GROUP 1

External students

Members

Paul Ollivier
Stephen Maher
Jinxi Luo

.

Contents

1. Project Goal

A key element of the Altmetric research scoring system and focus is the emphasis on attention and more specifically, social media attention. This approach contrasts with more traditional research scoring methodologies which focus more typically on research citations and the number and frequency of published research.

The goals of this project are twofold. The first is to evaluate a range of research scoring dimensions (including subject area, journal, institution, citation counts and peer scoring) against social attention measures in order to identify key drivers for the Altmetric score. The second goal utilises this evaluation to identify practical and effective uses of the Altmetric analysis.

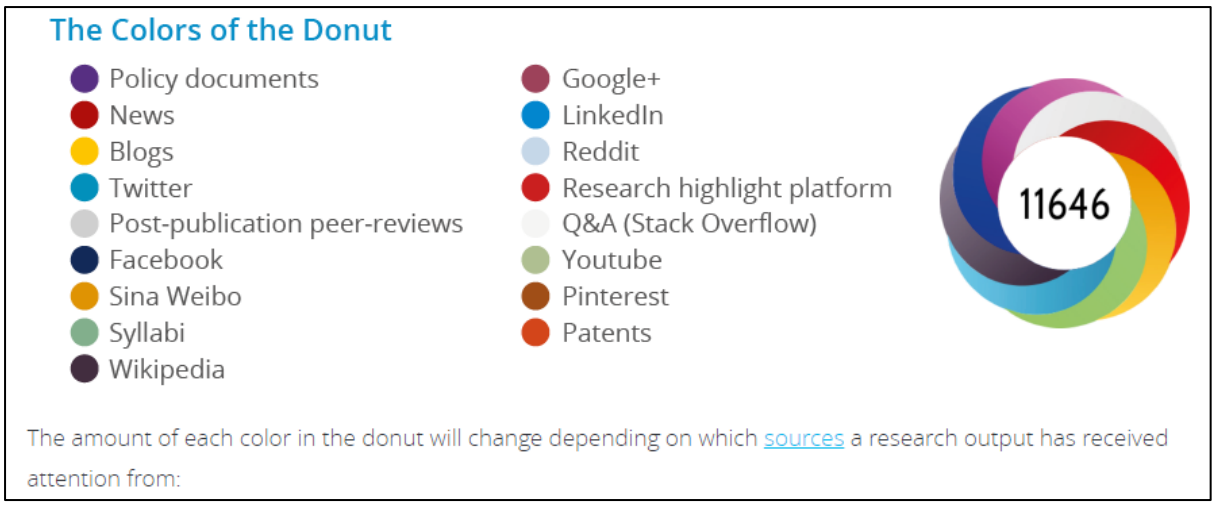
2. Project Process

Alternative metrics are developing branches for scoring research and similar publications in the field of scientometrics and the broader field of bibliometrics. More traditional measures of researcher and institutional research output and quality such as Journal Impact Factors or the h-index, have a propensity to focus on citation counts and number of papers published. In contrast, alternative metrics utilise web-based data, along with the data sources used by the more traditional measures, in order to add a social media (attention) based dimension to the analysis of researcher and institutional output.

The Altmetric organisation produces what is most likely the highest profile alternative metric incorporating social media elements and their output is characterised by the rainbow donut (Figure 1). Altmetric claims to "collect and collate" web-based activity relating to published research along with traditional citation and peer review activity, in order to produce "a single visually engaging and informative view of the online activity surrounding [your] scholarly content."[1]

[1] Altmetric, <https://www.altmetric.com/about-our-data/the-donut-and-score/>, viewed 10 April 2020.

Figure 1: The Altmetric Donut



Source: Altmetric, <https://www.altmetric.com/about-our-data/the-donut-and-score/>, viewed 10 April 2020

While the Altmetric donut is clearly visually striking, subsuming traditional measures of researcher and institutional output and quality into a broader measure begs the question as to the value of the Altmetric score (or similar social attention scoring systems) for these traditional uses. It also raises the question as to potential uses beyond those reflected in the traditional measures. To this point, Altmetric profiles a range of uses for measure and analytics on their website. These uses include[2]:

- allowing publishers to determine how "research is being shared and discussed online";
- supporting research institutions understand and interpret "the attention surrounding [the] institution's research";
- supporting "commercial strategy with up-to-the minute business intelligence" on a researcher's research and research from other relevant sources;
- allowing researchers to "track and demonstrate the reach and influence of [their] work to key stakeholders";
- supporting sponsors of research in "[monitoring] and [reporting on the online discussion surrounding the work [they] fund".

This paper will focus on the Altmetric attention score and sets out a framework for evaluating the Altmetric attention score and its constituent elements. The paper will:

- establish a conceptual framework for the analysis and define relevant focus points for the university;
- briefly review data assets available for the analysis and highlight key results in preparing data for use;
- summarise the results of the initial data explorations, and
- outline the plan for the first iteration of the analysis.

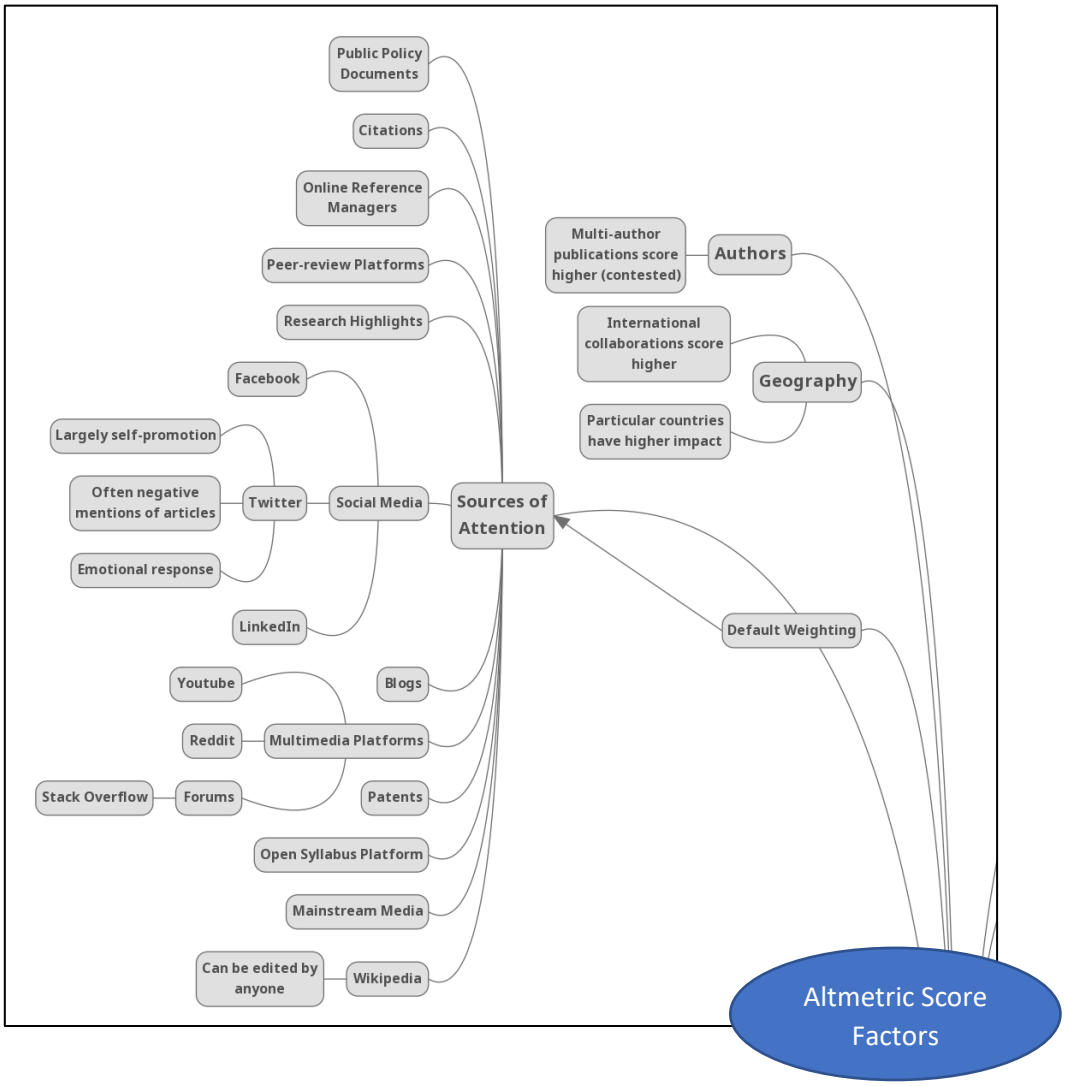
[2] Altmetric, <https://www.altmetric.com/audience/>, viewed 10 April 2020.

3. Project Problem Space

A conceptual landscape was developed for this analysis through an extensive literature review. The literature review focussed on the concepts of alternative metrics and scientometrics more broadly. The conceptual landscape, centred on the Altmetric scoring system, is visualised in Appendix 1.

Clearly, research scoring is a broad and well-established field. And while alternative metrics and scoring providers such as Altmetric are relatively new entrants into the field, their conceptual coverage is by no means narrow. This analysis is iterative and will start by focussing on alternative metrics and the Altmetric scoring system more specifically (Figure 2). The key reason for this is that too broad an analysis risks derailing the project through lack of focus. However, it is necessary to be mindful of the applications for scoring metrics and the analysis will look to provide context to the value of the Altmetric scoring system relative to traditional metrics.

Figure 2: Altmetric Scoring Conceptual Landscape (upper left quadrant, Appendix 1)



4. Project Data Assets

Three data sets have been sourced to support the first iteration of this analysis. The principal dataset supporting this analysis comes from Altmetric and comprises data pertaining the Top 100 ranked papers (by Altmetric score) for the years 2013 through to 2019. Two additional datasets included in the analysis come from Plum Analytics[3] and Bornmann and Haunschild[4] which will be used to critique the interpretation of the Altmetric data. A key issue pertaining to all datasets is one of "passive data quality". That is, it is not feasible to check the methodology used and the value recorded in individual fields within all datasets. Hence there is a passive reliance on each of the contributors

The Altmetric dataset and the Plum Analytic dataset contain a range quantitative and qualitative fields encompassing identifying information such as:

- Identifying information (DOI and other centralised database ID's);
- Research paper names, authors, publishers, supporting institutions;
- Traditional research scores;
- Social attention measures, and
- Altmetric scores.

While the datasets are extensive, a range of data quality issues were identified and are systematically addressed in this analysis. Within the Altmetric dataset, key observations were:

- Inconsistent attribute names through 2013-2019
- Missing data and column information through 2013-2019
- Missing values within columns
- Extensive gaps in available attributes
- Poorly defined attributes where a definition cannot be inferred

Table 1 outlines general broad categories in which attributes in the Altmetric datasets belong and highlights the data quality issues described above and relevant attributes which may be used for analysis.

Notwithstanding the issues outlined for the Altmetric dataset, these can be, and were, accommodated through aligning attributes across years, aligning categorical variable values (within attributes) and removing attributes deemed to be of little or no value (further detail in Appendix 2). This was achieved by establishing context for each attribute by year. With these adjustments, attributes relevant to the analysis become more clearly identifiable and data gaps are reduced.

[3] Plum Analytics 2016, PlumX Altmetrics & Sci-Hub Downloads, Figshare.
[4] Bornmann, L & Haunschild, R 2018, 'Do altmetrics correlate with the quality of papers? A large-scale empirical study based on F1000Prime data.', PLoS ONE, vol. 13, no. 5, p. E0197133.

Table 1: Altmetric Dataset: Date Quality Summary

Attribute Deviation Table	2013	2014	2015 & affiliation	2016 & affiliation	2017 & affiliation	2018	2019	Consistent Attributes
Altmetric Info	Missing Data Sources Undescribed Data	No Altmetric Data	Missing Data Sources	Missing Data Sources Undescribed Data		Undescribed Data	Undescribed Data	Popular Data Sources AAS
Unique Identifiers				Extra ID Data		Extra ID Data	Extra ID Data	DOI Altmetric ID ArXiv ID PubMed ID Journal ISSNs
Publication Info				Undescribed Data		Undescribed Data	Undescribed Data	Category Title Publication Date AccessType Author Description Journal
Affiliation Info	No Affiliations		Undescribed Data	Undescribed Data		Undescribed Data		Affiliated Institution Country of Institution

For the remaining datasets (Plum Analytics and Bornmann and Haunschild), less work was required to make these fit for use.

The Plum Analytics dataset is standalone and only covers a single period (six months to early 2016). All fields were fully documented by the aggregator and corrections where applied, were also documented. A range of fields were partially filled, and these will be removed from the final dataset. An additional two fields relating to publishing cost will also be removed.

The Bornmann and Haunschild dataset is also standalone and the dataset covers a three-year span (2011 top 2013). All fields were fully documented by the aggregators. All fields are complete, but it must be noted that individual papers and subject areas, etc. have been excluded from the dataset.

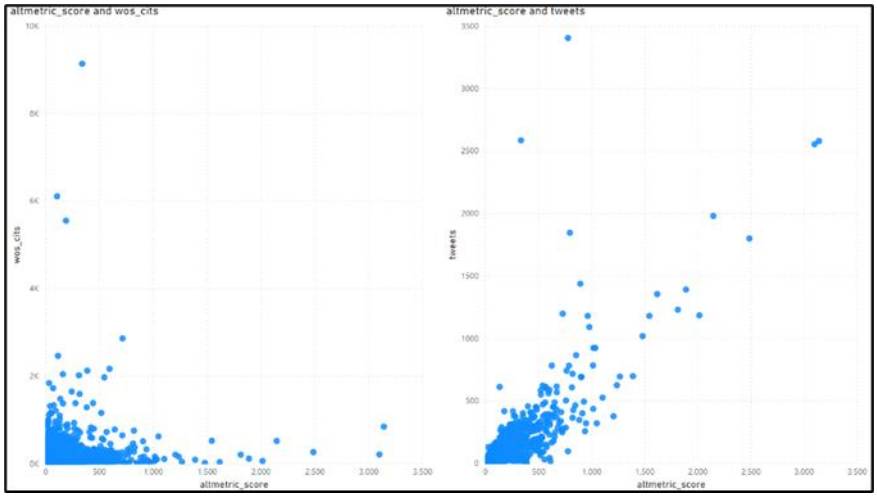
Despite both these datasets being well documented by their respective aggregators, there remains a dependence on their underlying data collection methodologies and processes.

5. Exploratory Analysis

There were a few key observations from the Altmetric and other datasets:

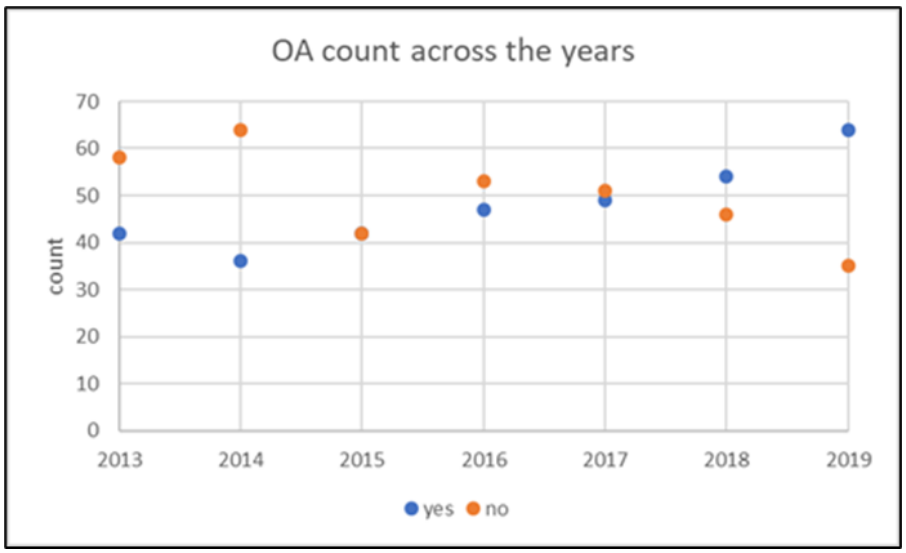
- There appears to be a weak, at best, relationship between traditional measures of quality (such as citations, downloads, and other factors) and social attention scores. The Bornmann and Haunschild dataset was reasonably definitive on this point, but this was also observed in the other datasets (Figure 3, Appendix 3: Figures 19, 20 and 21).
- Research available through Open Access and Free publications has improved Altmetric scores in later years (Figure 4).
- A few key subject categories appear to dominate social mentions (Appendix 3: Figures 15, 16, and 17).
- A few key publications appear to dominate social mentions.

Figure 3: Web of Science Citations and Tweets vs. Altmetric Score



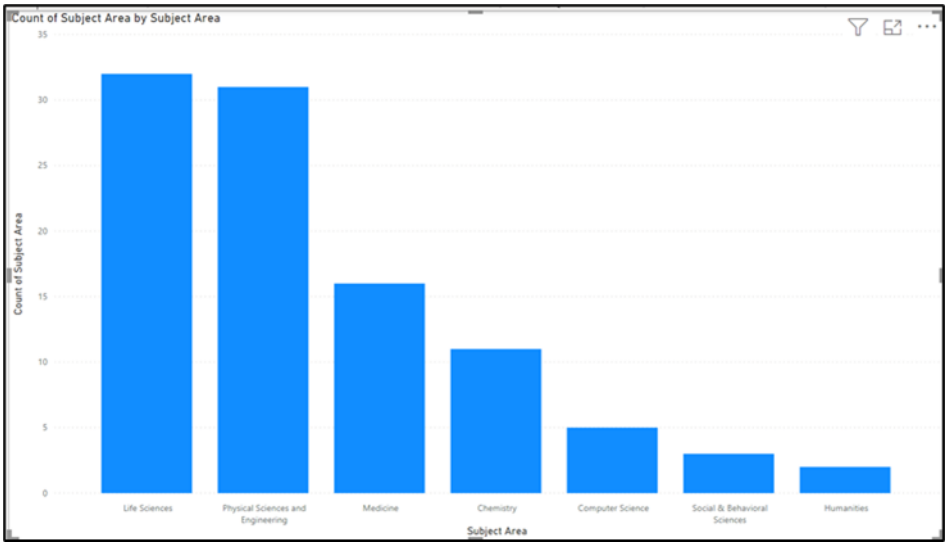
In addition to these observations, it was also noted that in the period 2013 through to 2016 (in the Altmetric data), most categories for non-open access publications had in general a larger sum Altmetric score. This characteristic was also present in the Plum Analytics dataset. From 2017 onwards, it was noted that a greater number of publications became open source (Figure 4, Appendix 3: Figure 8).

Figure 4: Change in Counts of Open and Non-open Access Type in the Altmetric Top 100 Publications, 2013 to 2019



Past 2016, this bias to a higher aggregate Altmetric score appears to have been unwound, with subject matter appearing to more a more important indicator of higher aggregate Altmetric scores. This subject characteristic (that is, a few subjects dominating social mentions) was also present in the Plum Analytics (Figure 5). It is worth noting that the Plum Analytics dataset is based on a set of most downloaded papers over a six-month period. As such, it may be more reflective of quality than the Altmetric datasets (driven by attentions scores) but still suggests that subject matter is a determinant of attention. For example, Medical Sciences were prominent and appeared to score highly on Altmetric scores post-2016. In a similar vein, hard physical sciences appeared to have more social mentions in the Plum Analytics dataset, while Life Sciences had greater downloads.

Figure 5: Plum Analytics: Count of Subject Area by Subject Area

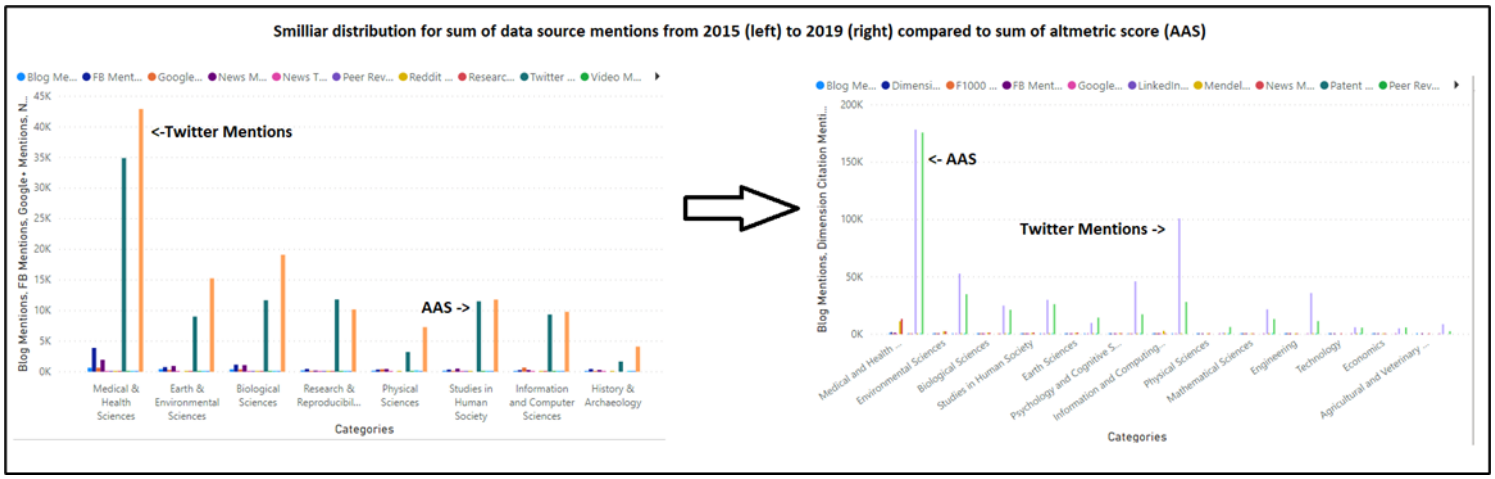


In terms of social mentions, twitter mentions has the highest mentions per category, with top three categories being, Medical and Health, Studies in Human Society and Biological Sciences (Altmetrics

dataset). Moreover, there appears to be a definite relationship between twitter mentions and Altmetric scores (Altmetric and Bornmann and Haunschild datasets). But the relationship is imperfect, and this could be explained by the Altmetric score constituent weightings. Further, within each data source (used by Altmetrics), there are additional rules applied that source. This modification to the Altmetric score is not reflected in the current data, as a plot of data source counts with their default altmetric weights shows that in most instances the expected sum altmetric score is far exceeded by the sum of two individual data sources: News and Twitter mentions (Figure 6, Appendix 3: Figures 11 and 12).

There does not appear to be a trend in Altmetric score by publication category, though there is some indication that time of year (for publication may have an impact). With respect to publications more specifically, Nature and Science have consistently been the top journals (by sum Altmetric score) in the Altmetric dataset. However, Springer Nature was the top publisher in 2018 and 2019.

Figure 6: Distribution in the Sum of Altmetric Data Sources and AAS per Category from 2015 to 2019



Based on this exploratory analysis, the data suggest that focussing on subjects ranked highly by Altmetrics may deliver a higher Altmetrics score. This means a focus on subject matter which delivers proportionally higher social mentions. In addition, it may also be worthwhile to focus promotion (via social media) for research on twitter and News sources, while increasing the number of affiliated institutions per research paper. These practices may help increase the Altmetric attention scores that are received for research. It should be noted however, that there is a clear distinction between attentions and quality and further study is proposed in the analysis plan to determine the significance of these results.

6. Analytical Plan

The analytical plan is designed to evaluate the relationships between the attributes (variables) in each dataset. As outlined in Project Goal, the focus for the analytical plan is to identity collective relationships (clustering), direct liner relationships (regression modelling) and indirect liner relationship (factor analysis). The core focus will be on the Altmetric dataset for its overall size and time span.

The remaining two datasets will be used to provide supporting evidence (or otherwise) for any results derived from the Altmetric dataset.

6.1 Data Source: Altmetric Data Set 2013 – 2019

Below is a list of variables which will be used for analysis grouped by type. A list of analytical techniques is proposed for the given variables with consideration given to the research goal. Not all proposals will be pursued but are considered if either a problem is encountered in the initial analysis or the initial analysis indicates areas of specific interest.

Discrete (quantitative)
(Altmetric Attention Score) (Blog Mentions) (F1000 Mentions) (FB Mentions) (Google+ Mentions) (News Mentions) (Patent Mentions) (Policy Mentions) (Reddit Mentions) (Twitter Mentions) (Video Mentions) (Wikipedia Mentions)
Binary (qualitative)
(ArXiv ID) (Dimensions ID) (DOI)(Journal ISSNs) (PubMed ID)
Nominal (qualitative)
(Category) (Country) (Journal) (OA)

Proposed analysis 1: Clustering [J S]
Variables: (Blog Mentions) (F1000 Mentions) (FB Mentions) (Google+ Mentions) (News Mentions) (Patent Mentions) (Policy Mentions) (Reddit Mentions) (Twitter Mentions) (Video Mentions) (Wikipedia Mentions) (Altmetric Attention Score) (Category)
Goal: To determine which altmetric data source is most used for a category and to see if that has a positive effect on altmetric score.
Models

- K means
    - Altmetric score and Altmetric data source counts

- K prototype / K mode

- Some combination of all those variables of interest.

Timeframe: Three weeks

Proposed analysis 2: Association [J]
Variables: (Altmetric Attention Score) (Category) (Country) (ArXiv ID) (Dimensions ID) (DOI)(Journal ISSNs) (PubMed ID) (OA)(Journal)
Goal: To determine common associations that may exist between altmetric scores and a publication's category, affiliated institution's country, journal of the publication, ID and access type.
Method: With the Apriori Algorithm: support, confidence and lift will be used to evaluate a rule.
Timeframe: Three weeks

Proposed analysis 3: Multiple Linear Regression [J]
Dependent variable: (Altmetric Attention Score)
Independent variables: (Blog Mentions) (F1000 Mentions) (FB Mentions) (Google+ Mentions) (News Mentions) (Patent Mentions) (Policy Mentions) (Reddit Mentions) (Twitter Mentions) (Video Mentions) (Wikipedia Mentions)
Hypothesis: Altmetric Attention Score is independent of all Altmetric data sources including Dimension and Mendeley mentions. To determine which data source institutions should focus on advertising their publication on to increase research engagement.
Model Assumptions

- **Check Independence of observations:** Each observation should be independent due to the nature of data. However, residual analysis and Durbin Watson statistic will be used to check.
- **Check Linearity with dependent and independent variables:** Scatter plots can be used to check. Data will be transformed or removed if they do not satisfy.
- **Check for Homoscedasticity:** Residual analysis will be used to check. Variables transformed if needed
- **Check for normally distributed residuals:** Normal Q-Q plots used to check.

