay in the updates of both Altmetric.com data and Lagotto with regards to Mendeley data. There are also huge differences between Reddit counts reported by Lagotto and Altmetric.com (the differences is between 1 to 176,779 counts). Reddit reported by Lagotto is a compiled score of both mentions and comments while Reddit by Altmetric.com include the original posts without the comments. Tweets and Facebook counts reported by these two sources also present substantial differences. For tweet counts via Lagotto there is a limitation of 100 tweets per DOI (needing update code to collect more than 100 tweets). Another reason may be due to the different methodology in collecting tweets: Altmetric.com collects online mentions of scholarly papers from Twitter in real-time. They track links (by resolving URLs) to papers within tweets. Tweets reported by Altmetric.com includes public comments and retweets but no favorites. Facebook counts are aggregated (sum of shares, likes and comments) in Lagotto while Altmetric.com reports only public posts.

Conclusions and discussions:

Some of the main reasons for the different counts relate to the different methods in collecting and processing metrics by the different providers. How each provider queries from sources also matters (using DOI or other metadata), using different APIs (for example for Facebook and Twitter) or possible time lags in the data collection or updating issues. Furthermore, if the data provider is reporting the public Facebook counts or public tweets or compiling all the retweets or favorites in one metric or as a separate value also cause differences in the counts. There are also issues with tracking DOIs from difference registration agencies. Moreover, there are issues with the quality of metadata for which

altmetrics are collected, for example differences in publication dates between WoS and Crossref. Other problems include accessibility issues (e.g. with Twitter), resolving DOIs to URLs issues (e.g. differences across publisher platforms in resolving DOIs to journal landing pages, cookies problems, access denies), etc.

These results emphasize the need for adhering to best practices in altmetric data collection both by altmetric providers and the publishers. Future steps include developing guidelines and recommendations regarding altmetric data collection to introduce transparency and consistency across providers. NISO in 2015 has initiated a working group on altmetrics data quality and the group has developed a draft code of conduct for collection, processing, dissemination and reuse of altmetric data that can contribute to solve many of these issues.