Timeframe: Three weeks

Proposed analysis 4: Multiple Linear Regression [J]
Dependent variable: (Altmetric Attention Score)
Independent variables: (ArXiv ID) (Dimensions ID) (DOI)(Journal ISSNs) (PubMed ID)
Hypothesis: Altmetric Attention Score is independent of a publication's ID. To determine if making a publication available on certain data bases will increase research engagement.
Model Assumptions

- **Check Independence of observations:** Each observation should be independent due to the nature of data. However, residual analysis and Durbin Watson statistic will be used to check.
- **Check Linearity with dependent and independent variables:** Scatter plots can be used to check. Data will be transformed or removed if they do not satisfy.
- **Check for Homoscedasticity:** Residual analysis will be used to check. Variables transformed if needed

- **Check for normally distributed residuals:** Normal Q-Q plots used to check.

Timeframe: Three weeks

6.2 Plum Analytics Data

Below is a list of variables which will be used for analysis grouped by type. A list of analytical techniques is proposed for the given variables based on the research goal. Not all proposals will be pursued but are considered if either a problem is encountered in the initial analysis or the initial analysis indicates areas of specific interest.

Discrete (quantitative)
(Sci-Hub downloads) (Captures) (Social media) (Mentions) (Usage)
Nominal (qualitative)
(Subject Area) (Type)

Proposed analysis 1: Clustering [S]
Variables: (Sci-Hub downloads) (Year) (Captures) (Social media) (Mentions) (Usage)
Goal: To determine along which characteristics, if any, the data demonstrates clustering.
Model

- K means

Proposed analysis 2: Factor Analysis and Principal Components Analysis [S]
Variables: (Subject Area) (Type) (Sci-Hub downloads) (Captures) (Social media) (Mentions) (Usage)
Goal: To look for factors, if any, along which the data aligns and data associations.
Timeframe: Three weeks

6.3 Bornmann and Haunschild Data

Below is a list of variables which will be used for analysis grouped by type. A list of analytical techniques is proposed for the given variables based on the research goal. Not all proposals will be pursued but are considered if either a problem is encountered in the initial analysis or the initial analysis indicates areas of specific interest.

Discrete (quantitative)
(pubyear) (wos_cits) (sco_cits) (tweets) (me_readers) (total_f1000_score)
Continuous (quantitative)
(altmetric_score) (citescore) (item_ijif)

Proposed analysis 1: Clustering [S]
Variables: (pubyear) (wos_cits) (sco_cits) (tweets) (me_readers) (total_f1000_score) (altmetric_score) (citescore) (item_ijif)
Goal: To determine along which characteristics, if any, the data demonstrates clustering.

Model

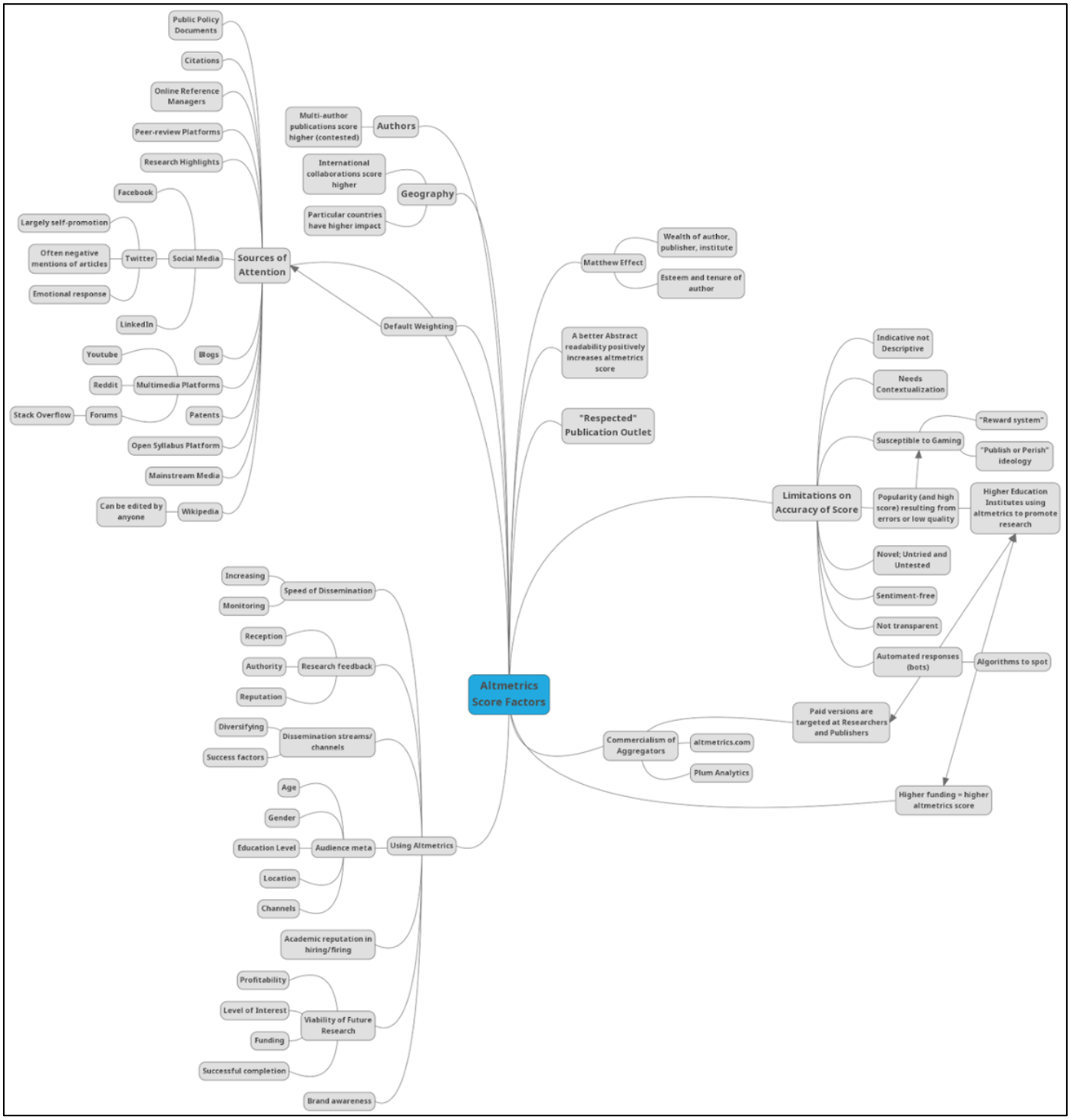
- K means

Proposed analysis 2: Factor Analysis and Principal Components Analysis [Stephen]
Variables: (pubyear) (wos_cits) (sco_cits) (tweets) (me_readers) (total_f1000_score) (altmetric_score) (citescore) (item_ijif)
Goal: To look for factors, if any, along which the data aligns and data associations.
Timeframe: Three weeks

7. Appendix 1: Conceptual Landscape

Conceptual landscape visualisation of the Altmetric Scoring System. This will be iteratively developed through the project.

Figure 7: Altmetric Scoring Conceptual Landscape



8. Appendix 2: Data Review and Preparation

Three datasets have been sourced for this analysis. The first dataset comprises data underpinning the Altmetric Top 100 Scores for the years 2013 through 2019. The second dataset is sourced from Plum Analytics[5] and contains attention and quality data for the top 100 Sci-Hub downloads for late 2015 through to early 2016. The third dataset is sourced from a paper published by Bornmann and Haunschild[6], which focuses on the relationships between attention score (Twitter and Altmetric) and research paper quality.

While the primary focus of this analysis is on the Altmetric dataset, the remaining two datasets provide an opportunity to critically evaluate the results of the Altmetric analysis.

8.1 Data Source: Altmetric Data Set 2013 – 2019

Table 2 contains the attribute names for each Altmetric dataset by year and divided into four categories. Attribute names throughout the years are similar in meaning with the primary differences being capitalisation and variations in naming. Attribute values were checked by comparing values across years. This does not impact the analysis except and may be confusing, hence a convention was adopted to correct for this.

Only data of relevance that has been found to not be of high quality described in this section. Quality determination was based on the dimensions of data quality. The Altmetric Info section primarily describes the Altmetric attention score and the attention counts collected from Altmetric data sources. One column of altmetric attention scores were float values, this will be corrected by conversion to a whole number and rounded up as otherwise the data would be invalid. URL Links to the altmetric donut and score are also included for a given publication (small/medium image and badge URL), this information is not useful for our project as it provides no additional information and it is not feasible to access each link individually to check the contents.

Context for the following attributes in Altmetric Info data are unknown: they are all categorial attributes with factor levels of Yes or No. More research may be done to determine their meaning. At present, they are deemed not relevant to the current research.

- Checked?
- Stayin in?
- Had Corrections?

[5] Plum Analytics 2016, PlumX Altmetrics & Sci-Hub Downloads, Figshare.
[6] Bornmann, L & Haunschild, R 2018, 'Do altmetrics correlate with the quality of papers? A large-scale empirical study based on F1000Prime data.', PLoS ONE, vol. 13, no. 5, p. E0197133.

The Unique Identifiers column contains ID attributes that are used to track a publication back to a database containing a listing of the publication. As they are IDs, they are also unique and may be used as keys to connect data sets, it was found that the following connections exist when connecting affiliation data sets to their respective years:

- (2015aff) altmetric_id references (2015) id
- (2016aff) DOI references (2016) DOI
- (2017aff) DOI references (2017) DOI
- (2017aff) arXivID references (2017) arXvID

Apart from this use, the use of ID attributes in the analysis may be difficult as it is infeasible to reference each ID to their data bases for error checking, obtaining extra information or to connect datasets across different years. However, a publication having a unique identifier (ID) is useful, as authors that push to get their publication(s) on different databases may improve their capacity to distribute their publication(s) to a wider audience. One potential, yet to fully considered, is some form binary analysis where the analysis focusses on whether a publication has or does not have a particular ID. This decision will reflect the fact that a many of the ID columns have empty values. It is not feasible to fill in these values with an appropriate ID as it either may not exist may be difficult to find. It is also difficult to check if all IDs are correct, while some have fixed digit lengths others are mix of numbers and letters that have different formats depending on organisation supplying the ID.

The Publication Info section contains information on publications, when the research was published, author names, the publishing journal, type of access and links to the publication. Attributes of interest are the publications category, publishing journal, date published and type of access.

There were few problems with Category type attributes except for 2013 where factors levels of the category were poorly specified and had levels differing from those in subsequent years. There were also a few values missing. As Category type attributes are of interest, missing entries where manually replaced with an appropriate category. To mitigate the poor factor specification, affected attributes will be aggregated with the same attributes in subsequent years was performed to reduce the amount of bias added to the datasets and analysis.

There were few problems with publication date other than inconsistent attribute names and format. Most columns only needed their data format to be changed in Excel to one standard format. One column had specific 00:00:00 time appended to the data that was removed. The main issue was with the 2019 publishing times which were listed as 5-digit numbers. It Is unclear how this may be decoded. Hence the column may not be of use. It is noted however, that it could be a representation of time in terms of the order in which each publication was published based on numerical order.

Table 2: Altmetric dataset data understanding

Year	Altmetric Info	Unique Identifiers	Publication Info	Affiliation Info
2013	(Altmetric Attention Score) (Reddit threads)(Bloggers)(Tweeters)(Google+ authors)(F1000reviews)(Pinterest posts)(News outlets) (Q&A site users) (Facebook walls) (Mendeley readers) (CiteULike readers) (Small image) (Medium image)	(DOI) (PMID) (ArXiv ID) (Journal ISSNs)	(Category 1) (Title) (Journal) (URL) (OA)	
2014		(DOI) (Altmetric ID)	(Title) (Journal) (URL) (Published) (Author name) (Access type) (Category)	(Institution) (Country) (Region)
2015 Affiliation		(altmetric_id)		(Name) (Lat) (Lng) (Country) (type)
2015	(Score) (News_tweet_ratio) (Count news) (Count blog) (Count twitter) (Count facebook) (Count peer review) (Count weibo) (Count google plus) (Count reddit) (Count research hi) (Count video) (Count wikipedia)	(Doi) (id)	(Title) (Journal) (Date) (Authors) (Rank) (Link) (Journal rank) (Categories) (Open access)	(Countries) (Institutions)
2016 Affiliation		(DOI)	(name) (confidence) (search_type)	(Affiliation) (Grid_id) (Grid_name) (Grid_city) (Grid_state) (Grid_country)
2016	(Score) (Checked for gaming/incorrect mentions) (Checked?) (Stayin in?) (Had corrections?) (Total_posts_msm) (total_posts_blog) (total_posts_policy) (total_posts_tweet) (total_posts_peer_review) (total_posts_wikipedia) (total_posts_weibo) (total_posts_fbwall) (total_posts_gplus) (total_posts_linkedin) (total_posts_rdt) (total_posts_pinterest) (total_posts_f1000) (total_posts_video) (total_posts_qna)	(DOI) (Altmetric ID)	(Position) (title) (Journal) (Publication Date) (link) (sharedit) (OA) (Content Type) (subject) (Blurb)	(Has affiliations?)
2017 Affiliation		(Dimensions ID) (DOI) (ArXiv ID) (PubMedID)	(Last name) (First name)	(Affiliation) (Country)
2017	(Score) (Number of news stories) (Number of blog posts) (Number of policy documents) (Number of tweets) (Number of peer reviews) (Number of weibo posts) (Number of Facebook posts) (Number of Wikipedia pages) (Number of Google+ posts) (Number of Linkedin posts) (Number of Reddit posts) (Number of pins) (Number of F1000 posts) (Number of Q&A posts) (Number of videos) (Number of syllabi) (Number of Mendeley readers)	(DOI) (PubMed ID) (ArXiv ID) (Altmetric ID)	(Title) (Journal) (Date) (Subject) (Description) (OA)	
2018	(News mentions) (Blog mentions) (Policy mentions) (Twitter mentions) (Patent mentions) (Peer review mentions) (Weibo mentions) (Facebook mentions) (Wikipedia mentions) (Google+ mentions) (LinkedIn mentions) (Reddit mentions) (Pinterest mentions) (F1000 mentions) (Q&A mentions) (Video mentions) (Syllabi mentions) (Number of Mendeley readers) (Number of Dimensions citations) (Altmetric Attention Score) (Badge URL)	(Journal ISSNs) (Handel.net IDs) (ADS Bibcode) (ArXiv ID) (RePEc ID) (SSRN) (URN) (PubMed ID) (PubMedCentral ID) (DOI)	(Rank) (Title) (Description) (Journal/Collection Title) (Publisher) (Subjects) (Subjects (FoR)) (Publication Date) (DOI URL) (Details Page URL) (Open_Access) (Authors) (Publisher-preferred links)	(Affiliations (GRID)) (Funders) (Unique countries)
2019	(Altmetric Attention Score) (News mentions) (Blog mentions) (Policy mentions) (Twitter mentions) (Patent mentions) (Peer review mentions) (Weibo mentions) (Facebook mentions) (Wikipedia mentions) (Google+ mentions) (LinkedIn mentions) (Reddit mentions) (Pinterest mentions) (F1000 mentions) (Q&A mentions) (Video mentions) (Syllabi mentions) (Number of Mendeley readers) (Number of Dimensions citation)	(Journal ISSNs) (Handel.net IDs) (ADS Bibcode) (ArXiv ID) (RePEc ID) (SSRN) (URN) (PubMed ID) (PubMedCentral ID) (DOI)	(Title) (Journal/Collection Title) (Open_Access) (Description) (Publication Date) (DOI URL) (Details Page URL) (Publisher) (Subjects) (Authors) (sharedit) (Publisher-preferred)	(Affiliations (GRID))

Open access information can be generalised as either Yes or No for a publication and factor levels which had a similar meaning were grouped as such. There was one column which only contained values detailing if a publication is open access. Hence, it was decided that empty values will be treated as 'not open access'. In the analysis section, this decision was reasonable as it matches distributions in subsequent years.

The remaining attributes are of less interest for the analysis. Data quality was not considered, and any missing values present in those columns were replaced with NULL to explicitly represent data non-existence:

- Title, Author name, Blurb, Description
    - These attributes contain mostly text information that may be analysed in the analysis plan with association rules.
- URL, Link, sharedit, DOI URL, Details page URL, Publisher preferred links
    - Links are not useful to the analysis, though they can be used to support fact checking and repair of the dataset.
- Confidence, search_type
    - Confidence is a float value, and search type is a string. The specific context beyond these attributes is unknown, hence they cannot be used.
- Rank/position
    - Has no purpose other than numbering publications as they are listed by default.
    - This information, however, was helpful to see that 2018 had more than 100 publications. The excess was removed to reduce inconsistencies during analysis and because the entries were missing values for most of the attributes of interest.

The Affiliation Info section contains attributes that describe affiliated institutions for a publication, such as name of the institution, geographic location, and the type of institution. Attributes of interest are location information and type of institution. Specifically, location country as it is the most complete attribute for all years. Only 2016 provides specific city and state locations as attributes, around half of the entries are empty and so would not be useful for analysis.

Country attributes for most years are of good data quality, however there are some empty values. As it is not feasible to manually enter the appropriate institution, a NULL entry was used. This should not materially affect the analysis as only the 2015 Affiliations data set had this issue and around 30% of the entries were empty.

Type attribute had some empty values, but these were few. It was not possible to judge, without further research, the factor level to which an institution belonged. Hence, a new level was created and set to "Private" as a replacement for missing values. It was possible to delete Affiliations with missing values as it will not affect the number of publications. This was not done as those missing values may be crucial to the analysis and the missing values may be due to an unnamed category.

Any remaining attributes were deemed to be of interest. As a result, data quality was not considered and any missing values that exists in those columns were replaced with NULL to explicitly represent that the data does not exist. These were:

- Institution/affiliation/affiliations GRID/Funders/name/Funders
    - These attributes contain mostly text information that may be analysed in the analysis plan with association rules.
- Has affiliations?
    - Limited purpose as all entries were 'Yes'.

8.2 Data Source: Plum Analytics: PlumX Top 100

Parkhill[7] published a blog post detailing the characteristics of the top 100 Sci-Hub downloads through the period September 2015 to February 2016. The data supporting this research was published by Plum Analytics.

Sci-Hub is best described as a source of "pirated" research papers which sources copyrighted material from legitimate journal publishers using stolen credentials and functionally acts as file sharing service for research papers. Notwithstanding the illegitimate nature of the Sci-Hub service, Sci-Hub experiences a high volume of downloads from its service (Sci-Hub claimed an estimated 28.0m downloads in 2019). The data supporting Parkhill's blog post is a compilation of PlumX metrics for the top 100 most-downloaded DOIs in the Sci-Hub.

There are 14 attributes and 100 rows within the PlumX Sci-Hub Top 100 downloaded papers dataset. Attributes contain information which uniquely identify each paper, along with measures for estimated value lost to legitimate publishers, subject area, type of research (article, review, etc.), year published, captures (number of bookmarks), social media measures and downloads. The DOI and Title variables are unique identifiers. Sci-Hub downloads, Year, Captures, Social Media, Mentions and Usage are discrete variables. Subject Area is a categorical variable.

While the dataset does not include an Altmetric score per se, the social media measures can be viewed as equivalent measures. Similarly, downloads are viewed as an equivalent to traditional quality measures as these are not present in the dataset either.

[7] Parkhill, M. 2016, "Sci-Hub: The Academic Cat is Out of the Bag", Plum Analytics, viewed and downloaded 27 March 2020, <https://plumanalytics.com/sci-hub-academic-cat-bag-post/>.

8.3 Data Source: Bormann and Haunschild

Bornmann and Haunschild analysed the relationship between alternative metrics (altmetric) and a range of measures of research quality[8]. The analysis focussed on papers research papers reviewed within the F1000prime framework and was thus able to leverage peer reviews (a measure of quality). In addition, as each paper in their analysis was uniquely identifiable (using DOI), the authors matched each paper with additional data from CiteScore, Mendeley and Scopus (citation measurement organisations), along with social attention data and an Altmetric score from Altmetric.

Their dataset comprises 11 attributes and 33,683 rows of data from research papers reviewed on F1000prime between 2011 and 2013. Attributes are divided into three distinct types: year of publication, measures of research paper quality and social attention measures. There are eight attributes reflecting measures for research paper quality (Journal Impact Factor, CiteScore citation index, Web of Science citations, Scopus citations, the sum of the F1000prime review scores and Mendeley reader counts). There are two measures each for Web of Science and Scopus: the year following publication and the sum of the three years following publication. Social attention is measured across two attributes: number of tweets referencing the research and the altmetric score for the research. All attributes comprise either discrete or continuous variables.

The authors noted that the data supporting the social attention scores was sourced from Altmetric and was current as of June 04, 2016. As Altmetric measures continuously evolve through time, current social attention scores as measured by Altmetric are likely to have changed.

The Bornmann and Haunschild paper is particularly relevant to this analysis as the authors used principal component analysis (PCA) and factor analysis (FA) in their research the intention was to determine whether a relationship (if any) existed between quality measures and social attention measures. Their results suggested that there was at best, a very limited relationship between the two.

[8] Data source: https://figshare.com/articles/Do_altmetrics_correlate_with_the_quality_of_papers_-_underlying_data/6120158

8.4 Data Preparation

The initial review of the Altmetric 2013 - 2019 datasets showed that attributes names and factor levels were inconsistent between the years. It was also noted that were missing values in attributes that were of interest. This section details the process to correct these problems, eliminate confusion, and add consistency to the data so that it is in a format suitable for analysis.

The Bornmann and Haunschild dataset had already been extensively cleaned by the authors and requires no further work. Similarly, the PlumX Analytics dataset also required no further work. These datasets are standalone datasets and will be used to corroborate or refute the findings from the Altmetric datasets.

8.5 Data Preparation: Altmetric

The first change made to the data was to standardise attribute names. A set of conventions was established for naming attributes such that underscores were converted to spaces, the first letter is capitalised and the whole word is capitalised if it is an acronym. These changes improve the quality of the information that can be derived from the data by removing ambiguity and providing certain comparisons between attribute names across the data sets. In addition to this base change, there were two other general issues which were addressed:

1. Attributes which are alternative metric information and end with the suffix 'Mentions' are recorded unique instances of any activity by users on each specific social media platform. There is, however, no information as to what constitutes 'activity'. It is assumed that context for each 'Mention' is equivalent by attribute. For example, it is assumed that the 2013 dataset attribute 'Facebook Walls' is equivalent to the 2015 dataset attribute 'count_facebook'.
2. The Category attribute has been re-factored to the levels established in the 2013 and 2015 datasets.

The final modifications to the data were to further improve readability and useability. Unless otherwise noted, these modifications comprised:

- Replacement of missing values with NULL;
- Standardising time formats;
- Setting factor levels for some attributes;
- Removing empty columns, and
- Removing irrelevant entries.

For the analysis plan, the ID and sharedit attributes have been converted to nominal variables. Finally, entries which had a link or ID have had this replaced with "Yes" and entries which were empty (in these columns) have been with replaced with "No".

8.6 Data Preparation: Plum Analytics

The Plum Analytics dataset is a standalone dataset and it is not anticipated that it will be linked to the Altmetrics dataset. Minimal changes were made to the PlumX Top 100 for this analysis and these comprised:

- Removal of the "Publisher Cost" and "Sci-Hub x Pub Cost" (measure of revenues forgone through pirating) columns as costings will not be part of this analysis.
- Values in the ISBN and PMID columns to be replaced with "Yes" and empty cells were replaced with "No".

8.7 Data Preparation: Bornmann and Haunschild

The Bornmann and Haunschild dataset was compiled from wide range of data sources. It is a standalone dataset and it is not anticipated that it will be linked to the Altmetric dataset. The sources for the dataset were:

- F1000prime;
- CiteScore;
- Scopus;
- Web of Science, and
- Altmetric.

The Journal Impact Factor used by Bornmann and Haunschild was sourced from a database managed by Haunschild. In their paper, Bormann and Haunschild detailed the process with which they cleansed and merged their datasets. The data as it stands is suitable for immediate use.

8.8 Data Dictionary and Quality Assessment

Table 3: Altmetric Data Dictionary

Attribute Name	Definition	Format	Format exceptions
ADS Bibcode	Bibliographic code to identify publications in the ADS data base. Astrophysics data system bibliographic code.	String (YYYYJJJJJVVVVMPPPPA)	NULL or string value
Affiliation	Provides the name of affiliated institution/s to the publication	String	NULL or string value with comma delimiter
Altmetric Attention Score	The altmetric attention score for a publication A value based on weights assigned to social media mentions of an article or similar published document, including peer reviews, Wikipedia citations, discussions on research blogs, mainstream media coverage, bookmarks, and mentions on social networks such as Twitter.	Positive Integer including 0	
Altmetric ID	ID to reference a publication on the altmetric site	Positive Integer	Must be 8 digits
ArXiv ID	ID to reference a publication on arxiv	String (YYMM.NNNNN)	Null or string value
Authors	Name of the author/s of the publication	String (FirstName SecondName Initials)	Null or string value with comma delimiter
Blog Mentions	Count of unique mentions of a publication on a blog source	Positive Integer Including 0	
Category	Fits a publication's topic into one of the following areas: Agricultural and Veterinary Sciences, Biological Sciences, Chemical Sciences, Earth Sciences, Engineering, Environmental Sciences, History and Archaeology, Information and Computing Science, Medical and Health Sciences, Physical Sciences, Psychology and Cognitive Sciences, Studies in Human Society, Being human, Current events, Medical matters, Offbeat, Real-life science fiction and The world we live in.	String	
Checked For Gaming/Incorrect Mentions	Describes if the altmetric score and results for a publication has been checked to not be affected by gaming and incorrect mentions	String (Yes or No)	
CiteULike Mentions	Count of unique mentions of a publication on the CiteULike service	Positive Integer including 0	
Content Type	Fits a publication's type into one of the following areas: Analysis, Article, Brief report, Correspondence, Early Release Article, Editorial, For Debate, Letter, Original Investigation, Perspective, Policy Forum, Pre-print, Recommendations statement, Report, Review, Short Article, Short Communication, Special Communication and Special Report.	String	
Country	The country of the affiliated Institution	String	NULL or string value with ';' delimiter
Description	Describes what the publication is about	String	

Attribute Name	Definition	Format	Format exceptions
Dimensions ID	ID reference for a publication in the dimensions database	String(pub.NNNNNNNNNN)	NULL or string value
DOI	Digital object identifier for a publication	String	NULL or string value
F1000 Mentions	Count of unique mentions of a publication on the f1000 research	Positive Integer Including 0	
FB Mentions	Count of unique mentions of a publication on Face Book in posts on a set or curated Facebook pages	Positive Integer including 0	
First Name	First name of one author of the publication	String	
Google+ Mentions	Count of unique mentions of a publication on Google Plus	Positive Integer including 0	
Grid City	City of the affiliation institution	String	NULL or string value
Grid Country	Country of the affiliated institution	String	NULL or string value
Grid Name	Name of the affiliated Institution	String	
Grid State	State of the affiliated Institution	String	NULL or string value
Had Corrections	UNKNOWN		
Handle.net IDS	ID reference for a publication on Handel.Net Corporation for National Research Initiatives persistent identifier for information resources	String	NULL or string value
Had Affiliation Data	Describes if the publication has data on its affiliated institutions	String (Yes or No)	
Institution	Another name for Grid Name	String	
Journal	Describes the Journal to which the publication is associated with	String	
Journal ISSNs	ISSN of the journal. International Standard Serial Number: A journal can have more than one ISSN - for example, one for print and one for on-line publication. Some publications do not have an ISSN.	String	NULL or string value
Journal Rank	UNKNOWN		
Last Name	Last name of an author of the publication	String	
Link	A link to where the publication can be accessed online	String	NULL or string value
LinkedIn Mentions	Count of unique mentions of a publication on LinkedIn	Positive Integer including 0	
Mendeley Mentions	Count of unique mentions of a publication on Mendeley	Positive Integer including 0	
MSM Mentions	Count of unique mentions of a publication on MSM	Positive Integer including 0	
News Mentions	Count of unique mentions of a publication from news sources	Positive Integer including 0	
News Tweet Ratio	Ratio of news and twitter count	Float	
OA	Describes if a publication is open access or not	String (Yes or No)	
Patent Mentions	Count of unique mentions of a publication in patents across nine jurisdictions	Positive Integer Including 0	
Peer Review Mentions	Count of unique mentions of a publication from peer review comments for any paper with a DOI, PubMed ID, or ArXiv ID	Positive Integer including 0	
Pinterest Mentions	Count of unique mentions of a publication on Pinterest	Positive Integer including 0	
PubMed ID	ID reference for a publication on Pub Med	String	NULL or string value
Policy Mentions	Count of unique mentions of a publication on sources of policy	Positive Integer including 0	

Attribute Name	Definition	Format	Format exceptions
Published	Describes when the publication was published	String (DD/MM/YY)	String or numeric value
PubMedCentral ID	ID reference for a publication on PubMedCentral	String	NULL or string value
Q&A Mentions	Count of unique mentions of a publication on Q&A sources	Positive Integer including 0	
Reddit Mentions	Count of unique mentions of a publication on Reddit	Positive Integer including 0	
Region	Describes region in the world where the affiliated institution/s are	String	
RepPEc ID	ID reference for a publication on RepPEc Research Papers In Economics permanent identifier attributed to a person	String	NULL or string value
Research Hi Mentions	UNKNOWN		
Sharedit	A link to the publication from sharedit	String	NULL or string value
SSRN	SSRN ID for a publication Social Science Research Network.	String	NULL or string value
Syllabi Mentions	Count of unique mentions of a publication on Syllabi	Positive Integer including 0	
Title	The title of a publication	String	
Twitter Mentions	Count of unique mentions of a publication on Twitter Number of registered users that tweet or retweet a post that links to a trackable scholarly product	Positive Integer including 0	
Type	Describes the type of the affiliated institution into: Company, Education, Facility, Government, Healthcare, Nonprofit, Other, Private	String	
URN	URN ID for a publication	String	NULL or string value
Video Mentions	Count of unique mentions of a publication on video sources in description of YouTube videos	Positive Integer including 0	
Weibo Mentions	Count of unique mentions of a publication on Weibo (No longer done, but historical data kept)	Positive Integer including 0	
Wikipedia Mentions	Count of unique mentions of a publication on Wikipedia	Positive Integer including 0	
Dimension Citation Mentions	Count of unique citations of a publication on Dimensions	Positive Integer including 0	

Table 4: Altmetric Attribute Characteristics

Attribute Name	Is Compete?	Is Timelines?	Is Valid?	Is Consistent?	Is Unique?
ADS Bibcode	No, has NULLs	Yes	Yes	No, has NULLs	No
Affiliation	No, has NULLs	Yes	Yes	No, has NULLs	No
Altmetric Attention Score	Yes	Yes	Yes	Yes	No
Altmetric ID	Yes	Yes	Yes	Yes	Yes
ArXiv ID	No, has NULLs	Yes	Yes	No, has NULLs	Yes
Authors	No, has NULLs	Yes	Yes	No, has NULLs	No
Blog Mentions	Yes	Yes	Yes	Yes	No
Category	Yes	No, some factors are not specific	Yes	Yes	No
Checked for Gaming/Incorrect Mentions	Yes	Yes	Yes	Yes	No
CiteULike Mentions	Yes	Yes	Yes	Yes	No
Content Type	Yes	Yes	Yes	Yes	Yes
Country	Yes	Yes	Yes	Yes	Yes
Description	Yes	Yes	Yes	Yes	No
Dimensions ID	No, has NULLs	Yes	Yes	No, has NULLs	Yes
DOI	No, has NULLs	Yes	Yes	No, has NULLs	Yes
F1000 Mentions	Yes	Yes	Yes	Yes	No
FB Mentions	Yes	Yes	Yes	Yes	No
First Name	Yes	Yes	Yes	Yes	No
Google+ Mentions	Yes	Yes	Yes	Yes	No
Grid City	No, has NULLs	Yes	Yes	No, has NULLs	No
Grid Country	No, has NULLs	Yes	Yes	No, has NULLs	No
Grid Name	No, has NULLs	Yes	Yes	No, has NULLs	No
Grid State	No, has NULLs	Yes	Yes	No, has NULLs	No
Had Corrections	Yes	No	Yes	Yes	No
Handle.net IDS	No, has NULLs	Yes	Yes	No, has NULLs	Yes
Had Affiliation Data	Yes	Yes	Yes	Yes	No
Institution	No, has NULLS	Yes	Yes	No, has NULLS	No
Journal	Yes	Yes	Yes	Yes	No
Journal ISSNs	No, has NULLs	Yes	Yes	No, has NULLs	Yes
Journal Rank	Yes	No	Yes	Yes	No

Attribute Name	Is Compete?	Is Timelines?	Is Valid?	Is Consistent?	Is Unique?
Last Name	Yes	Yes	Yes	Yes	No
Link	Yes	Yes	Yes	Yes	Yes
LinkedIn Mentions	Yes	Yes	Yes	Yes	No
Mendeley Mentions	Yes	Yes	Yes	Yes	No
MSM Mentions	Yes	Yes	Yes	Yes	No
News Mentions	Yes	Yes	Yes	Yes	No
News Tweet Ratio	Yes	No	Yes	Yes	No
OA	Yes	Yes	Yes	Yes	No
Patent Mentions	Yes	Yes	Yes	Yes	No
Peer Review Mentions	Yes	Yes	Yes	Yes	No
Pinterest Mentions	Yes	Yes	Yes	Yes	No
PubMed ID	No, has NULLs	Yes	Yes	No, has NULLs	Yes
Policy Mentions	Yes	Yes	Yes	Yes	Yes
Published	Yes	Yes	Yes	No, has numbers	No
PubMedCentral ID	No, has NULLs	Yes	Yes	No, has NULLs	Yes
Q&A Mentions	Yes	Yes	Yes	Yes	No
Reddit Mentions	Yes	Yes	Yes	Yes	No
Region	Yes	Yes	Yes	Yes	No
RepPEc ID	No, has NULLs	Yes	Yes	No, has NULLs	No
Research Hi Mentions	Yes	No	Yes	Yes	No
Sharedit	No, has NULLs	Yes	Yes	No, has NULLs	No
SSRN	No, has NULLs	Yes	Yes	No, has NULLs	No
Syllabi Mentions	Yes	Yes	Yes	Yes	No
Title	Yes	Yes	Yes	Yes	No
Twitter Mentions	Yes	Yes	Yes	Yes	No
Type	Yes	No, some factors are not specific	Yes	Yes	No
URN	No, has NULLs	Yes	Yes	No, has NULLs	No
Video Mentions	Yes	Yes	Yes	Yes	No
Weibo Mentions	Yes	Yes	Yes	Yes	No
Wikipedia Mentions	Yes	Yes	Yes	Yes	No
Dimension Citation Mentions	Yes	Yes	Yes	Yes	No

Data Quality (Bornmann and Haunschild):

- Data is consistent across all fields
- Provenance is known
- Sources (by organisation) are well established and known
- Authors are well established in the field of scientometrics
- Dependence on authors sourcing processes

Data is deemed to be of high

Table 5: Bornmann and Haunschild Data Dictionary

Dataset	Owner	Variable Name	Definition	Format	Uniqueness
Supporting data for the Paper "Do altmetrics correlate with the quality of papers? A large-scale empirical study based on F1000Prime data"	Lutz Bornmann and Robin Haunschild (compiled from various, but trusted, sources)	pubyear	Year of publication for paper: Papers from 2011-2013 in F1000; Papers with a DOI excluded. Papers not referenced by Altmetrics excluded.	Integer (yyyy)	No
		wos_cits_3	Lagged (over 3 years) and smoothed Web of Science citation counts	Integer	No
		wos_cits	Web of Science citation count for paper	Integer	No
		sco_cits_3	Lagged (over 3 years) and smoothed Scopus citation counts	Integer	No
		sco_cits	Scopus citation count for paper	Integer	No
		tweets	Altmetrics tweet counts for paper	Integer	No
		me_readers	Altmetrics Mendeley counts for paper	Integer	No
		altmetric_score	Altmetrics score for paper	Float	No
		citescore	CiteScore citation score	Float	No
		item_ijif	Clarivate Analytics Journal Impact Factor	Float	No
		total_f1000_score	Sum of F1000 evaluation scores	Integer	No

Data Quality (Plum Analytics):

- Data is consistent across all fields
- Provenance is known
- Data dictionary and additional notes on dataset were provided
- Sources (by organisation) are well established and known
- Generating organisation (Plum Analytics) are well established and recognised
- Dependence on Plum Analytics sourcing processes

Data is deemed to be of high quality

Table 6: Plum Analytics Data Dictionary

Dataset	Owner	Variable Name	Definition	Format	Uniqueness
The PlumX Altmetrics & Sci-Hub Dataset (Top 100, 2016)	Plum Analytics	DOI	DOI is the unique identifier assigned to each paper on Sci-Hub	String	Yes
		Title	Title of the research paper	String	Yes
		Sci-Hub Downloads	These values were calculated by summing the number of rows for each DOI in the Sci-Hub dataset. The top 100 download counts were then identified and added to https://plu.mx/scihub100.	Integer	No
		Publisher Cost	This is the cost to download a PDF from the publisher's site (as of May 10, 2016).	Float	No
		SciHub x Pub Cost	The publisher cost was multiplied with the Sci-Hub Downloads count to estimate the dollar amount that these downloads may be costing publishers.	Float	No
		Subject Area	Each DOI was investigated and sorted into a general subject area by Noella Natalino, MLIS, Product/Content Manager for Plum Analytics.	String	No
		ISBN	International Standard Book Number	Integer (empty if not assigned/known)	Yes
		PMID	PubMed identifier is a unique integer value (PubMed is an index). 72 of the top 100 DOIs are available via PubMed.	Integer (empty if not assigned/known)	Yes
		Year	The publication year of the research paper	Integer (yyyy)	No
		Type	The type of research paper: Article (62), Review (25), Letter (6), Book Chapter (4), Conference Paper (2), Book (1)	String	No
		Captures	Captures track when end users bookmark, favorite, or save an item for future use. Downloaded May 10, 2016. Visit https://plu.mx/scihub100 for the most recent counts.	Integer	No
		Social Media	Social media metrics are the +1s, likes, shares, and tweets about research. Downloaded May 10, 2016. Visit https://plu.mx/scihub100 for the most recent counts.	Integer	No
		Mentions	Mentions are the blog posts, comments, reviews, and wikipedia links about research. Downloaded May 10, 2016. Visit https://plu.mx/scihub100 for the most recent counts.	Integer	No
		Usage	Usage metrics are the # of clicks, downloads, views, and library holdings for research. Downloaded May 10, 2016. Visit https://plu.mx/scihub100 for the most recent counts.	Integer	No

9. Appendix 3: Exploratory Data Visualisations

9.1 Altmetric 2013 – 2019 Visuals

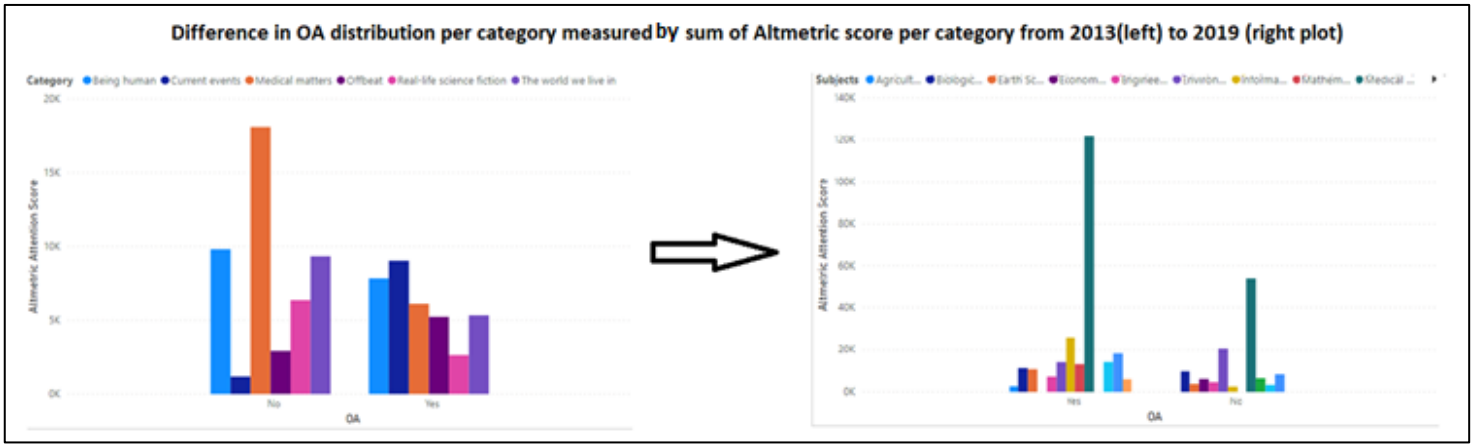


Figure 8: Difference in Open Access A distribution per Category as Measured by Sum of Altmetric Score per Category from 2013 to 2019

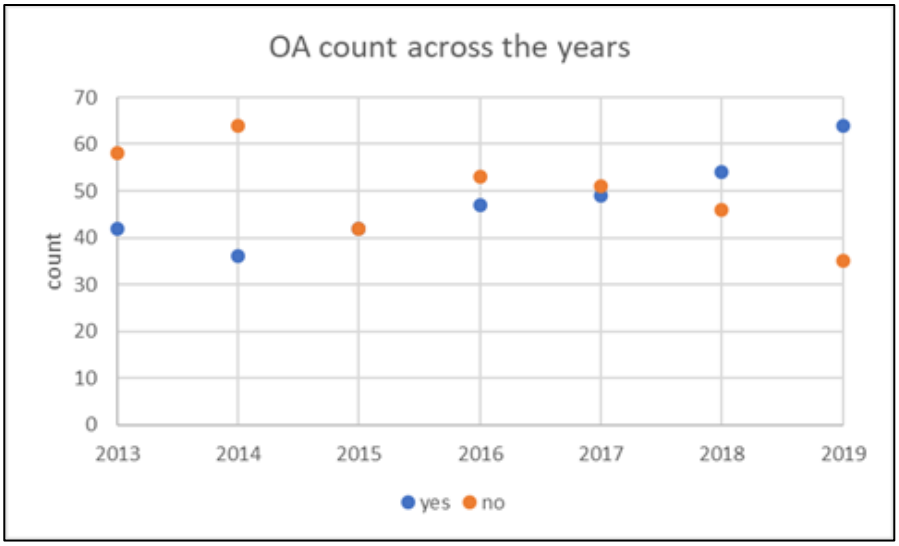


Figure 9: Change in Counts of Open and Non-open Access Type in the Top 100 Publications, 2013 to 2019

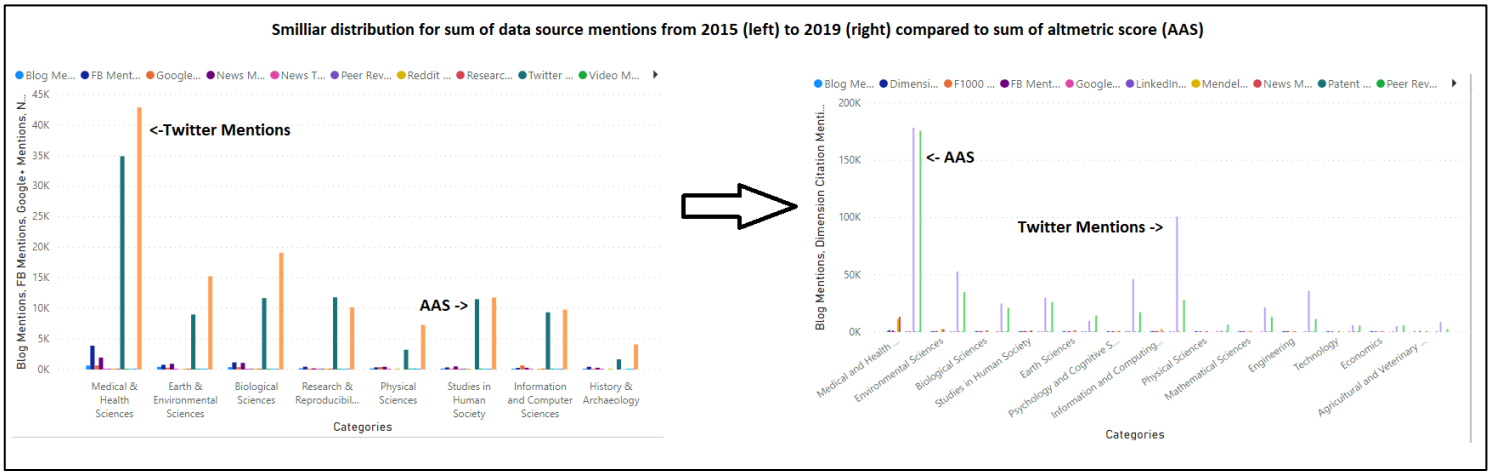


Figure 10: Distribution in the Sum of Altmetric Data Sources and AAS per Category from 2015 to 2019

	AAS	Blog	F1000	FB	Google+	LinkedIn	Mendeley	News	PeerReview	Pinterest	Policy	Q&A	Reddit	Syllabi	Twitter	Video	Weibo	Wikipedia	Dimension	Patent
2013	83391	1149	24	6493	1513	0	5043	2185	0	13	0	1	245	0	63697	0	0	0	0	0
2015	120458	2172	0	7685	2765	0	0	5674	72	0	0	0	293	0	93156	26	349	261	0	0
2016	225848	2233	34	4869	2027	0	0	0	39	0	16	20	266	0	100814	54	0	185	0	0
2017	268729	2037	29	6141	879	0	17689	26096	13	0	6	4	259	0	161685	167	0	206	0	0
2018	295916	1971	33	2586	751	0	20202	20445	0	0	12	2	430	0	319750	106	0	154	1639	0
2019	362936	1801	52	0	92	0	21780	22985	0	0	23	5	474	0	521467	216	0	142	2469	0

Original mention counts

⇩

Blog	F1000	FB	Google+	LinkedIn	Mendeley	News	PeerReview	Pinterest	Policy	Q&A	Reddit	Syllabi	Twitter	Video	Weibo	Wikipedia	Dimension	Patent
5	1	0.25	1	0.5	0	8	1	0.25	3	0.25	0.25	1	1	0.25	1	3	0	3

Apply mention weighting

⇩

	AAS	Blog	F1000	FB	Google+	LinkedIn	Mendeley	News	PeerReview	Pinterest	Policy	Q&A	Reddit	Syllabi	Twitter	Video	Weibo	Wikipedia	Dimension	Patent
2013	83391	5745	24	1623.25	1513	0	0	17480	0	3.25	0	0.25	61.25	0	63697	0	0	0	0	0
2015	120458	10860	0	1921.25	2765	0	0	45392	72	0	0	0	73.25	0	93156	6.5	349	783	0	0
2016	225848	11165	34	1217.25	2027	0	0	0	39	0	48	5	66.5	0	100814	13.5	0	555	0	0
2017	268729	10185	29	1535.25	879	0	0	208768	13	0	18	1	64.75	0	161685	41.75	0	618	0	0
2018	295916	9855	33	646.5	751	0	0	163560	0	0	36	0.5	107.5	0	319750	26.5	0	462	0	0
2019	362936	9005	52	0	92	0	0	183880	0	0	69	1.25	118.5	0	521467	54	0	426	0	0

New mention counts

Figure 11: Excel Tables of Altmetric Data Source Counts Before (first table) and After (third table) Weightings (second table) are Applied

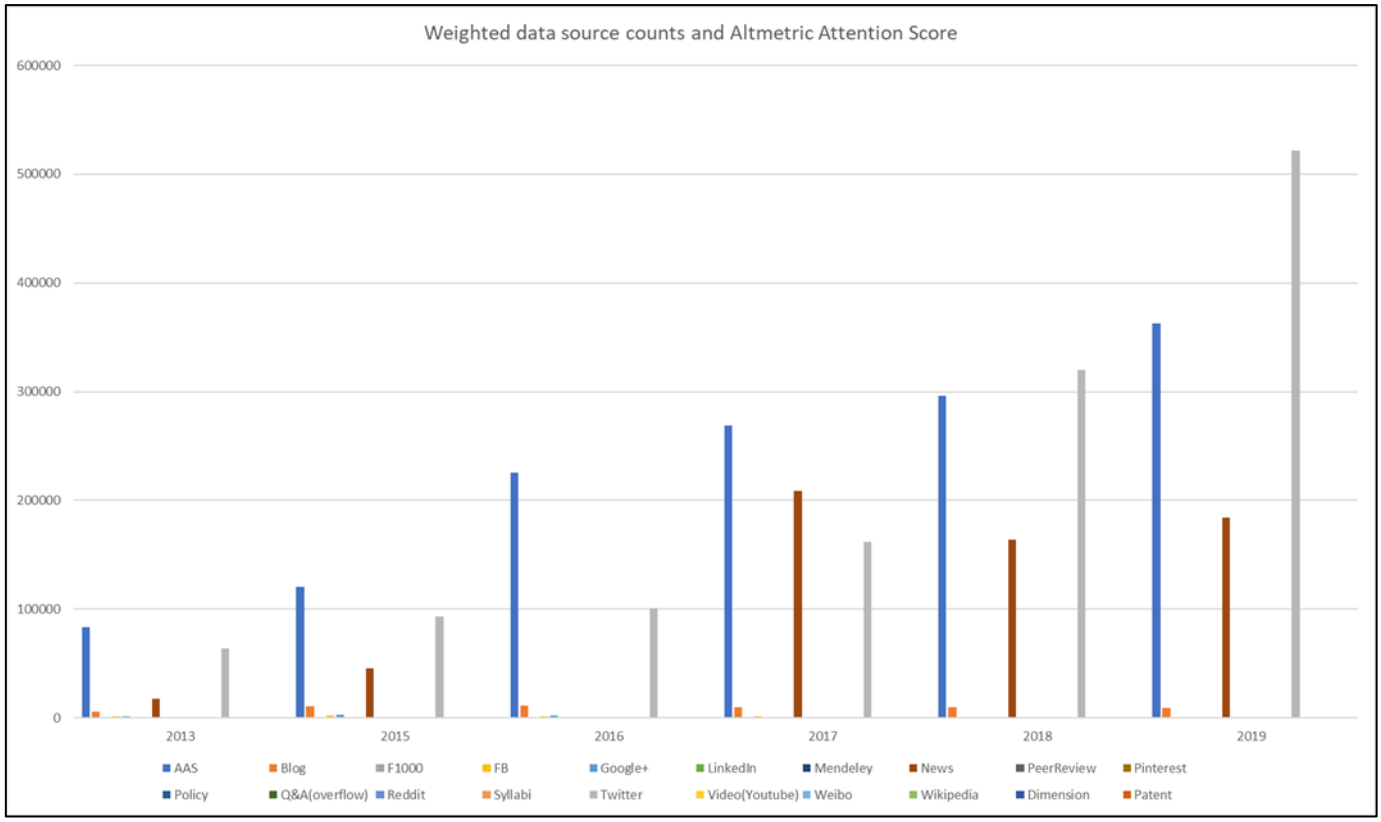


Figure 12: Graph of the Third Table in Figure 7, Comparing Sum AAS to the New Sum of Data Source Counts per Year

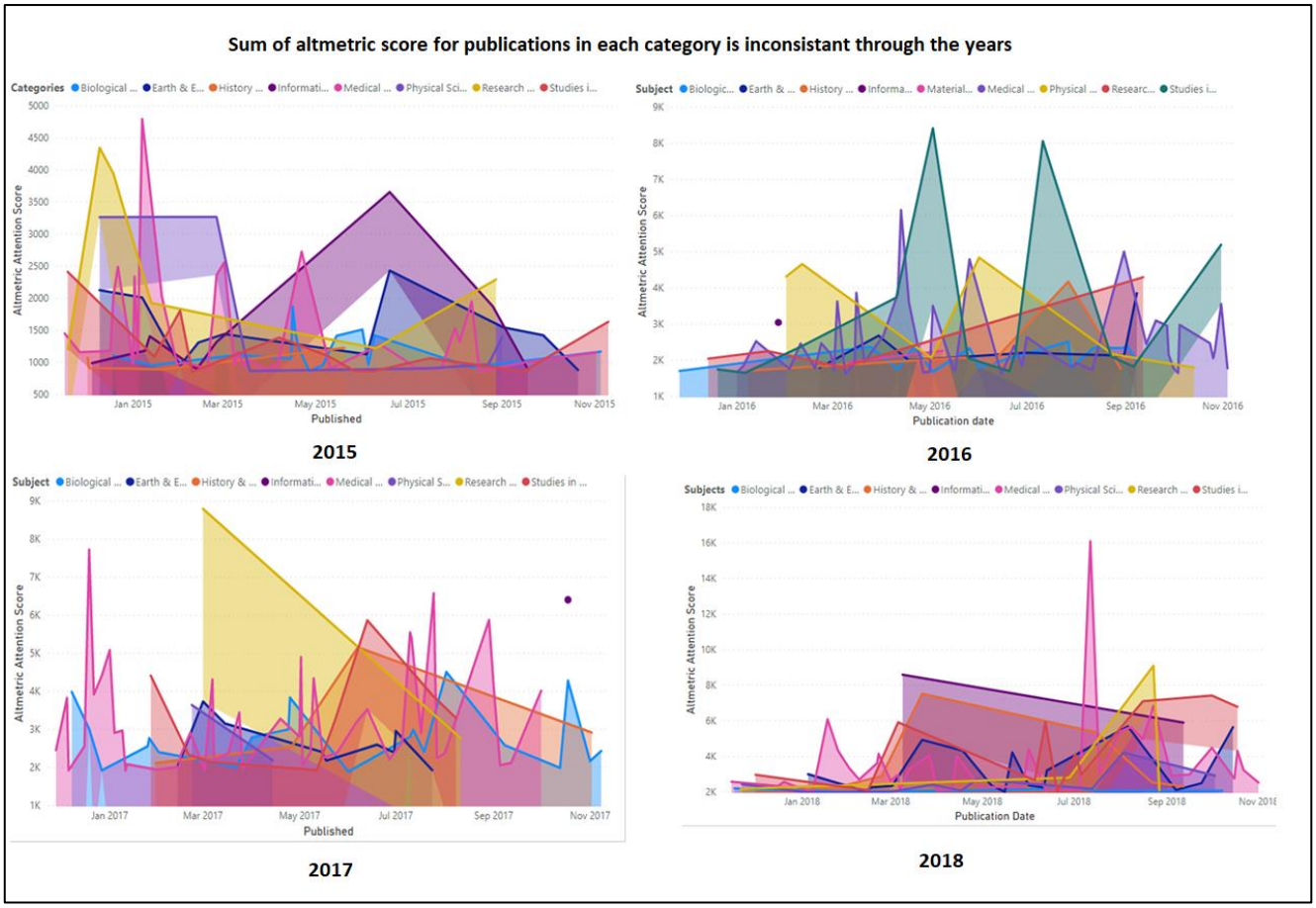


Figure 13: Distribution in AAS per Category Throughout the Year, for Four Consecutive Years 2015 to 2018

9.2 Plum Analytics Visuals

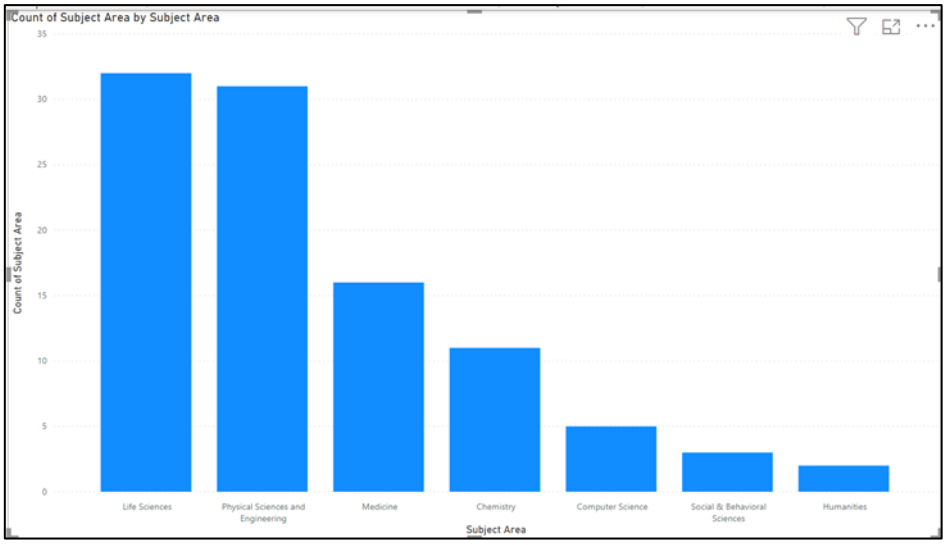


Figure 14: Count of Subject Area by Subject Area

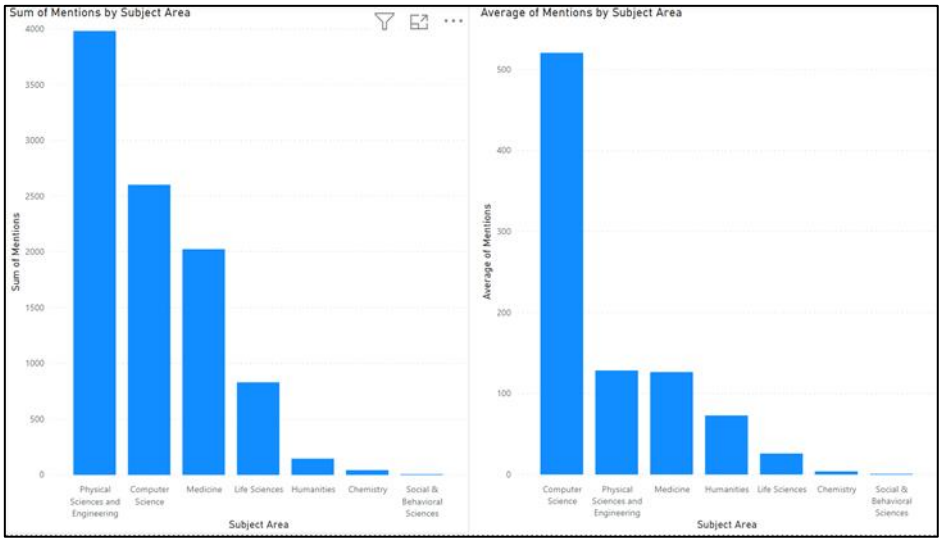


Figure 15: Sum (left) and Average (right) of Mentions by Subject Area

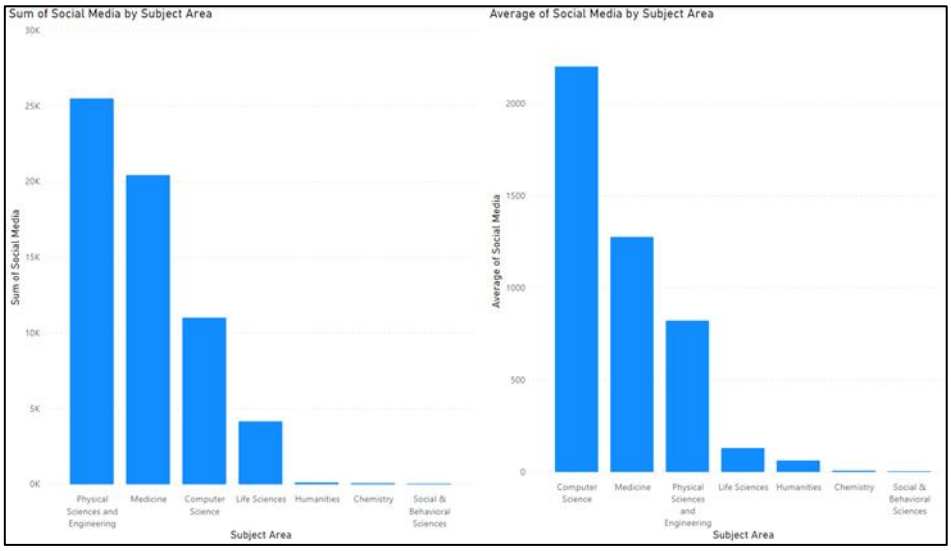


Figure 16: Sum (left) and Average (right) of Social Media by Subject Area

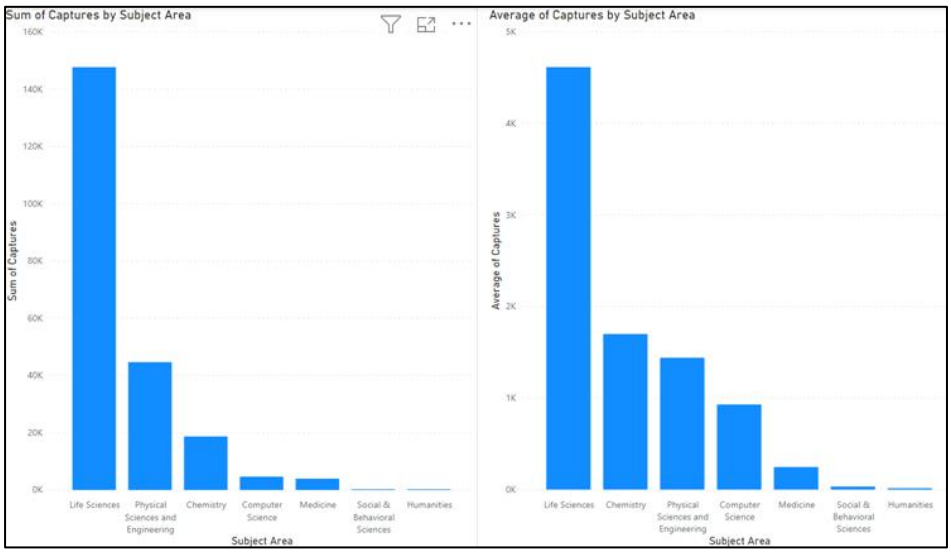


Figure 17: Sum (left) and Average (right) of Captures by Subject Area

9.3 Bormann and Haunschild Visuals

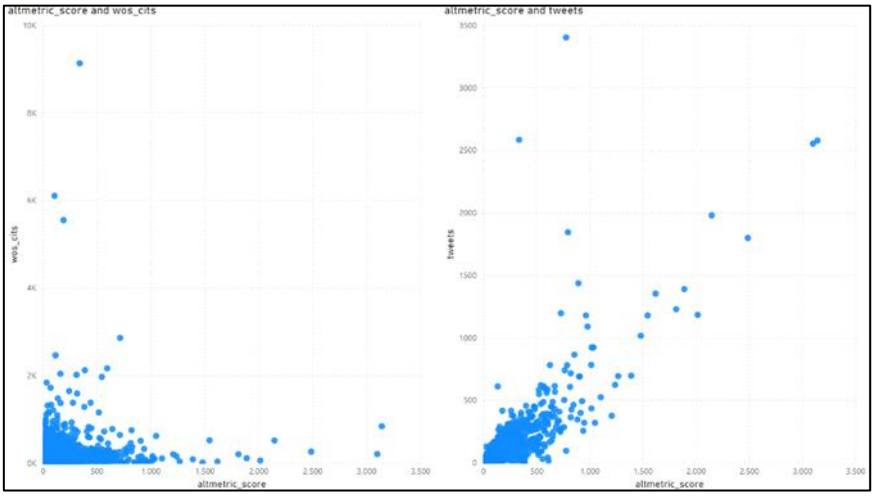


Figure 18: Scatter plot of Web of Science Citations (left) and Tweets (right) vs Altmetric Score

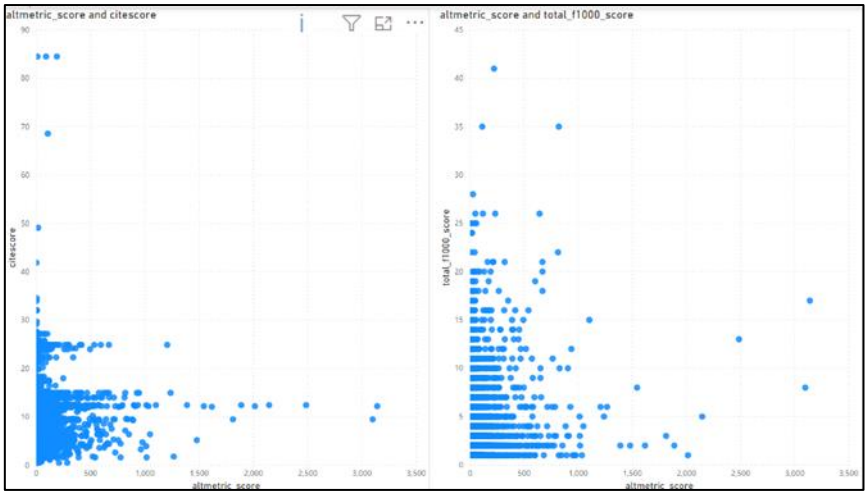


Figure 19: Scatter plot of Citescore (left) and F1000 Score (right) vs Altmetric Score

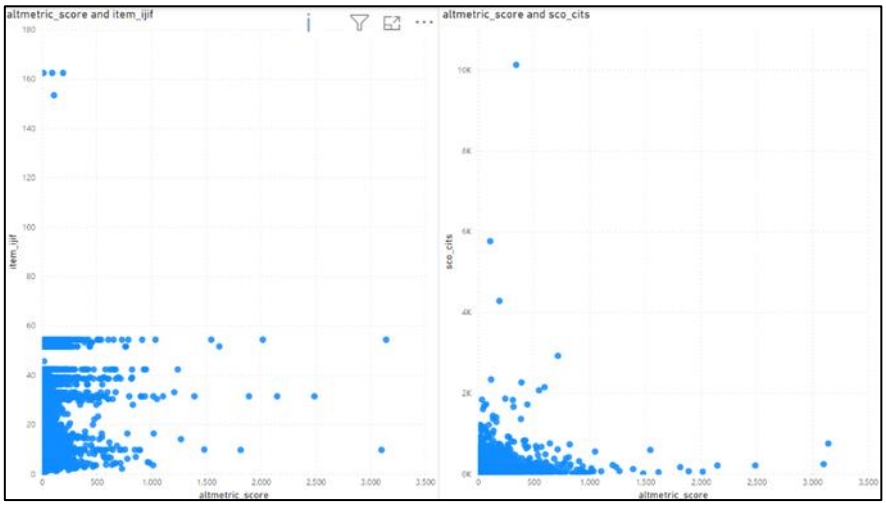


Figure 20: Scatter plot of Item (left) and Scopus Citations (right) vs Altmetric Score

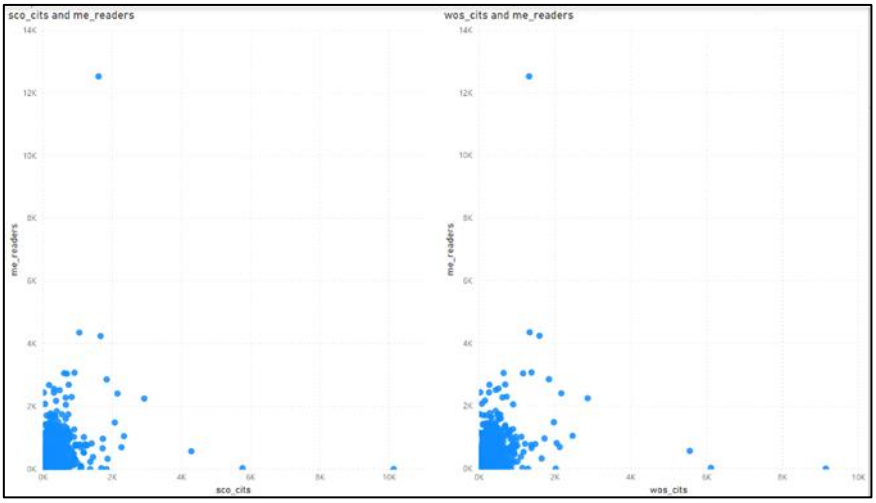


Figure 17: Scatter plot of Mendeley Readers vs Scopus Citations (left) and Mendeley Readers vs WoS Citations (right)

10. Appendix 4: Contribution Table

Contribution Table	**Stephen**	**Jinxi**	**Paul**
Project Process	x	x	
Project Problem	x	x	x
Project Data Assets	Altmetric, Plum Analytics and Bornmann Data	Altmetric Data	
Exploratory Analysis	Altmetric, Plum Analytics and Bornmann Data	Altmetric Data	
Analytical Plan	6.1 , 6.2, 6.3	6.1	
Appendix 1	x	x	x
Appendix 2	8.1, 8.2, 8.3, 8.4, 8.6, 8.7, 8.8	8.1, 8.4, 8.5, 8.8	
Appendix 3	9.2, 9.3	9.1	

11. Appendix 5: Jinxi Luo

Data Understanding

Below is a table containing the attribute names for each data set by year into the four categories. It can be noted that attribute names throughout the years are similar in meaning with the difference of capitalization and different usage of words; this is checked by noticing their values across years. This does not have much effect on analysis except it may be confusing at times so in the next section a convention is followed to correct for this. (Note that some attributes may not match their exact names in the data set due to Word's auto correct)

To keep things brief only data of relevance that have been found to not be of high quality is mentioned; determined based on the dimensions of data quality. Beginning in the Altmetric Info section it contains mostly attributes that describe the Altmetric attention score and the count of attention received from Altmetric data sources. It was noticed that one column of altmetric attention score are of float values, this will need to be corrected to be a whole number rounded up as otherwise the data would be invalid. URL Links to the altmetric donut and score are also included for a given publication(Small/medium image and badge URL), this information is not useful for our project as it provides no additional information and it is not feasible to access each link individually to check the contents.

It should be noted that the context to the following attributes in Altmetric Info are unknown, they are all categorial attributes with factor levels of Yes/No. More research may be done to determine their meaning; however, they are not relevant to our research so far.

- Checked?
- Stayin in?
- Had Corrections?

The Unique Identifiers column contains ID attributes that are used to track a publication back to a database containing a listing of the publication. As they are IDs, they are also unique and may be used as keys to connect data sets, it was found that the following connections exist when connecting affiliation data sets to their respective years:

- (2015aff) altmetric_id references (2015) id
- (2016aff) DOI references (2016) DOI
- (2017aff) DOI references (2017) DOI
- (2017aff) arXivID references (2017) arXvID

Besides from this usage, ID attributes has no feasible purpose as it is not possible for us to reference each id back to their data bases for error checking, obtaining extra information or to connect data sets across different years. However, the information of a publication having an ID from a data base is useful, as authors that purse to get their publication on different data bases will have more chance to distribute their publication to a wider audience. So, it may also be useful when performing analysis such as binary regression that we only consider if a publication has or has not a particular ID.

This decision reflects the fact that a lot of ID columns have empty values, it is not possible to fill in these values with an appropriate ID as it may not exist and is very difficult to find for all. It is also hard to check if all IDs are correct, while some have fixed digit lengths others are mix of numbers and letters that have different formats depending on a publication.

Table 7

Year	Altmetric Info	Unique Identifiers	Publication Info	Affiliation Info
2013	(Altmetric Attention Score) (Reddit threads)(Bloggers)(Tweeters)(Google+ authors)(F1000reviews)(Pinterest posts)(News outlets) (Q&A site users) (Facebook walls) (Mendeley readers) (CiteULike readers) (Small image) (Medium image)	(DOI) (PMID) (ArXiv ID) (Journal ISSNs)	(Category 1) (Title) (Journal) (URL) (OA)	
2014		(DOI) (Altmetric ID)	(Title) (Journal) (URL) (Published) (Author name) (Access type) (Category)	(Institution) (Country) (Region)
2015 Affiliation		(altmetric_id)		(Name) (Lat) (Lng) (Country) (type)
2015	(Score) (News_tweet_ratio) (Count news) (Count blog) (Count twitter) (Count facebook) (Count peer review) (Count weibo) (Count google plus) (Count reddit) (Count research hi) (Count video) (Count wikipedia)	(Doi) (id)	(Title) (Journal) (Date) (Authors) (Rank) (Link) (Journal rank) (Categories) (Open access)	(Countries) (Institutions)
2016 Affiliation		(DOI)	(name) (confidence) (search_type)	(Affiliation) (Grid_id) (Grid_name) (Grid_city) (Grid_state) (Grid_country)
2016	(Score) (Checked for gaming/incorrect mentions) (Checked?) (Stayin in?) (Had corrections?) (Total_posts_msm) (total_posts_blog) (total_posts_policy) (total_posts_tweet) (total_posts_peer_review) (total_posts_wikipedia) (total_posts_weibo) (total_posts_fbwall) (total_posts_gplus) (total_posts_linkedin) (total_posts_rdt) (total_posts_pinterest) (total_posts_f1000) (total_posts_video) (total_posts_qna)	(DOI) (Altmetric ID)	(Position) (title) (Journal) (Publication Date) (link) (sharedit) (OA) (Content Type) (subject) (Blurb)	(Has affiliations?)
2017 Affiliation		(Dimensions ID) (DOI) (ArXiv ID) (PubMedID)	(Last name) (First name)	(Affiliation) (Country)
2017	(Score) (Number of news stories) (Number of blog posts) (Number of policy documents) (Number of tweets) (Number of peer reviews) (Number of weibo posts) (Number of Facebook posts) (Number of Wikipedia pages) (Number of Google+ posts) (Number of Linkedin posts) (Number of Reddit posts) (Number of pins) (Number of F1000 posts) (Number of Q&A posts) (Number of videos) (Number of syllabi) (Number of Mendeley readers)	(DOI) (PubMed ID) (ArXiv ID) (Altmetric ID)	(Title) (Journal) (Date) (Subject) (Description) (OA)	
2018	(News mentions) (Blog mentions) (Policy mentions) (Twitter mentions) (Patent mentions) (Peer review mentions) (Weibo mentions) (Facebook mentions) (Wikipedia mentions) (Google+ mentions) (LinkedIn mentions) (Reddit mentions) (Pinterest mentions) (F1000 mentions) (Q&A mentions) (Video mentions) (Syllabi mentions) (Number of Mendeley readers) (Number of Dimensions citations) (Altmetric Attention Score) (Badge URL)	(Journal ISSNs) (Handel.net IDs) (ADS Bibcode) (ArXiv ID) (RePEc ID) (SSRN) (URN) (PubMed ID) (PubMedCentral ID) (DOI)	(Rank) (Title) (Description) (Journal/Collection Title) (Publisher) (Subjects) (Subjects (FoR)) (Publication Date) (DOI URL) (Details Page URL) (Open_Access) (Authors) (Publisher-preferred links)	(Affiliations (GRID)) (Funders) (Unique countries)
2019	(Altmetric Attention Score) (News mentions) (Blog mentions) (Policy mentions) (Twitter mentions) (Patent mentions) (Peer review mentions) (Weibo mentions) (Facebook mentions) (Wikipedia mentions) (Google+ mentions) (LinkedIn mentions) (Reddit mentions) (Pinterest mentions) (F1000 mentions) (Q&A mentions) (Video mentions) (Syllabi mentions) (Number of Mendeley readers) (Number of Dimensions citation)	(Journal ISSNs) (Handel.net IDs) (ADS Bibcode) (ArXiv ID) (RePEc ID) (SSRN) (URN) (PubMed ID) (PubMedCentral ID) (DOI)	(Title) (Journal/Collection Title) (Open_Access) (Description) (Publication Date) (DOI URL) (Details Page URL) (Publisher) (Subjects) (Authors) (sharedit) (Publisher-preferred)	(Affiliations (GRID))

The Publication Info section contains information on what a publication is about, when it was published, author names, what journal it belongs to, whether it is open access and links to the publication. Main attributes that will be of interest are the category on what a publication is about, journal It belongs to, when it was published and if it is open access.

The category attributes mainly had no data problems except in 2013 where factors levels of the category were not very specific and different in levels from those in the following years. There were also a few values missing. As the attribute is of interest, I manually determined the category for publications with missing values to my best judgment. To fix the nonspecific factor level problem, I decided to leave them as it is and combine the factors with those in the following years; as to reduce the amount of bias I am adding to the datasets and analysis.

There were not many problems with publication date except for inconsistent attribute names and format. Most columns only needed their data format to be changed in excel to one standard representation, one column had specific 00:00:00 time appended to the data that needed to be removed as that information was incoherent and added no further information. Only problem was with the 2019 publishing times that were listed to be 5-digit numbers. I am unsure of how that may be decoded to the standard date format. So, the column may not be of use; it is noted that it could be a representation of time in terms of order each publication was published based on numerical order.

Open access information can be generalized to being Yes or No for a publication, factor levels that had similar meaning can be grouped into such. There was one column that only contained values detailing if a publication is open access, it was decided that empty values will be treated to be not open access. In the analysis section we see that this decision was reasonable as it matches distributions in the following years. As otherwise the attribute for that year could not be used as it was not possible to replace a lot of the empty values in a reasonable and feasible way; by randomly adding Yes/No or checking information on links to the publication.

The remaining attributes are then not of much interest in this stage with some that may be considered in assignment 2. As a result, data quality is not considered and any missing values that exists in those columns are replaced with NULL to represent explicitly that the data does not exist.

- Title, Author name, Blurb, Description
  - These attributes contain mostly text information that may be analysed in assignment 2 with association rules or others. As they cannot be summarized through exploratory data analysis (visualizations and summary statistics)in this stage.
- URL, Link, sharedit, DOI URL, Details page URL, Publisher preferred links
  - Links are not useful to the analysis, only to fact check/repair data set
- Confidence, search_type
  - Confidences is a float value, and search type is a string. The specific context beyond these attributes are unknown so they are not used unless further knowledge is known about them.
- Rank/position
  - Has not much purpose other than numbering publications as they are listed by default.
  - This information, however, was helpful to see that 2018 had more than 100 publications. The excess was removed to reduce inconsistencies during analysis and because the entries were missing values for most of the attributes of interest.

The Affiliation Info section contains attributes that describe affiliated institutions for a publication, such information includes, name of the institution, geographic information of where that institution is and the type of institution. Attributes of interests are location information and type of institution. Specifically, Country for location, as it is the most complete attribute for all years, and I am more interested in specific locations instead of in more general such as regional, grid id, latitude and longitude information. Additionally, only 2016 provides specific city and state locations which are attributes where about half of the entries are empty and so would not be very significant for analysis.

Country attributes for most years are of good data quality, however there are some empty values, as it is not feasible to manually enter the appropriate institution, a NULL used. This should not affect analysis much as only 2015 affiliations data set had this issue and only about 30% of the entries were empty.

Type attribute also had some empty values and there was only a few, however, for this context it was not possible to judge without research on whether an institution belonged into a factor level. So, a new level was created to be "Private" to replace the missing values. It was possible to delete affiliations with missing values as it will not affect the number of publications, this was not done as those missing values may be crucial to analysis and the missing values may be due to an unnamed category.

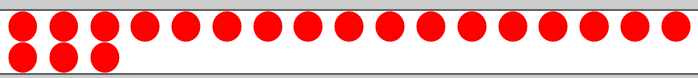
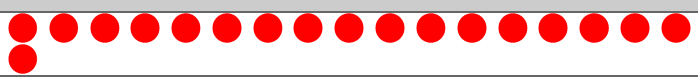
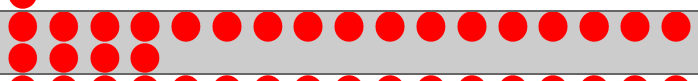
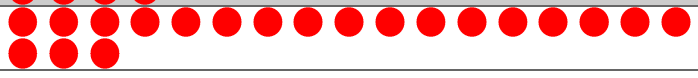
The remaining attributes are then not of much interest in this stage with some that may be considered in assignment 2. As a result, data quality is not considered and any missing values that exists in those columns are replaced with NULL to represent explicitly that the data does not exist.

- Institution/affiliation/affiliations GRID/Funders/name/Funders
  - These attributes contain mostly text information that may be analysed in assignment 2 with association rules or others. As they cannot be summarized through exploratory data analysis (visualizations and summary statistics)in this stage.
- Has affiliations?
  - Has not much purpose as it is all Yes.

These are the major data Issues that I have found that can be grouped into the data understanding stage, In Preparation and Analysis section, more context and further data quality assessments are made. From data understanding, the decision is to make less modifications to the data as possible that would result in a loss of information. And only modify to improve readably and analysis, this is done so that the data set can be used for group discussions more generally, analysis is less affected by biases that may have been introduced and in a business perspective it becomes less waste of resources.

Below is a more condensed version of table 1 which shows the counts of attributes for each column. The available attributes throughout the years differs enough for it to be hard to make connecting points across all years. Note the difference is more significant when considering what each attribute describes. So, analysis is done separately for each year and common attributes and patterns from each year are used to build on each other.

Table 8

Year	Altmetric Info	Unique Identifiers	Publication Info	Affiliation Info
2013				
2014				
2015 Affiliation				
2015				
2016 Affiliation				
2016				
2017 Affiliation				
2017				
2018				
2019				

Preparation and Analysis

First change made to the data is to standardize attribute names. A set of conventions to naming attributes is followed as such that underscores are converted to space, first letter is capitalized and the whole word is capitalized if it is an acronym. Below list more specific changes for each data set. ('->' means 'changed to')

2013

- Bloggers -> Blog Mentions
- F1000 reviews -> F1000 Mentions
- Q&A site users -> Q&A Mentions
- CiteULike readers -> CiteULike Mentions
- Category 1 -> Category
- Tweeters -> Tweet Mentions
- Pinterest posts -> Pinterest Mentions
- Facebook walls -> FB Mentions
- Reddit threads-> Reddit Mentions
- Google+ authors -> Google+ Mentions
- News outlets -> News Mentions
- Mendeley readers -> Mendeley Mentions
- PMID -> Pubmed ID

2014

- Author name -> Authors
- Access type -> OA

2015 Affiliation

- Name -> Affiliation

2015

- Id -> Altmetric ID
- Count_news -> News Mentions
- Count_facebook -> FB Mentions
- Count_google_plus ->Google+ Mentions
- Count_video -> Video Mentions
- Score -> Altmetric Attention Score
- Count_blog -> Blog Mentions
- Count_peer_review -> Peer Review Mentions
- Count_reddit ->Reddit Mentions
- Count_wikipedia ->Wikipedia Mentions
- Date -> Published
- Count_twitter -> Twitter Mentions
- Count_weibo -> Weibo Mentions
- Count_research_hi ->Research Hi Mentions
- Open_access ->OA

2016 Affiliation

• Name -> Authors	• Affiliations -> Institutions	

2016

• Score -> Altmetric Attention Score	• Publication Date -> Published	• Subject -> Categories
• Blurb -> Categories		

2017 Affiliation

• Missing header1 deleted (unknown context)	• Missing header2 -> Grid ID	• Affiliation -> Institution

2017

• Score -> Altmetric Attention Score	• Date -> Published	• #news stories -> News Mentions
• #blogs -> Blog Mentions	• #policydocs -> Policy Mentions	• #tweet -> Tweet Mentions
• #peerreview -> Peer Review Mentions	• #facebook -> FB Mentions	• #wikipedia -> Wikipedia Mentions
• #google+ -> Google+ Mentions	• #reddit -> Reddit Mentions	• #pins -> Pinterest Mentions
• #f1000 -> F1000 Mentions	• #Q&A -> Q&A Mentions	• #video -> Video Mentions
• #syllabi -> Syllabi Mentions	• #mendeley -> Mendeley Mentions	• Subject -> Categories

2018

• #dimensions citation -> Dimension Citation Mentions	• Open access -> OA	• Publication date -> Published
• Subjects-> Categories		

2019

• Open access -> OA	• #mendely readers -> Mendeley Mentions	• #dimension citation -> Dimension Citation Mentions
• Subjects -> Categories		

These changes made has no big impact on information we can derive from the data as the ambiguity of the attribute names across the data sets already removes any certainty, we can make about what specifically each attribute details. However, there are now certain generalizations on the data that we must address overall.

1. The attributes that are altmetric information and ends with the suffix 'Mentions' are recorded unique instances of any activity by users on that platform however we now do not have information as to what activity that may be. This, however, is still more of a clarification on the attribute of say 2013's "Facebook Walls" and 2015's "count_facebook". More information about what each activity could be is found at: https://help.altmetric.com/support/solutions/articles/6000060969-how-is-the-altmetric-score-calculated-
2. Category now is factored by the levels as seen in 2015 data sets in addition to the levels in 2013.

The second modification to the data is done as to further improve readability. However, in general unless stated, missing values are replaced with NULL.

2013

- Category's missing values are manually added based on personal judgment on what category fits. This is done as this attribute is of interest to our research and it is feasible to manually fill in missing values.
- It is assumed that for factors of the OA attribute that OA means Yes to being open access while missing values means that the publication was not. The certainty of this decision appears reasonable as the distribution of values is like the following years. Note, this substitution was done instead of leaving blank values as the attribute would be useless for analysis and it is not feasible to access the links to each publication individually to confirm OA information.
- The altmetric score is rounded to nearest whole

2014

- It is assumed for attribute OA that a value of open access means Yes and pay wall means No to open access.
- Published date format is standardized to dd/mm/yy

2015 Affiliation

- Affiliation's Type attribute with missing values are set to Private
- Published date is standardized

2015

- Columns countries and institutions are removed as they are empty

2016

- Published date is standardized
- Position column removed (no purpose)
- Value of Free in OA assumed to be Yes

2017

- Value of paywalled and free to view in OA assumed to be No and Yes respectably

2018

- Not OA assumed to be No
- Only the first 100 rows are kept
- Rank removed (no purpose)

2019

- Values of closed and free to read in OA assumed to be No and Yes respectably.

A data dictionary (first table) and quality assessment (second table) can then be made for the new attributes. Attributes of interest for analysis are highlighted yellow.

The analysis section for each year is then performed, note that no summary statistics were produced. This is because most of the attributes of interest are categorical variables which are better summarized with graphs. Altmetric attention score is the only numerical variable of interest and it is currently not relevant to be summarized by itself.

(There are a lot of graphs after the dictionary and data quality assessment, I recommend skipping to page 30 for the discussion)

Attribute Name	Definition	Format	Format exceptions
ADS Bibcode	Bibliographic code to identify publications in the ADS data base	String (YYYYJJJJJVVVVMPPPPA)	NULL or string value
Affiliation	Provides the name of affiliated institution/s to the publication	String	NULL or string value with comma delimiter
Altmetric Attention Score	The altmetric attention score for a publication	Positive Integer including 0	
Altmetric ID	ID to reference a publication on the altmetric site	Positive Integer	Has to be 8 digits
ArXiv ID	ID to reference a publication on arxiv	String (YYMM.NNNNN)	Null or string value
Authors	Name of the author/s of the publication	String (FirstName SecondName Initials)	Null or string value with comma delimiter
Blog Mentions	Count of unique mentions of a publication on a blog source	Positive Integer Including 0	
Category	Fits a publication's topic into one of the following areas: Agricultural and Veterinary Sciences, Biological Sciences, Chemical Sciences, Earth Sciences, Engineering, Environmental Sciences, History and Archaeology, Information and Computing Science, Medical and Health Sciences, Physical Sciences, Psychology and Cognitive Sciences, Studies in Human Society, Being human, Current events, Medical matters, Offbeat, Real-life science fiction and The world we live in.	String	
Checked For Gaming/Incorrect Mentions	Describes if the altmetric score and results for a publication has been checked to not be affected by gaming and incorrect mentions	String (Yes or No)	
CiteULike Mentions	Count of unique mentions of a publication on the CiteULike service	Positive Integer including 0	
Content Type	Fits a publication's type into one of the following areas: Analysis, Article, Brief report, Correspondence, Early Release Article, Editorial, For Debate, Letter, Original Investigation, Perspective, Policy Forum, Pre-print, Recommendations statement, Report, Review, Short Article, Short Communication, Special Communication and Special Report.	String	
Country	The country of the affiliated Institution	String	NULL or string value with ';' delimiter
Description	Describes what the publication is about	String	
Dimensions ID	ID reference for a publication in the dimensions database	String(pub.NNNNNNNNNN)	NULL or string value
DOI	Digital object identifier for a publication	String	NULL or string value
F1000 Mentions	Count of unique mentions of a publication on the f1000 research	Positive Integer Including 0	
FB Mentions	Count of unique mentions of a publication on Face Book	Positive Integer including 0	
First Name	First name of one author of the publication	String	
Google+ Mentions	Count of unique mentions of a publication on Google Plus	Positive Integer including 0	
Grid City	City of the affiliation institution	String	NULL or string value
Grid Country	Country of the affiliated institution	String	NULL or string value
Grid Name	Name of the affiliated Institution	String	
Grid State	State of the affiliated Institution	String	NULL or string value

Had Corrections	UNKNOWN		
Handle.net IDS	ID reference for a publication on Handel.Net	String	NULL or string value
Had Affiliation Data	Describes if the publication has data on its affiliated institutions	String (Yes or No)	
Institution	Another name for Grid Name	String	
Journal	Describes the Journal to which the publication is associated with	String	
Journal ISSNs	ISSN of the journal	String	NULL or string value
Journal Rank	UNKNOWN		
Last Name	Last name of an author of the publication	String	
Link	A link to where the publication can be accessed online	String	NULL or string value
LinkedIn Mentions	Count of unique mentions of a publication on LinkedIn	Positive Integer including 0	
Mendeley Mentions	Count of unique mentions of a publication on Mendeley	Positive Integer including 0	
MSM Mentions	Count of unique mentions of a publication on MSM	Positive Integer including 0	
News Mentions	Count of unique mentions of a publication from news sources	Positive Integer including 0	
News Tweet Ratio	Ratio of news and twitter count	Float	
OA	Describes if a publication is open access or not	String (Yes or No)	
Patent Mentions	Count of unique mentions of a publication in patents	Positive Integer Including 0	
Peer Review Mentions	Count of unique mentions of a publication from peer reviews	Positive Integer including 0	
Pinterest Mentions	Count of unique mentions of a publication on Pinterest	Positive Integer including 0	
PubMed ID	ID reference for a publication on Pub Med	String	NULL or string value
Policy Mentions	Count of unique mentions of a publication on sources of policy	Positive Integer including 0	
Published	Describes when the publication was published	String (DD/MM/YY)	String or numeric value
PubMedCentral ID	ID reference for a publication on PubMedCentral	String	NULL or string value
Q&A Mentions	Count of unique mentions of a publication on Q&A sources	Positive Integer including 0	
Reddit Mentions	Count of unique mentions of a publication on Reddit	Positive Integer including 0	
Region	Describes region in the world where the affiliated institution/s are	String	
RepPEc ID	ID reference for a publication on RepPEc	String	NULL or string value
Research Hi Mentions	UNKNOWN		
Sharedit	A link to the publication from sharedit	String	NULL or string value
SSRN	SSRN ID for a publication	String	NULL or string value
Syllabi Mentions	Count of unique mentions of a publication on Syllabi	Positive Integer including 0	
Title	The title of a publication	String	
Twitter Mentions	Count of unique mentions of a publication on Twitter	Positive Integer including 0	
Type	Describes the type of the affiliated institution into: Company, Education, Facility, Government, Healthcare, Nonprofit, Other, Private	String	
URN	URN ID for a publication	String	NULL or string value
Video Mentions	Count of unique mentions of a publication on video sources	Positive Integer including 0	
Weibo Mentions	Count of unique mentions of a publication on Weibo	Positive Integer including 0	
Wikipedia Mentions	Count of unique mentions of a publication on Wikipedia	Positive Integer including 0	
Dimension Citation Mentions	Count of unique citations of a publication on Dimensions	Positive Integer including 0	

Formatting issues has left this page blank

Attribute Name	Is Compete?	Is Timelines?	Is Valid?	Is Consistent?	Is Unique?
ADS Bibcode	No, has NULLs	Yes	Yes	No, has NULLs	No
Affiliation	No, has NULLs	Yes	Yes	No, has NULLs	No
Altmetric Attention Score	Yes	Yes	Yes	Yes	No
Altmetric ID	Yes	Yes	Yes	Yes	Yes
ArXiv ID	No, has NULLs	Yes	Yes	No, has NULLs	Yes
Authors	No, has NULLs	Yes	Yes	No, has NULLs	No
Blog Mentions	Yes	Yes	Yes	Yes	No
Category	Yes	No, some factors are not specific	Yes	Yes	No
Checked For Gaming/Incorrect Mentions	Yes	Yes	Yes	Yes	No
CiteULike Mentions	Yes	Yes	Yes	Yes	No
Content Type	Yes	Yes	Yes	Yes	Yes
Country	Yes	Yes	Yes	Yes	Yes
Description	Yes	Yes	Yes	Yes	No
Dimensions ID	No, has NULLs	Yes	Yes	No, has NULLs	Yes
DOI	No, has NULLs	Yes	Yes	No, has NULLs	Yes
F1000 Mentions	Yes	Yes	Yes	Yes	No
FB Mentions	Yes	Yes	Yes	Yes	No
First Name	Yes	Yes	Yes	Yes	No
Google+ Mentions	Yes	Yes	Yes	Yes	No
Grid City	No, has NULLs	Yes	Yes	No, has NULLs	No
Grid Country	No, has NULLs	Yes	Yes	No, has NULLs	No
Grid Name	No, has NULLs	Yes	Yes	No, has NULLs	No
Grid State	No, has NULLs	Yes	Yes	No, has NULLs	No
Had Corrections	Yes	No	Yes	Yes	No
Handle.net IDS	No, has NULLs	Yes	Yes	No, has NULLs	Yes
Had Affiliation Data	Yes	Yes	Yes	Yes	No
Institution	No, has NULLS	Yes	Yes	No, has NULLS	No
Journal	Yes	Yes	Yes	Yes	No
Journal ISSNs	No, has NULLs	Yes	Yes	No, has NULLs	Yes
Journal Rank	Yes	No	Yes	Yes	No
Last Name	Yes	Yes	Yes	Yes	No
Link	Yes	Yes	Yes	Yes	Yes
LinkedIn Mentions	Yes	Yes	Yes	Yes	No
Mendeley Mentions	Yes	Yes	Yes	Yes	No

MSM Mentions	Yes	Yes	Yes	Yes	No
News Mentions	Yes	Yes	Yes	Yes	No
News Tweet Ratio	Yes	No	Yes	Yes	No
OA	Yes	Yes	Yes	Yes	No
Patent Mentions	Yes	Yes	Yes	Yes	No
Peer Review Mentions	Yes	Yes	Yes	Yes	No
Pinterest Mentions	Yes	Yes	Yes	Yes	No
PubMed ID	No, has NULLs	Yes	Yes	No, has NULLs	Yes
Policy Mentions	Yes	Yes	Yes	Yes	Yes
Published	Yes	Yes	Yes	No, has numbers	No
PubMedCentral ID	No, has NULLs	Yes	Yes	No, has NULLs	Yes
Q&A Mentions	Yes	Yes	Yes	Yes	No
Reddit Mentions	Yes	Yes	Yes	Yes	No
Region	Yes	Yes	Yes	Yes	No
RepPEc ID	No, has NULLs	Yes	Yes	No, has NULLs	No
Research Hi Mentions	Yes	No	Yes	Yes	No
Sharedit	No, has NULLs	Yes	Yes	No, has NULLs	No
SSRN	No, has NULLs	Yes	Yes	No, has NULLs	No
Syllabi Mentions	Yes	Yes	Yes	Yes	No
Title	Yes	Yes	Yes	Yes	No
Twitter Mentions	Yes	Yes	Yes	Yes	No
Type	Yes	No, some factors are not specific	Yes	Yes	No
URN	No, has NULLs	Yes	Yes	No, has NULLs	No
Video Mentions	Yes	Yes	Yes	Yes	No
Weibo Mentions	Yes	Yes	Yes	Yes	No
Wikipedia Mentions	Yes	Yes	Yes	Yes	No
Dimension Citation Mentions	Yes	Yes	Yes	Yes	No

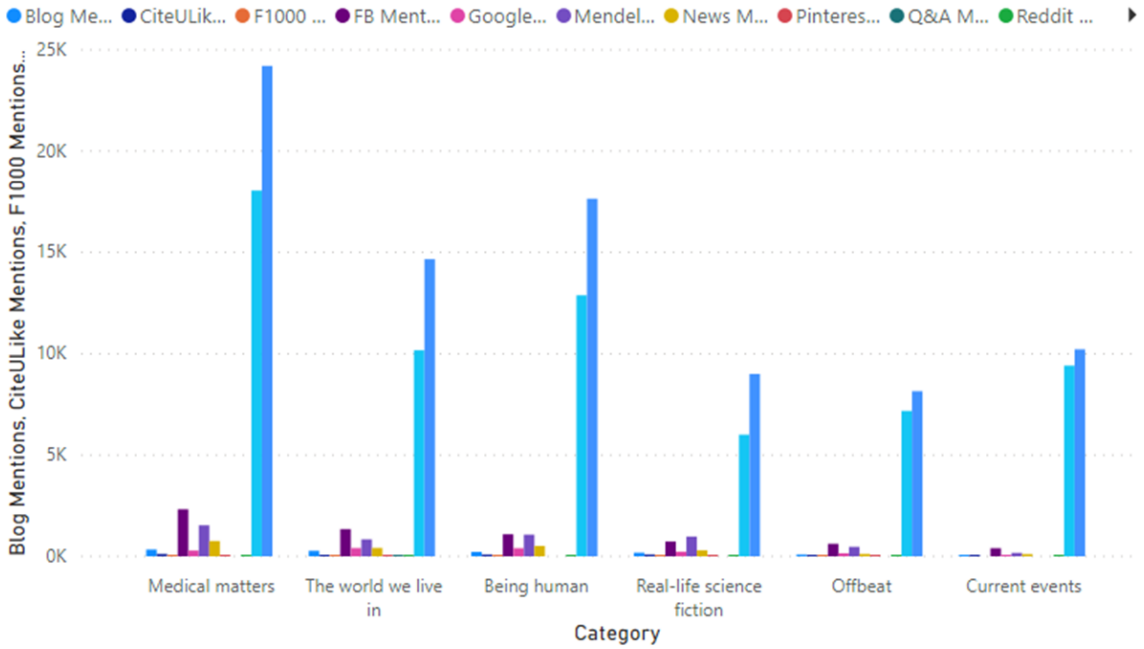
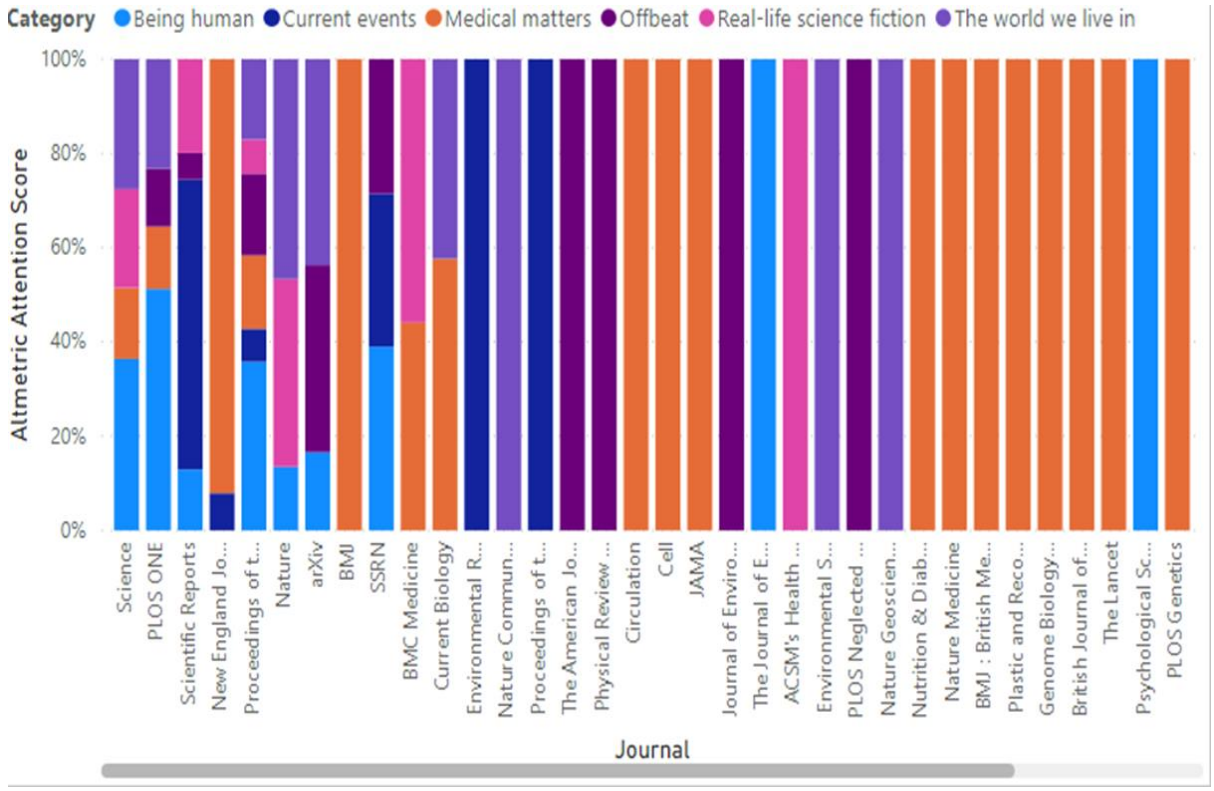


Figure 17: 2013

2013

Figure 18: 2013



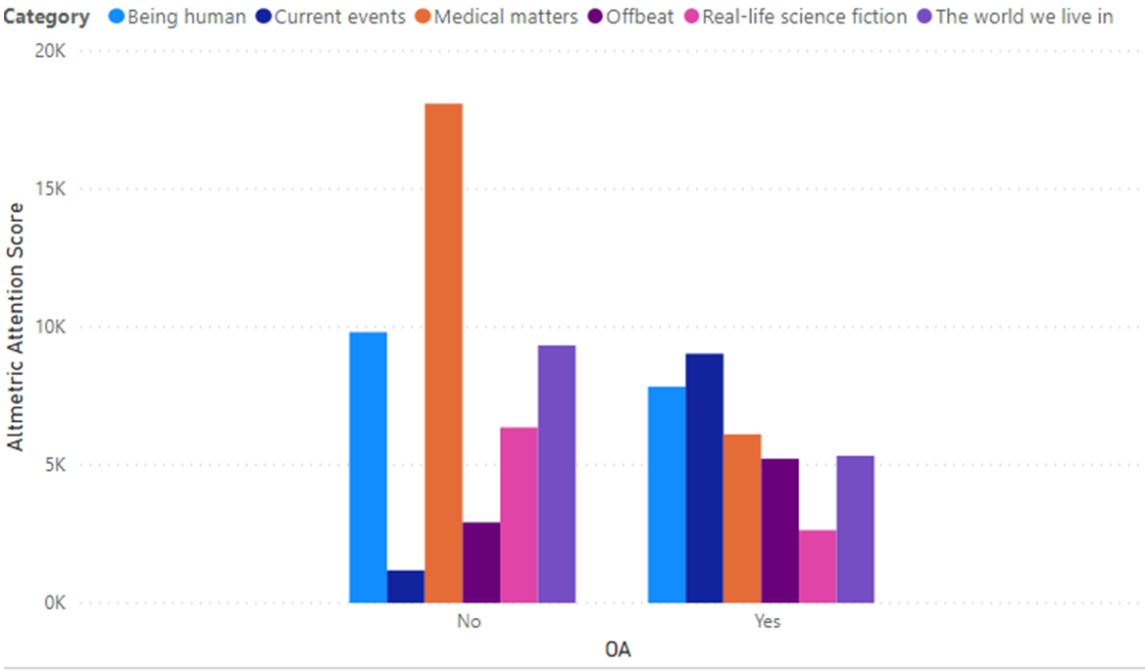


Figure 19: 2013

2014

Figure 20: 2014

2015

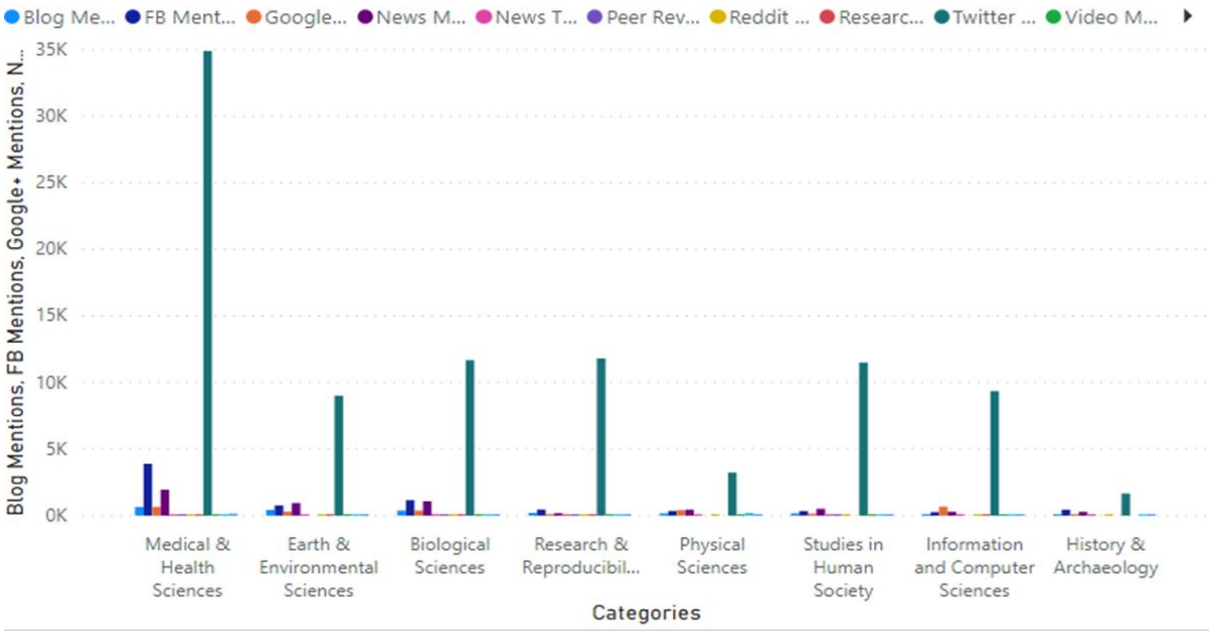


Figure 21: 2015

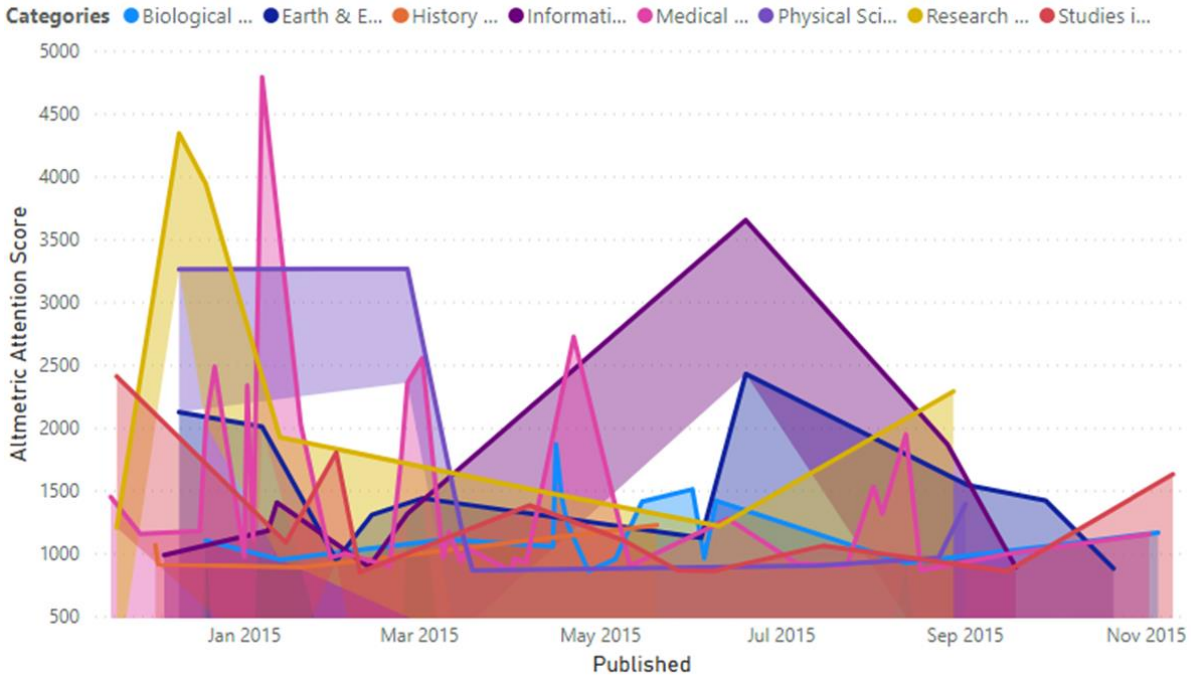


Figure 22: 2015

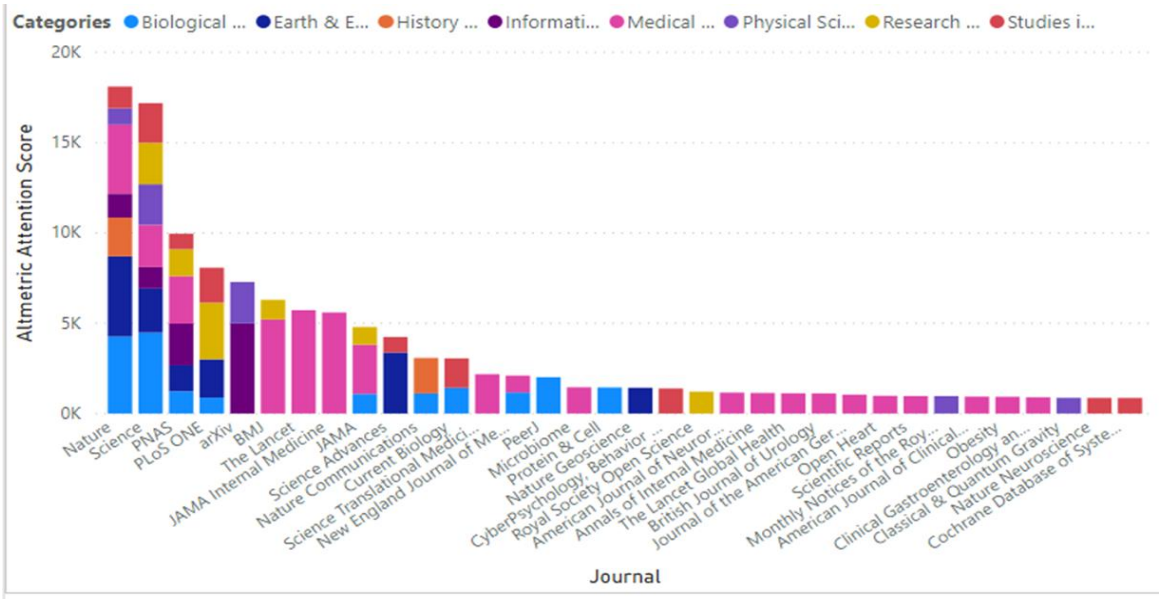


Figure 23: 2015

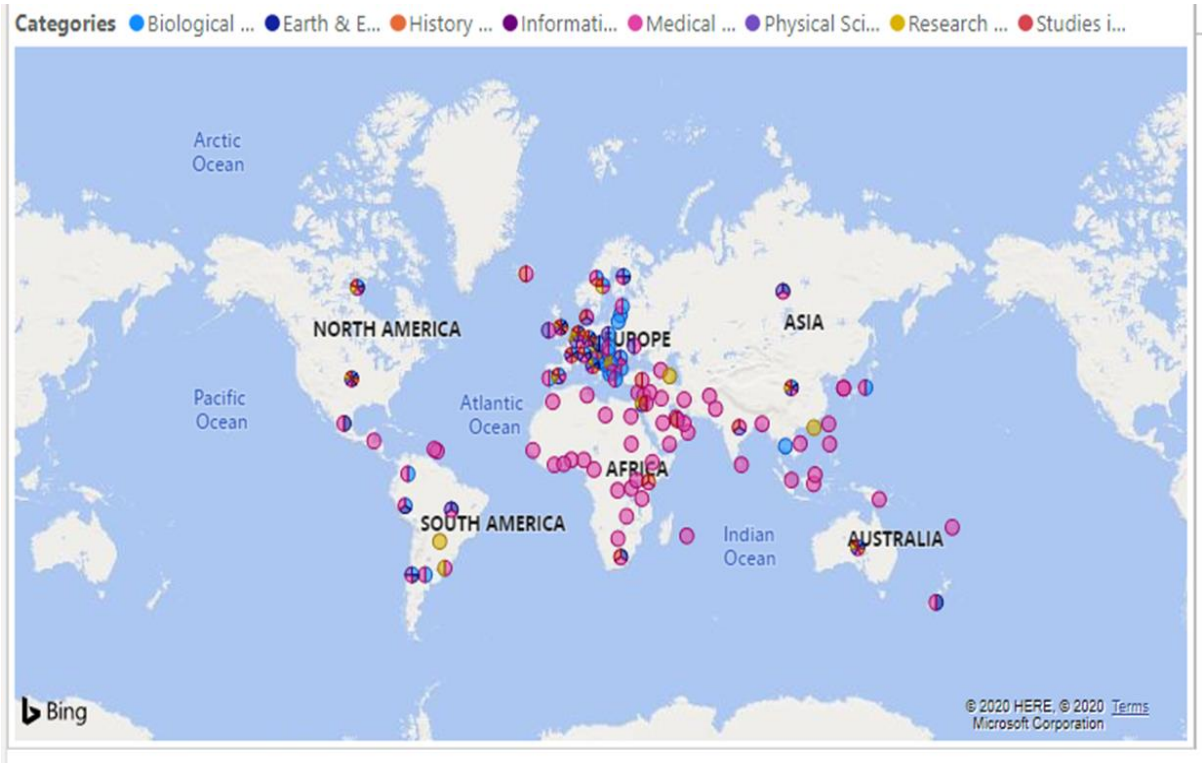


Figure 24: 2015

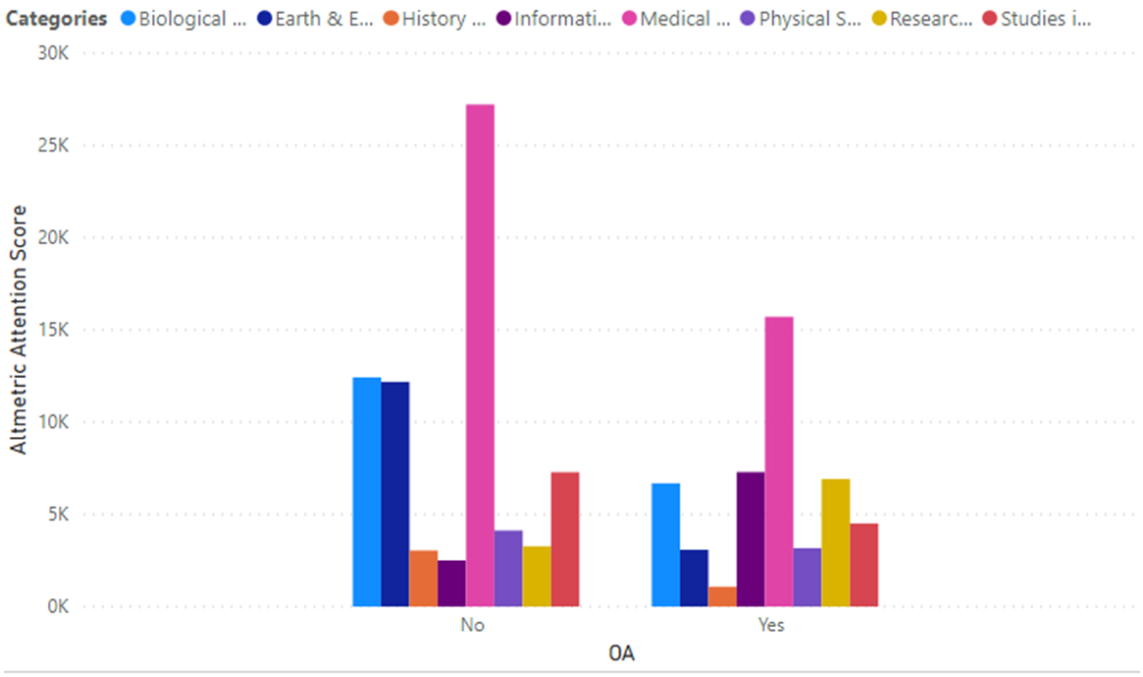


Figure 25: 2015

2016

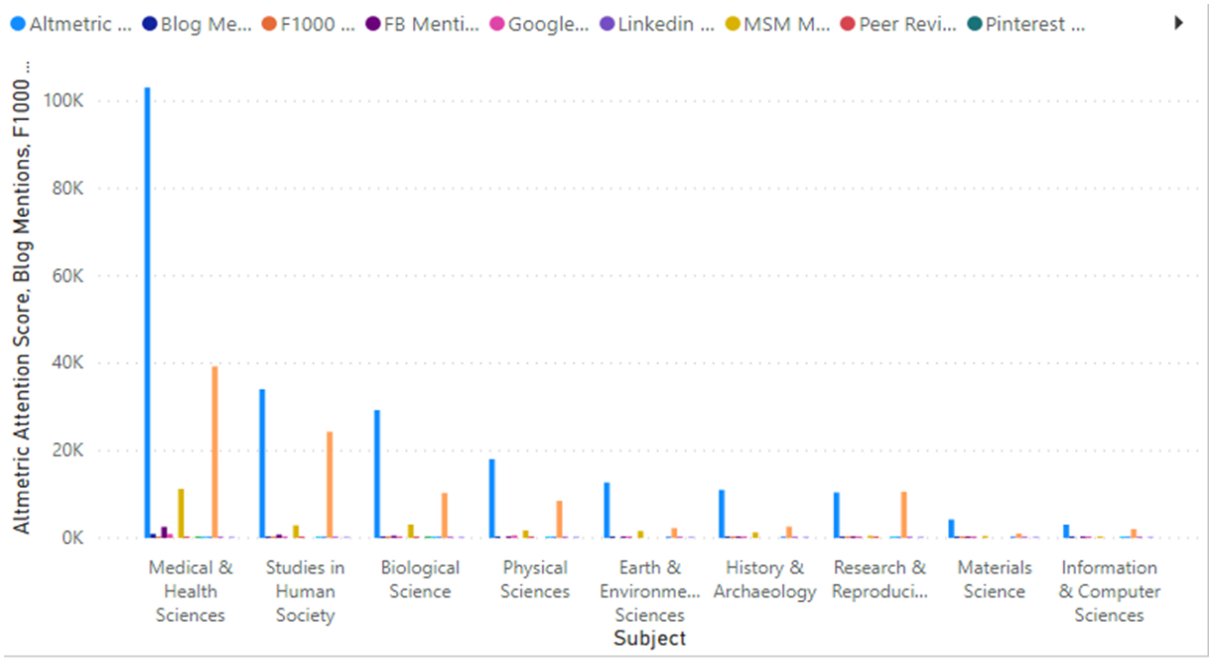


Figure 26: 2016

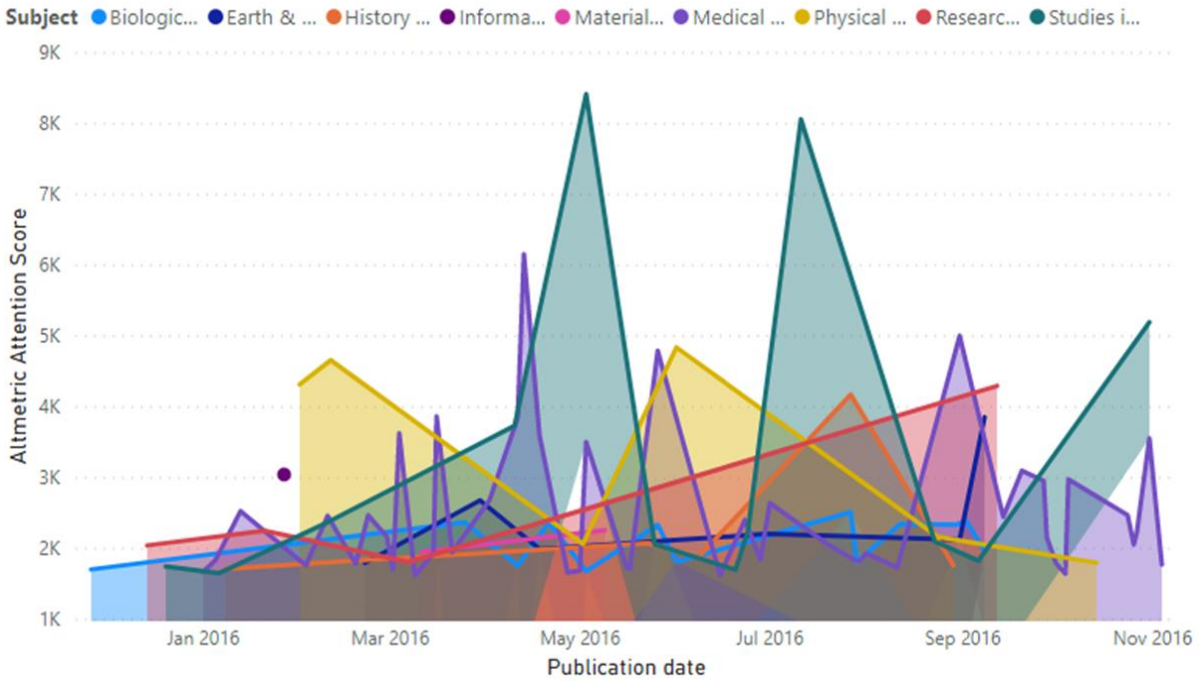


Figure 27: 2016

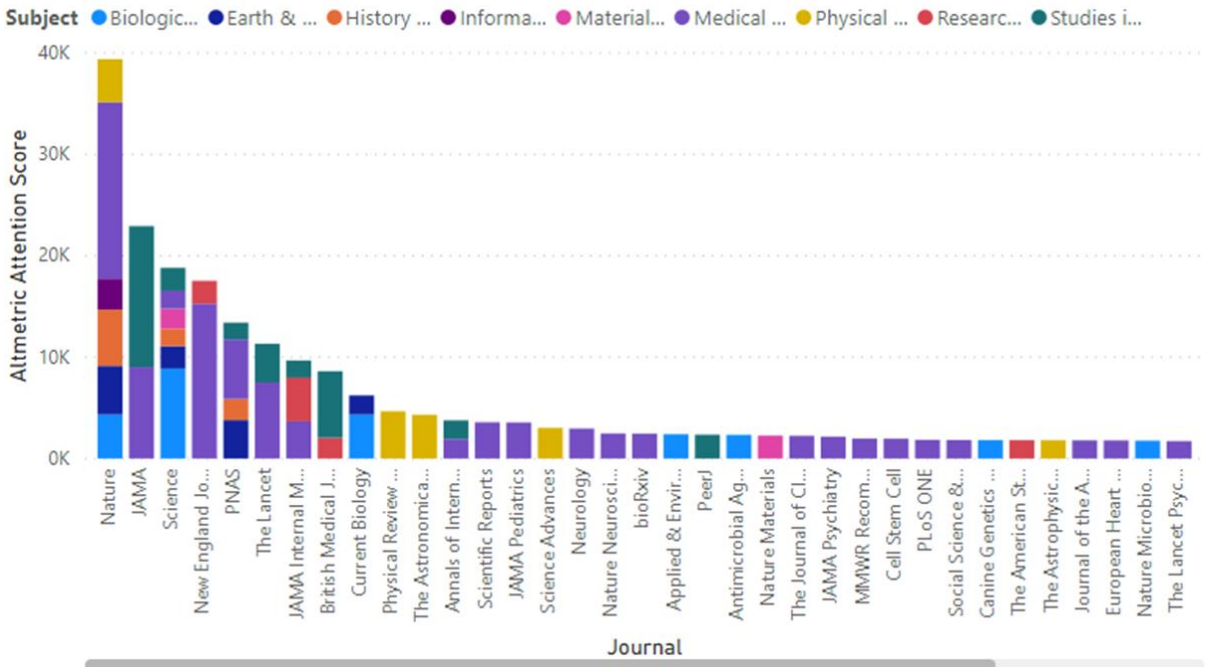


Figure 28: 2016

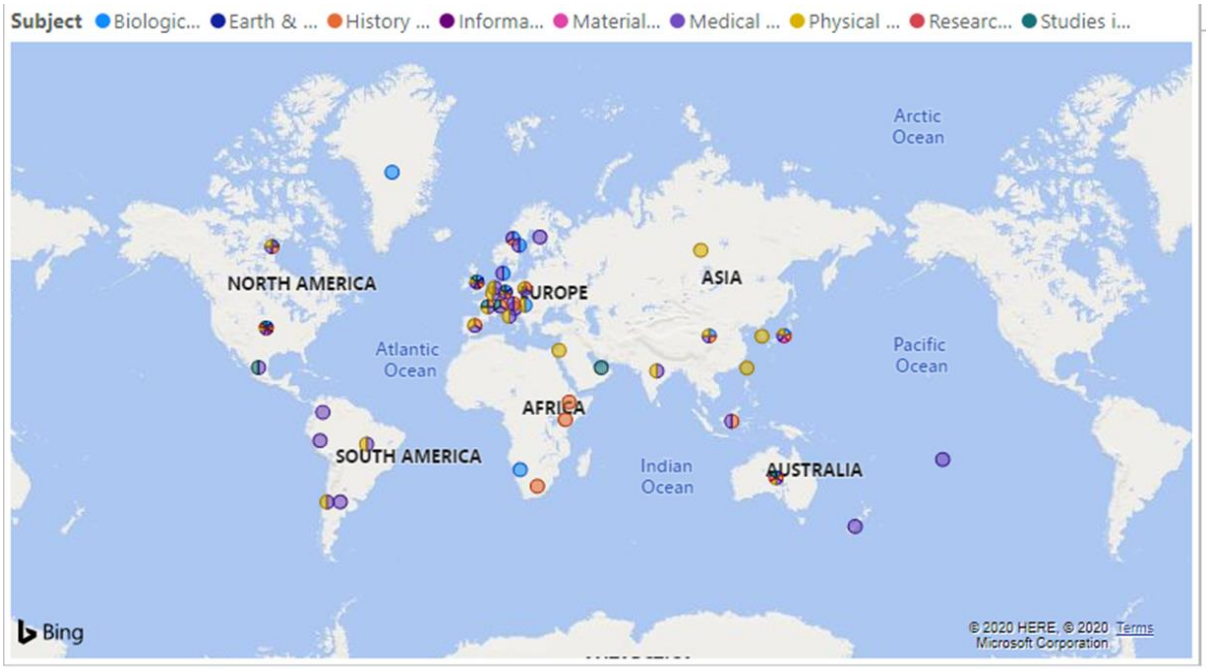


Figure 29: 2016

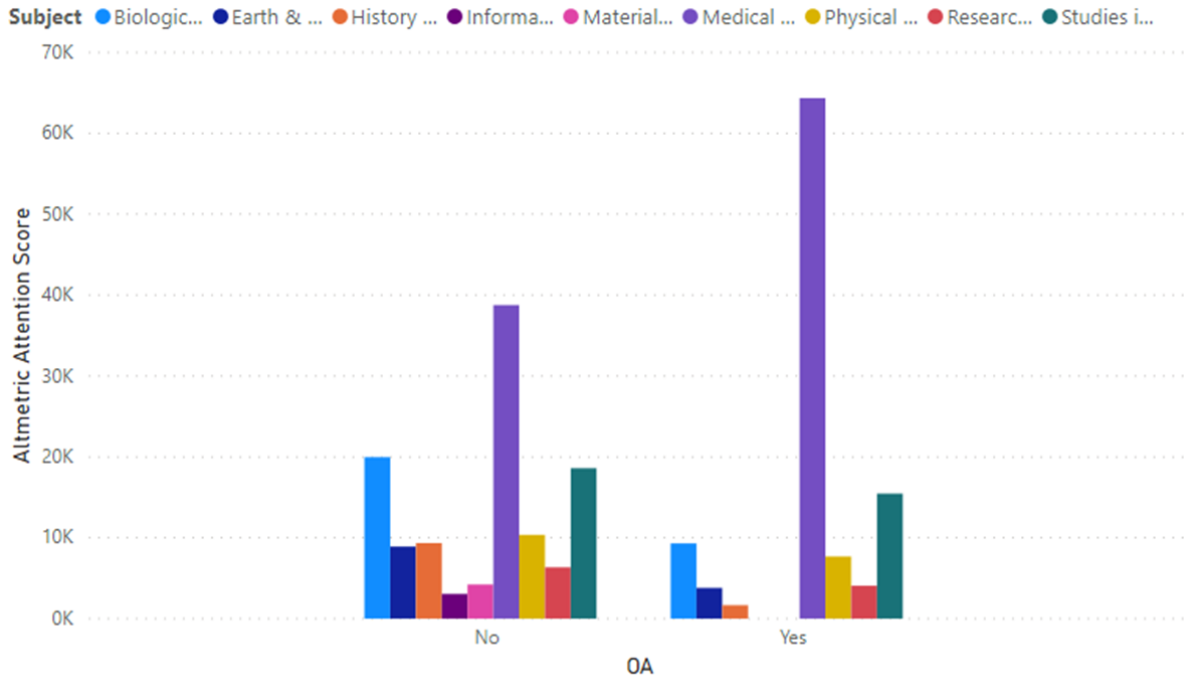


Figure 30: 2016

2017

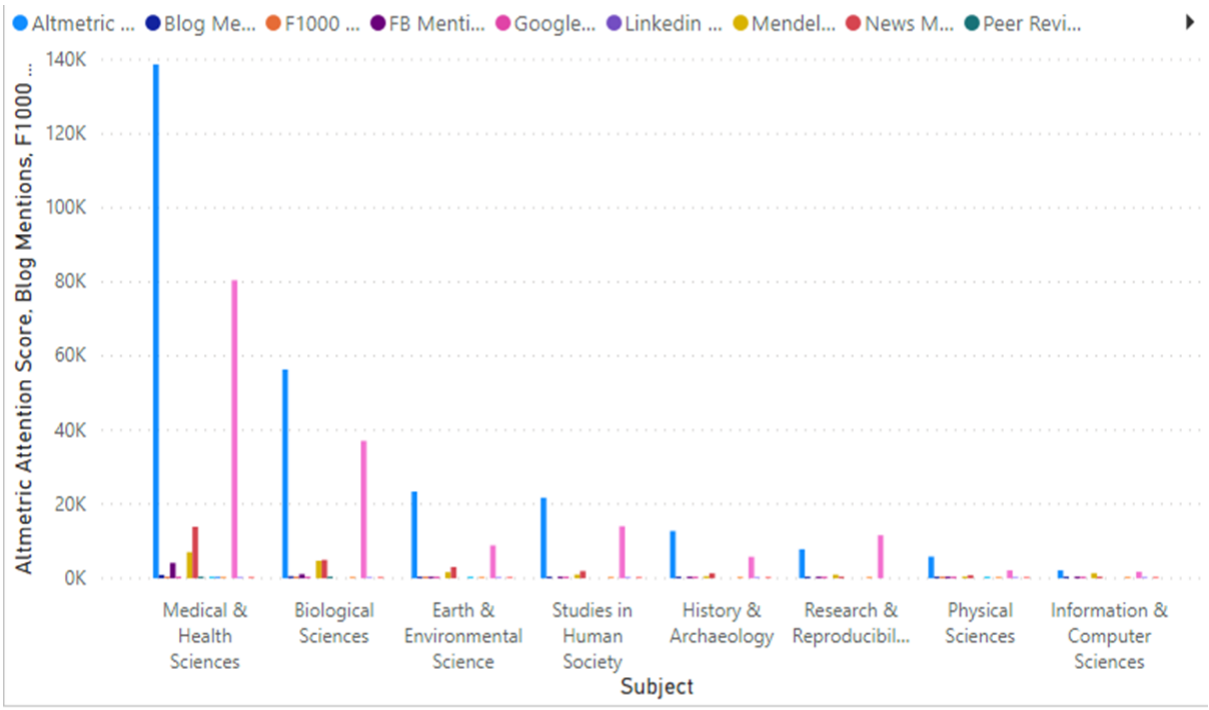


Figure 31: 2017

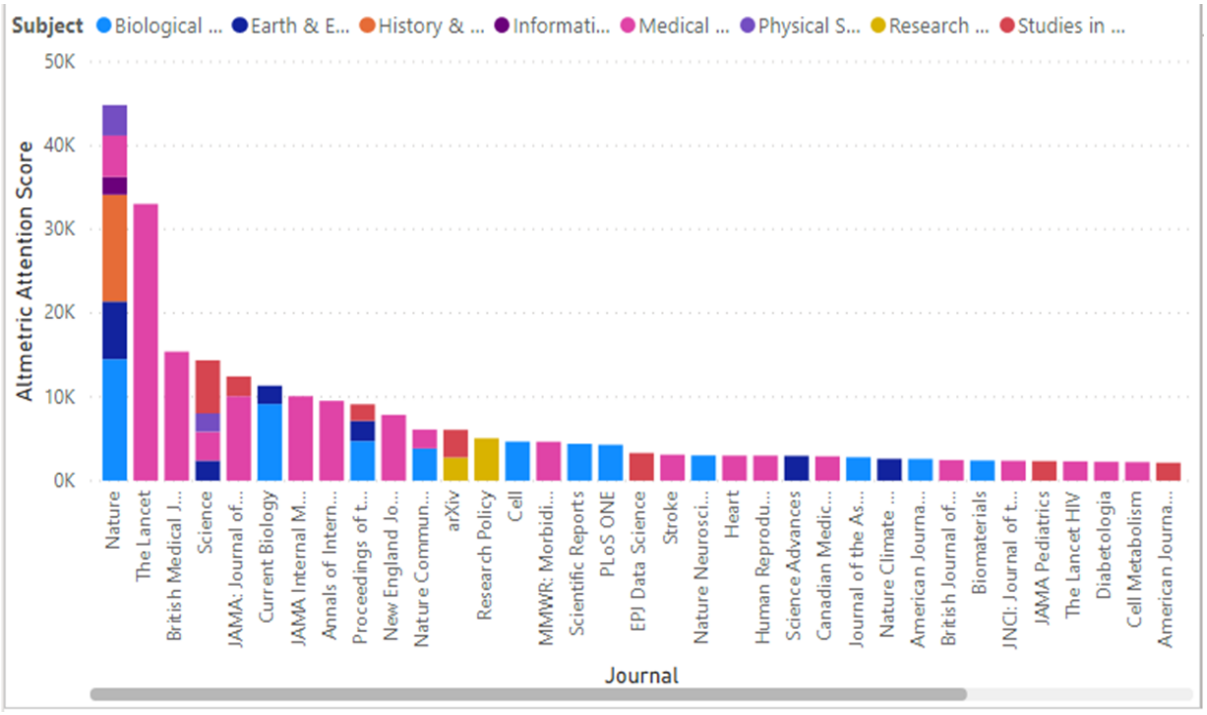


Figure 32: 2017

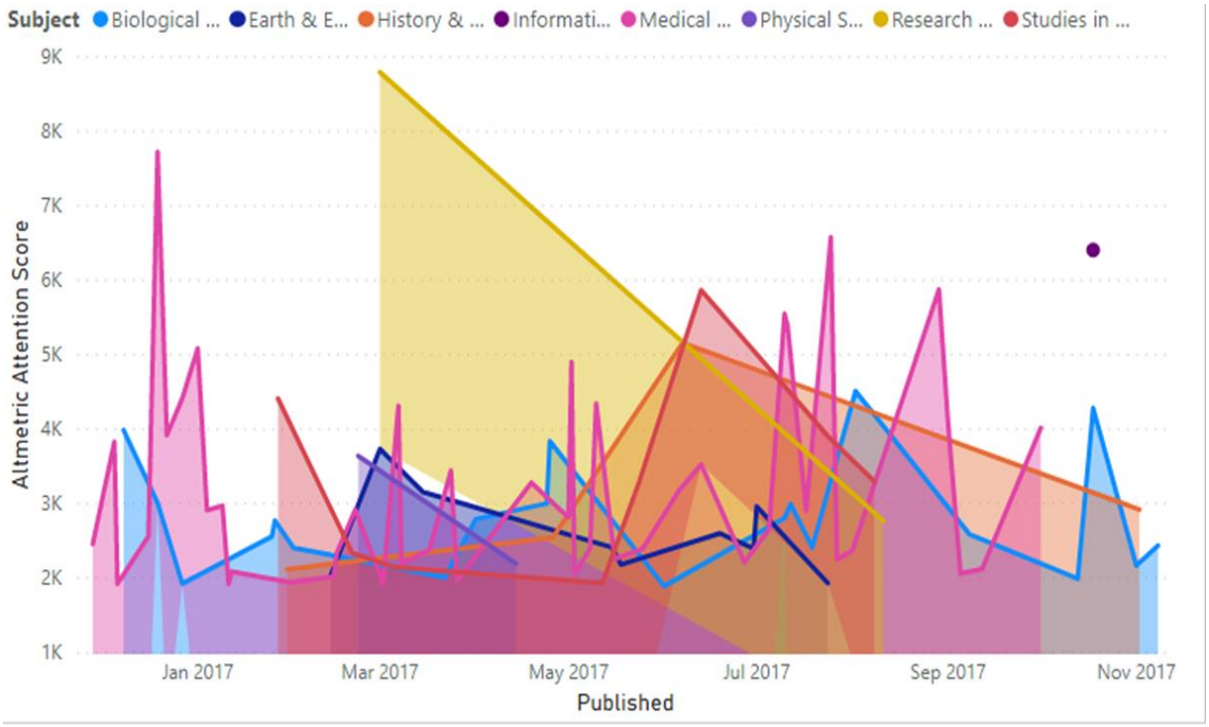


Figure 33: 2017

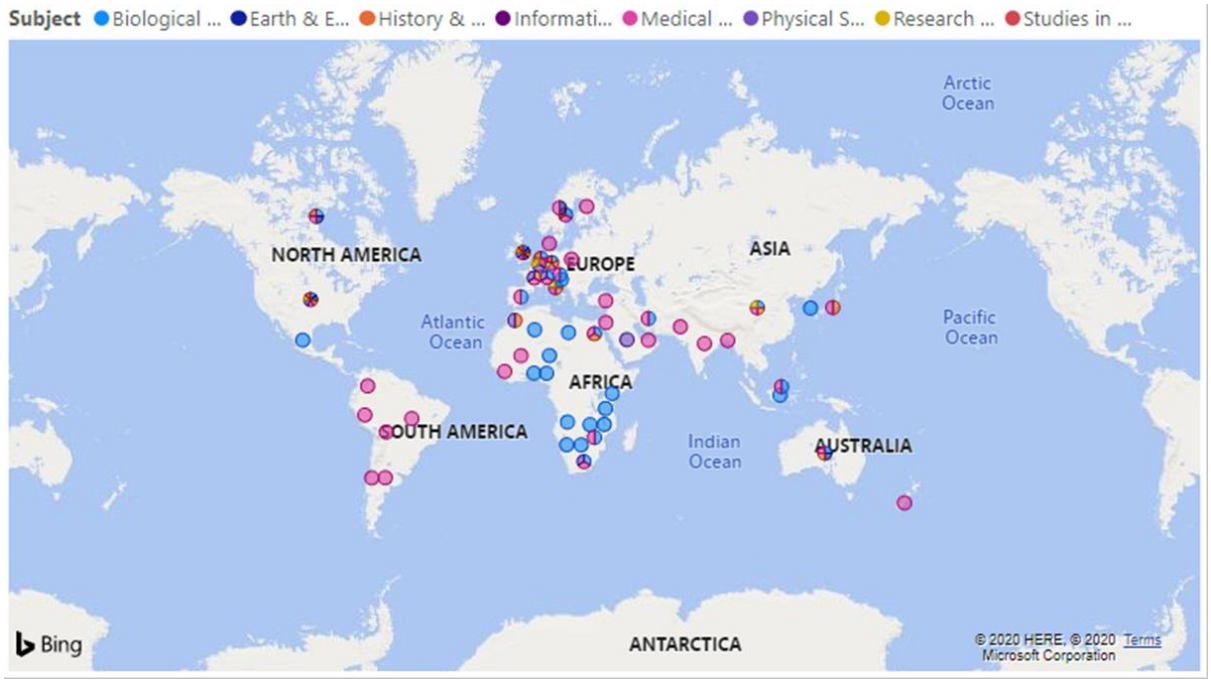


Figure 34: 2017

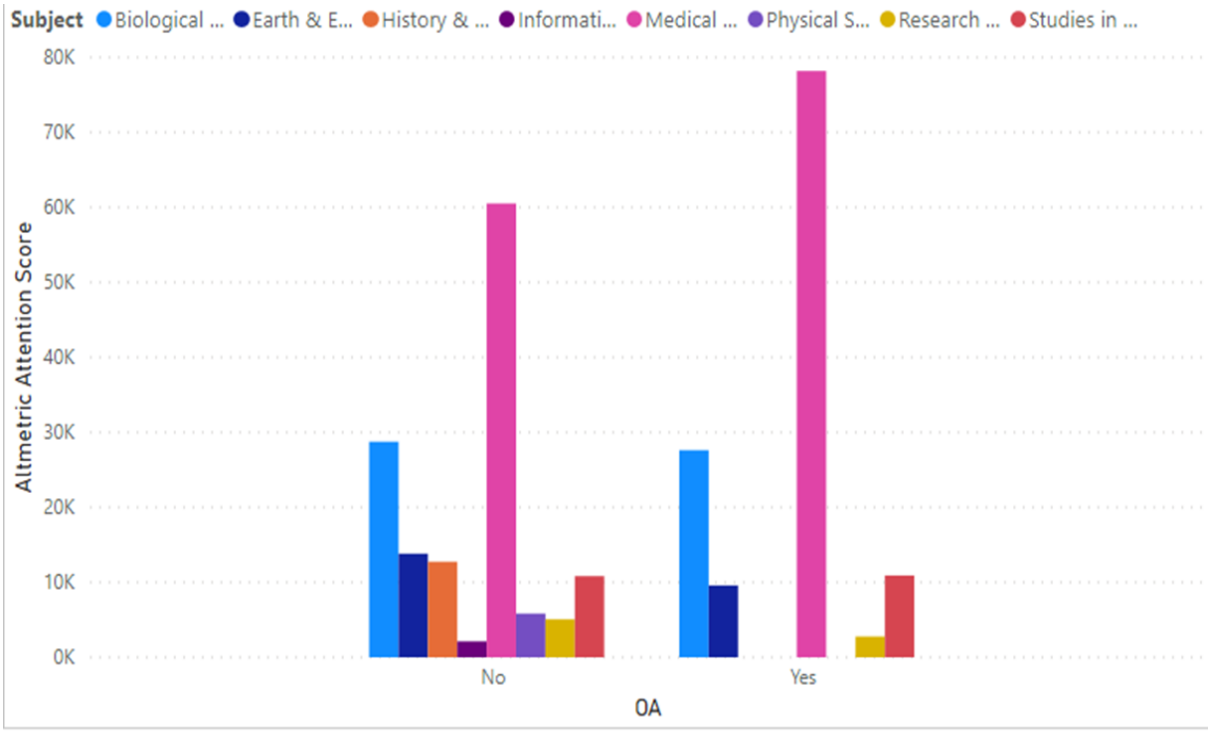


Figure 35: 2017

2018

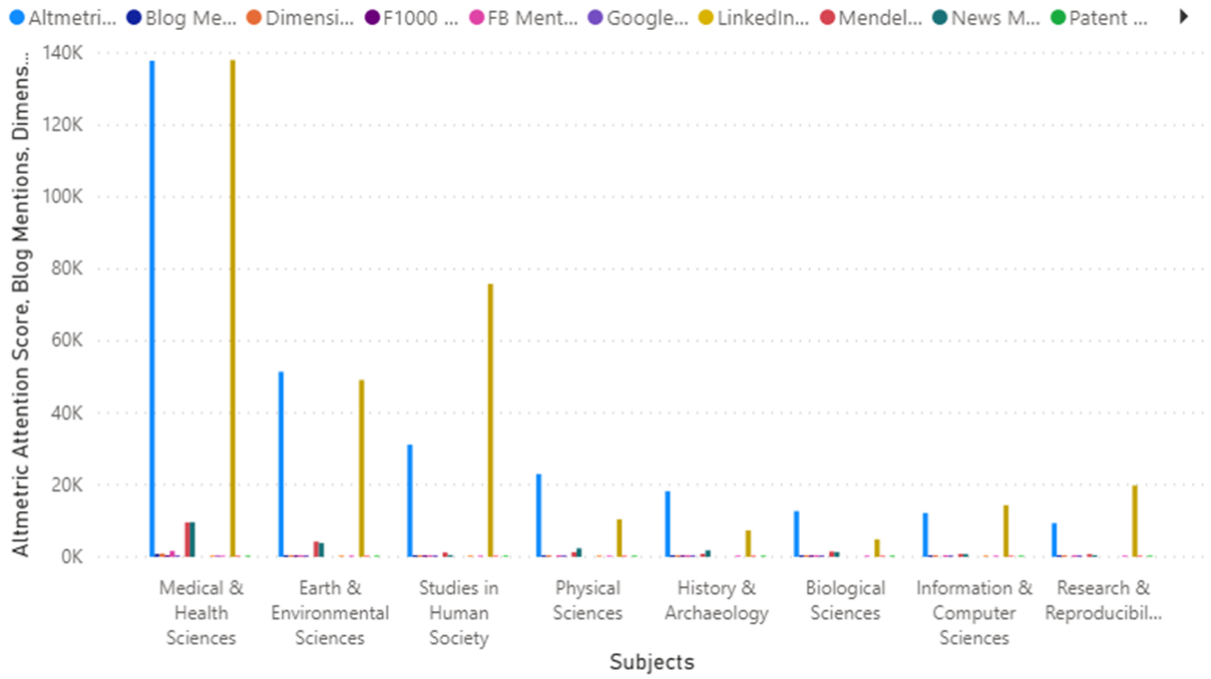


Figure 36: 2018

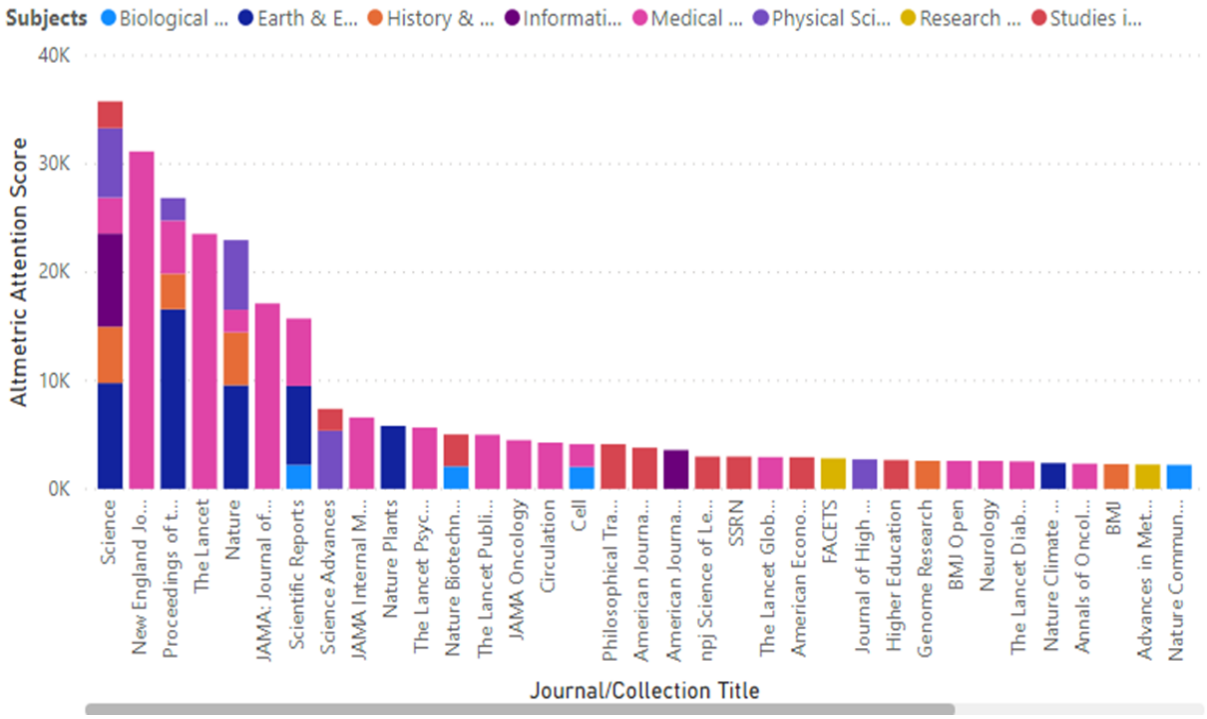


Figure 37: 2018

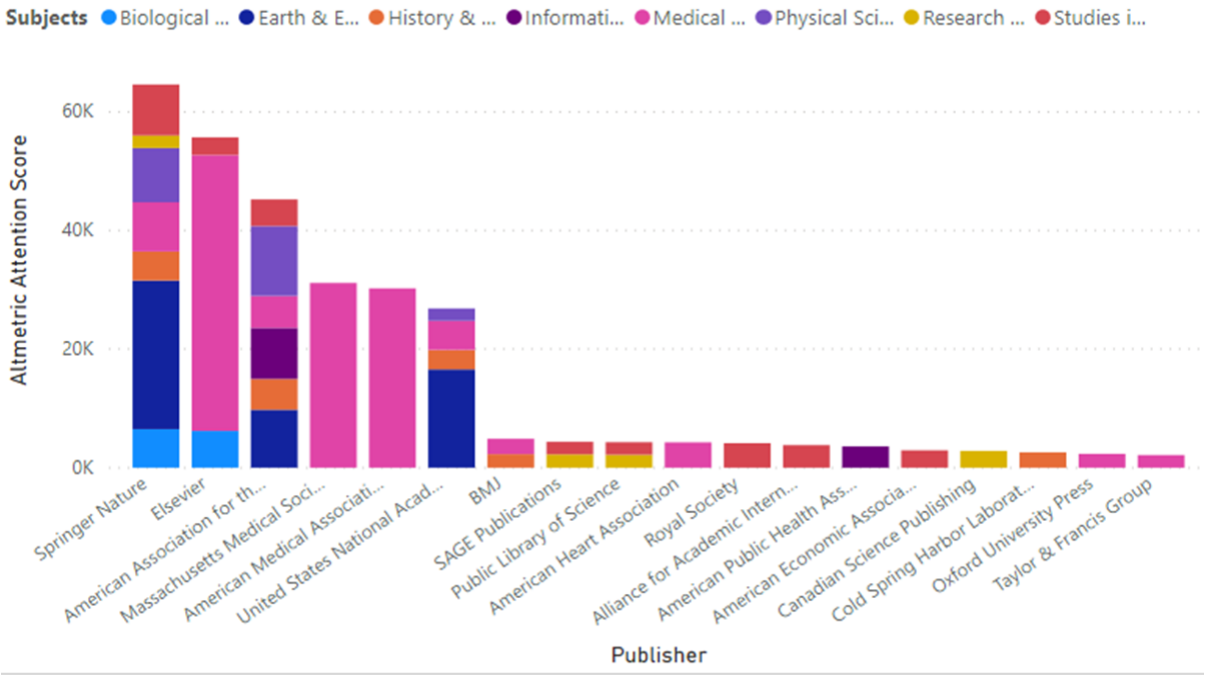


Figure 38: 2018

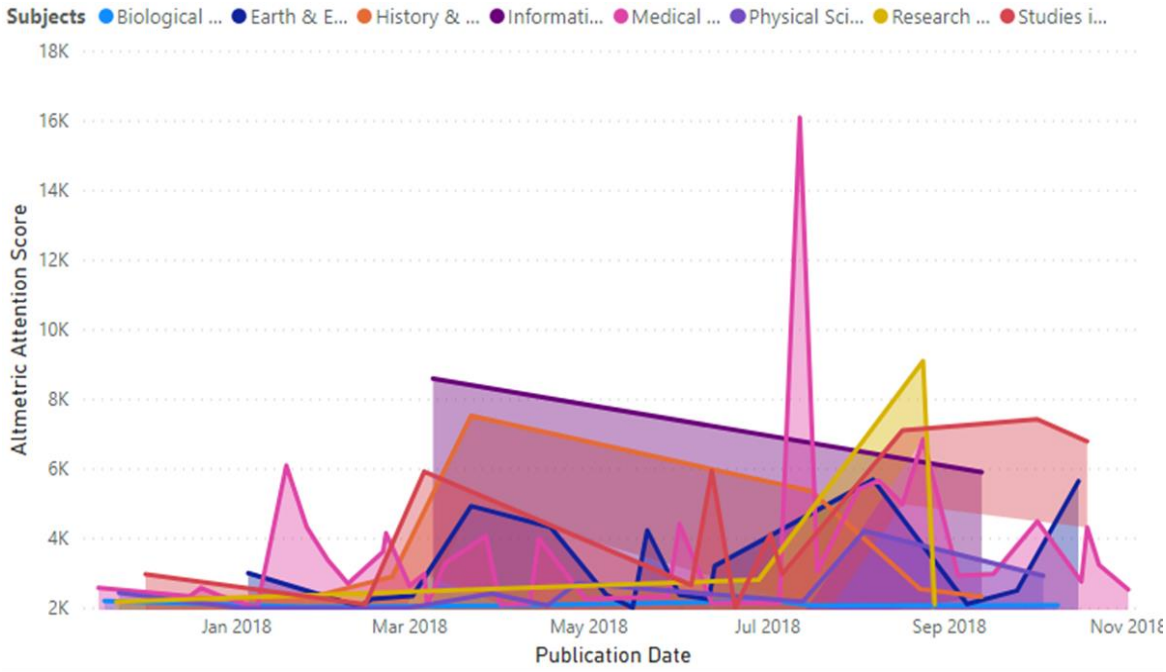


Figure 39: 2018

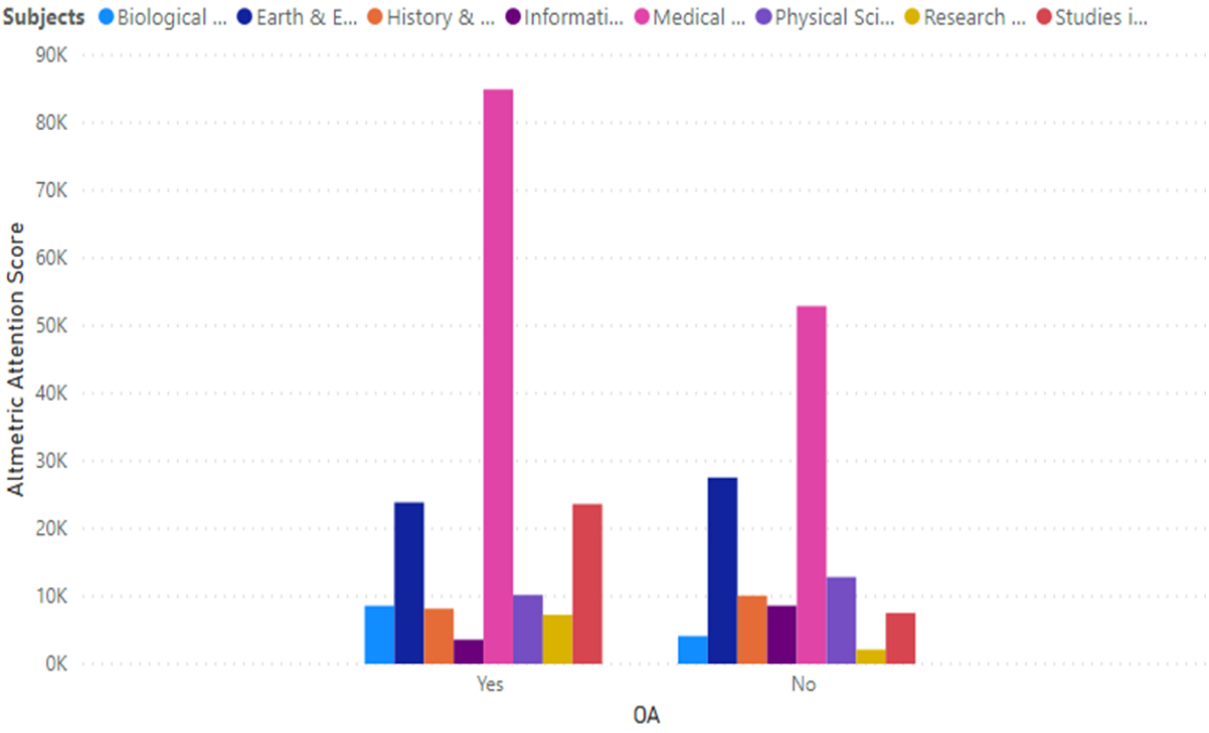


Figure 40: 2018

2019

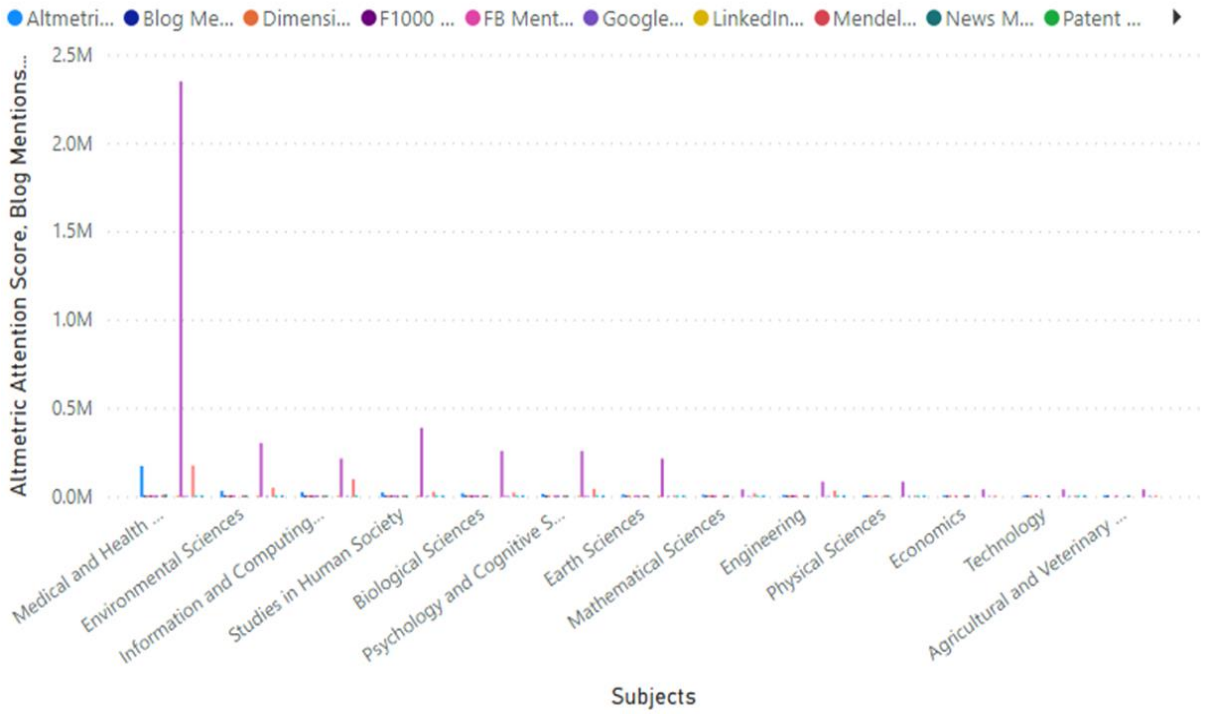


Figure 41: 2019

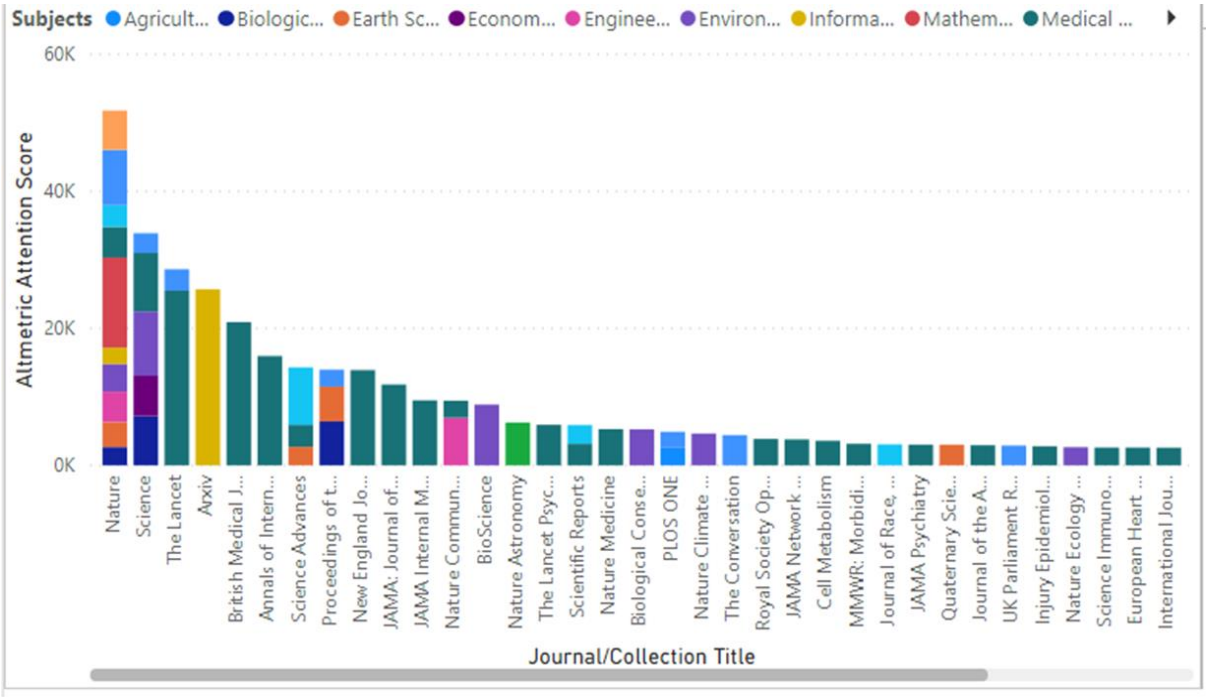


Figure 42: 2019

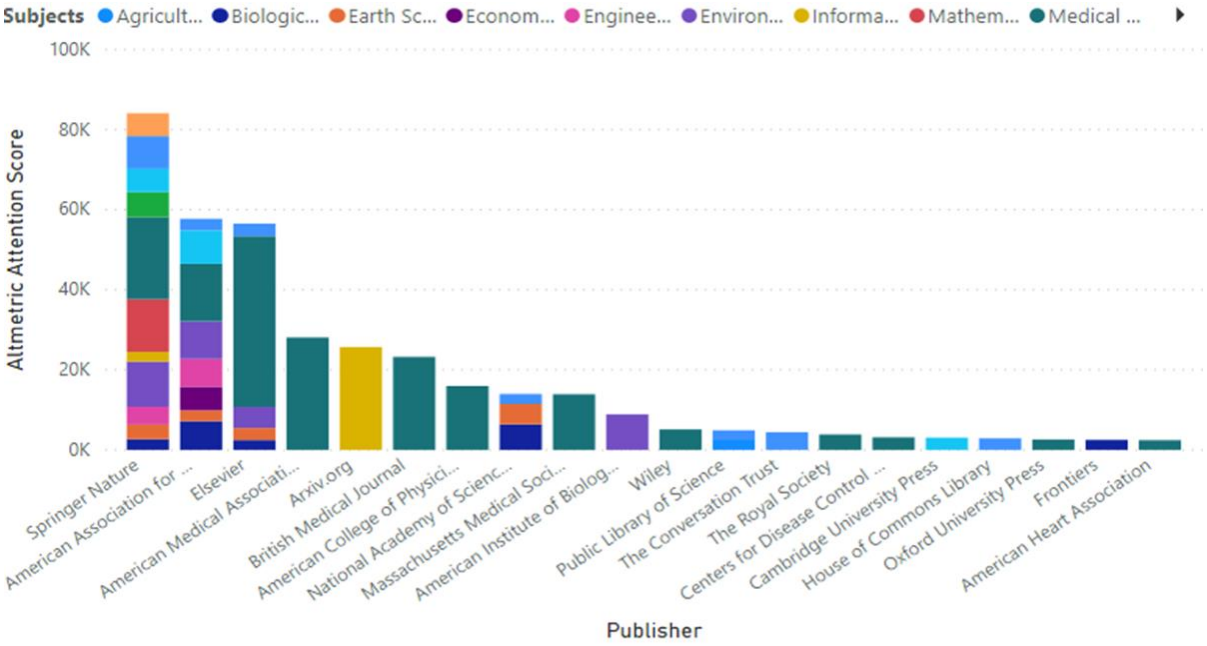


Figure 43: 2019

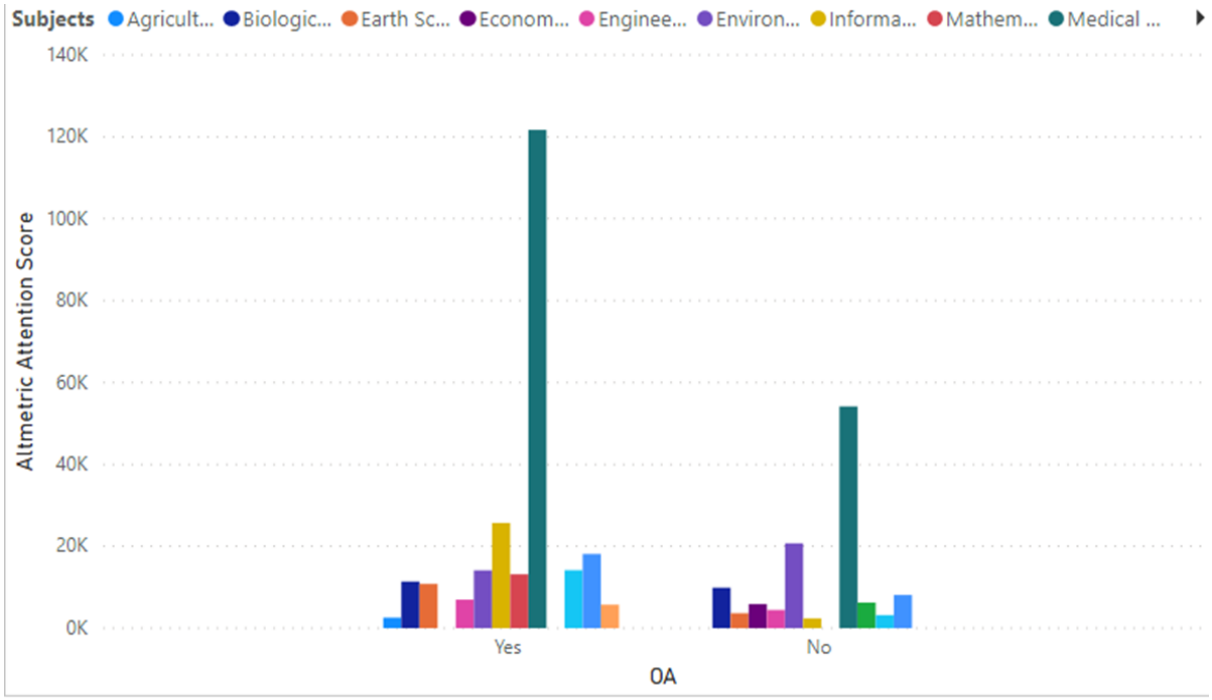


Figure 44: 2019

Discussion

The graphs above produce a summary for attributes of interest each year and potential correlations they may have with each other. The following dot points summarizes some of the results from the graphs:

- From 2013 up to 2016 the majority of category for non-open access publications had in general a larger sum altmetric score. Passing 2016 we see that the effect has grown increasingly opposite with Medical science being most prominent to this change.
    - This may be caused by a larger number of publications being open access in the recent years. Or open access publications are being more popular all together.
    - Additional graph of OA counts through the years supports the second statement, however we can not be sure as we lack sentiment data.

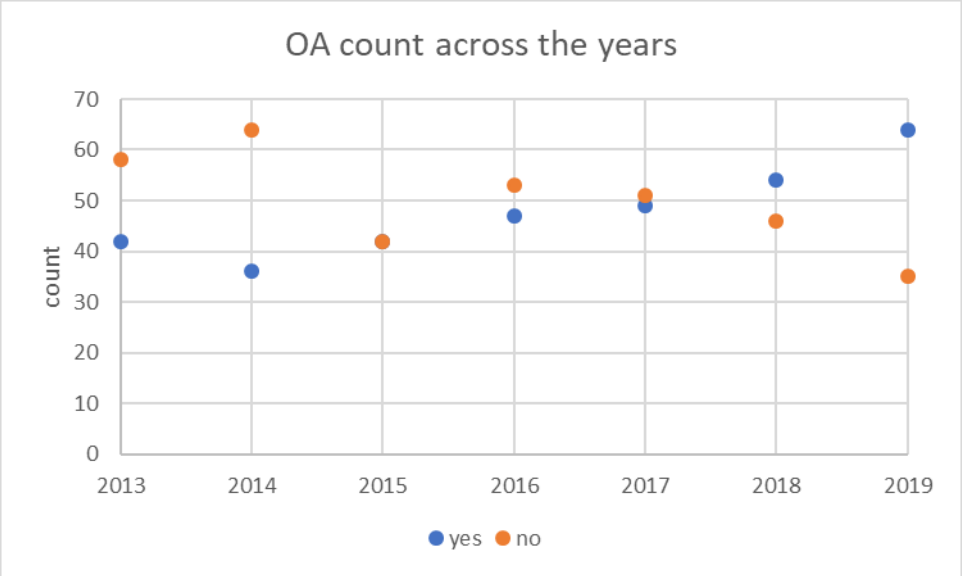


Figure 45

- In general twitter mentions has highest mentions per category, with top three categories being, Medical and Health, Studies in Human Society and Biological Sciences. The relationship between twitter mentions and sum of altmetric score seems to be proportional however there is variation.
    - This suggests that as the twitter weighting is 1, it may be reasonable to say that twitter count can some what predict sum of altmetric score.
    - This may be due to twitter being the soul source of attention for a publication and any fluctuations to the score being higher or lower is due to more mentions from other sources and less twitter mentions.
    - An investigation may be done finding the count of all data sources for a year and applying their altmetric weightings to then determine a ranking of which source has the greatest contribution to altmetric score.

Blog	F1000	FB	Google+	LinkedIn	Mendeley	News	PeerRevie	Pinterest	Policy	Q&A(over	Reddit	Syllabi	Twitter	Video(Yo	Weibo	Wikipedia	Dimensio	Patent
5	1	0.25	1	0.5	0	8	1	0.25	3	0.25	0.25	1	1	0.25	1	3	0	3

Figure 46: Weights for data sources

The graph below represents the results of this investigation and is a rough approximation. With News and Blog potentially being the second and third data sources that has greatest effect on altmetric score. Note that the weights used for each data source obtained from the altmetric site are for general cases and there exist variations to the weights depending on each data source which is not reflected in our current data source counts. This could explain why sum of the weighted data source counts for a year exceed the altmetric attention score.

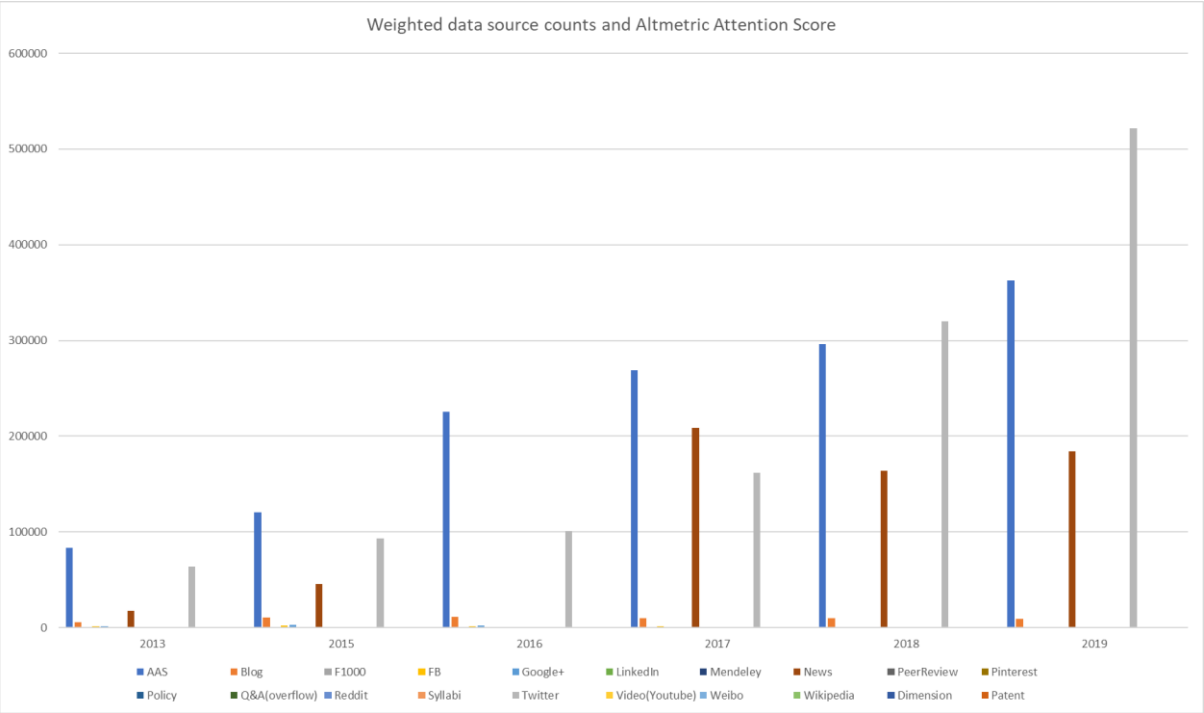


Figure 47

While these results using the sum of altmetric score does not fully describe the relationship between data sources and altmetric attention score for a publication. It does suggest that twitter, news and Blog data sources has the greatest contribution to the score. More research may be done analysing the correlation between each data source and altmetric attention score to support these arguments.

- There does not appear to be a trend for when a publication of a category might get a better altmetric score depending on when it is published in the year.
- Nature and Science has been consistently the top journals that have achieved highest sum altmetric score. While Springer Nature has been the top publisher in 2018 and 2019.
- Demographic data shows that from 2014 on-wards the top 100 publications had more affiliations. This does not necessary mean that a publication with more affiliations has a better altmetric score; without data on less successful publications below the top 100 we can't be sure.

Conclusion

In data understanding we discussed the potential to use further techniques to make use of ID data in order to determine if having certain IDs maybe positively affect altmetric scores. Descriptive data that were not considered in this analysis could also be of use to obtain more information about the category of a publication through data mining.

From a business perspective we can suggest for institutions to try and make their publications open access. Create publications in the topics relating to Medical Science and publishing their research through the Stringer Nature publisher or Science and Nature journals. It might also be worthwhile to focus most of the advertising for a publication on twitter and News sources and increase the number of affiliated institutions per publication. These practices should help increase the altmetric attention scores that an institution receives for a publication, and these changes may be done anytime of the year with no significant loss on impacting altmetric score.

12. Appendix 6: Stephen Maher

Week 4: Data Visualisations and Data Dictionary

Stephen Maher

Note: Data from the Altmetrics Top 100 2018 was also analysed – not included in visualisations as quite similar in overall outcome to 2019

Summary

The data for visualisations and initial analysis came from four sources: Altmetric 2019 and 2018 Top 100 datasets, a paper by Bormann and Haunschild which focussed on paper quality measures vs. Altmetric scores and a dataset from Plum Analytics examining social media characteristics of their top 100 downloads from SciHub (dataset was the subject of an analysis by Elbakyan and Bohannonin in their paper (2016) “Who's downloading pirated papers? Everyone”.

Visualisations (selected below) focussed on the relationships between quality and social media metrics, subject matter which predominated in the top 100 datasets, impact of paper accessibility, institution and journal (among a wider range of exploration).

Initial thinking is that research quality and social media metrics don’t necessarily mix. However, there appear to be some biases to institutions, types of journals and other factors featuring strongly in social media and Altmetric measures, suggesting further exploration may be useful.

It may be that Altmetrics best suggest avenues for promotion of specific types of research by an institution. Subject areas such as Medicine and some areas of the hard sciences appear to receive more social media attention, and this may be leveraged for promotional (either institutional or research) purposes.

Altmetrics 2019 Top 100 Data Review

Data Source

Altmetric Top 100 2019, *The 2019 Altmetric Top 100*, Altmetric, viewed and downloaded 27 March 2020,

<http://www.https://www.altmetric.com/top100/2019/>.

General Data Commentary

- Data cut-off: 15 November 2019
- Data coverage: 365 days to, but excluding, 15 November 2019
- Key Data Measure: Altmetric Score (Top 100 most discussed research)
- **All data is dependent on Altmetric data acquisition and analytical capabilities**
- Altmetrics claims to manually check each piece of discussed research in its list and that it removes news articles
- Altmetrics claims to use a combination of manual and automatic check of the following fields in the data: manually check the following: Open Access status, Publisher, Subject area, Author affiliation and Publication date (first published online)
- Altmetrics claims that it filters for "inorganic (i.e. spam) attention".
- Altmetrics note that for 2019 (and in a departure from prior years), the Top 100 has been expanded to include the following: systematic reviews, researchers' letters to the editor, case studies submitted to medical journals, and many other kinds of scholarly contributions.

Additional Data Commentary

- The columns: Handle.net IDs, ADS Bibcode, arXiv ID, RePEc ID, SSRN, URN, PubMedCentral ID, sharedit and Publisher-preferred links are either empty or largely empty.
- The columns: Patent Mentions, Peer Review Mentions, Weibo Mentions, Linkedin Mentions, Pinterest Mentions, Q@A Mentions and Syllabi Mentions either record only zero or have at most one value in a column.
- Both sets of columns described above should be dropped from the final analysis.
- Some fields available in 2018 have been excluded for 2019
- Open Access values for 2019 correspond to the Open Access values for 2018 but have name differences
- Some fields ceased being collected during 2019

- Where data is present in a column, it is consistent within the column.
- Data Provenance is known.
- Data sourcing organisation is well established and known.
- **Data is deemed to be of High Quality.**

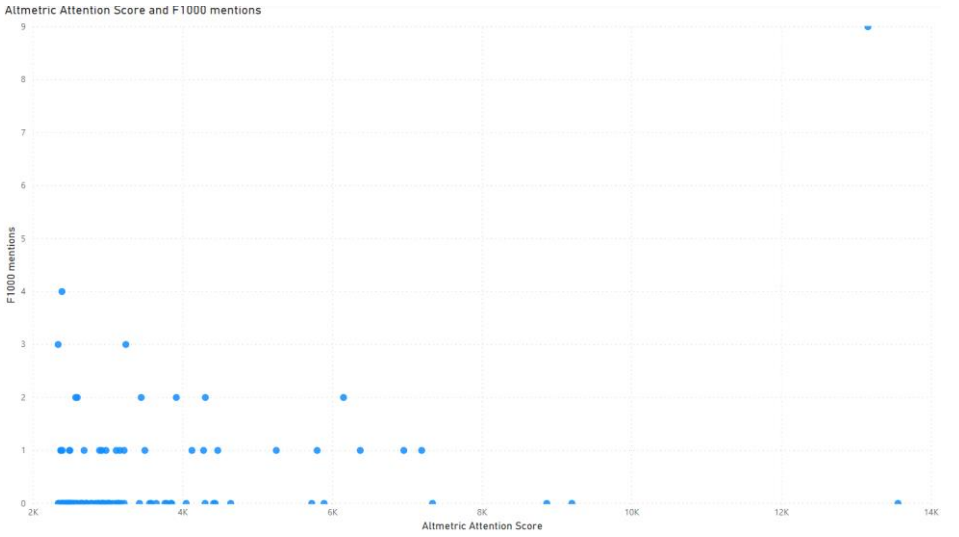
Data commentary from Altmetrics:2019 Top 100

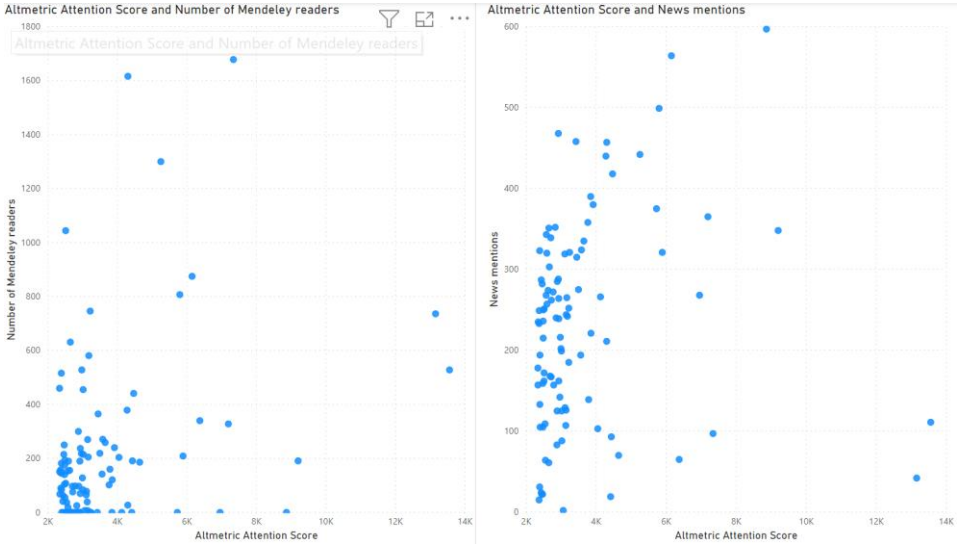
- Some correlation between Altmetric score and news mentions and twitter mentiomns

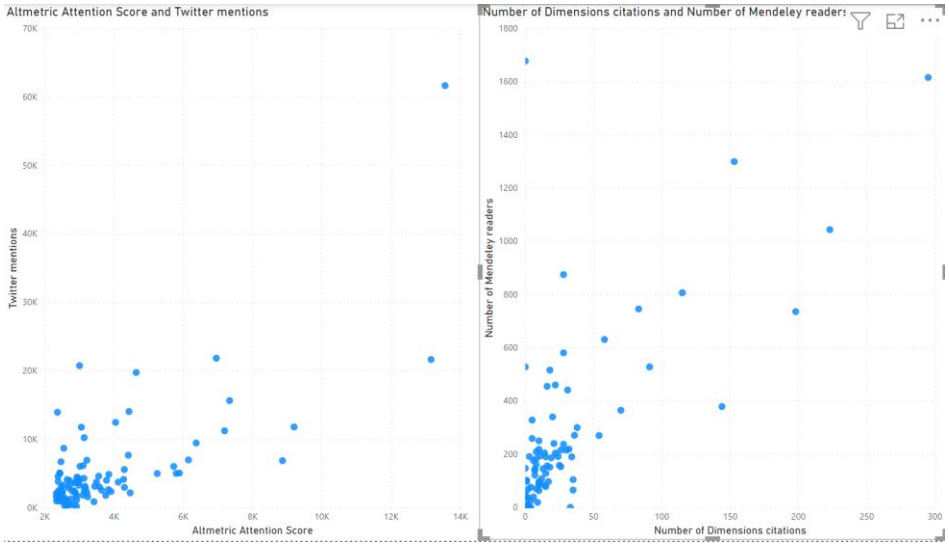
- Free to read appears to have higher average twitter mentions
- Mathematical papers heavily cited, and high average attention score and high average F1000 citations
- More generally, some subjects and journals feature more prominently in the Top 100
- Filters were applied for Open Access and Publisher
- More broadly, weak (at best) evidence of a relationship between published paper quality and its Altmetric score
- Using Twitter as a proxy for social media attention, a few fields stand out as receiving significantly more attention

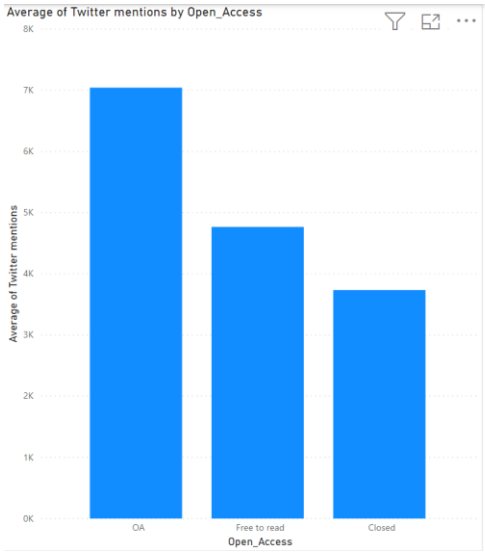
Altmetric Top 100-2019: Data Dictionary

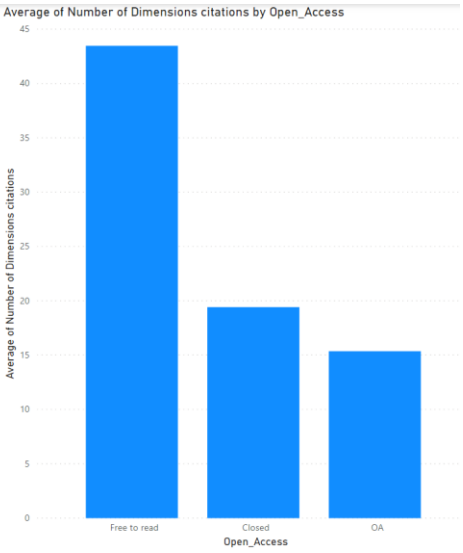
Dataset	Owner	Variable Name	Definition	Format	Uniqueness
Altmetric Top 100-2019	Altmetrics	Altmetric Attention Score	A values based on weights assigned to social media memtions of an article or similar published document, including peer reviews, Wikipedia citations, discussions on research blogs, mainstream media coverage, bookmarks, and mentions on social networks such as Twitter.	Integer	No
		Title	Title of a published paper or other similar document (e.g. editorial)	String	Yes
		Journal/Collection Title	Title of the Journal in which the paper was published	String	No
		Journal ISSNs	International Standard Serial Number: A journal can have more than one ISSN - for example, one for print and one for on-line publication. Some publications do not have an ISSN.	String(s) (empty or comma separated)	No
		Open_Access	Level of public accessibility	String (three possible values: OA, closed, Free to read)	No
		Handle.net IDs	Corporation for National Research Initiatives persistent identifier for information resources	String (Empty if not assigned/known)	Yes
		ADS Bibcode	Astrophysics data system bibliographic code	String (Empty if not assigned/known)	Yes
		arXiv ID	Article identifier scheme used by arXiv - a free distribution service and an open archive for scholarly articles maintained by Cornell University	Float (Empty if not assigned/known)	Yes
		RePEc ID	Research Papers In Economics permanent identifier attributed to a person	No values assigned/known	No
		SSRN	Social Science Research Network	No values assigned/known	No
		URN	Uniform Resource Name	No values assigned/known	No
		PubMed ID	PubMed identifier is a unique integer value (PubMed is an index)	Integer (empty if not assigned/known)	Yes
		PubMedCentral ID	PubMedCentral (full text archive) Identifier	String (Empty if not assigned/known)	Yes
		News mentions	The number of mainstream online news and magazine outlets that reference a research output	Integer	No
		Blog mentions	The number of times a scholarly output has been linked to from a blog	Integer	No
		Policy mentions	The number of times a research output has been cited in policy documents from government bodies or NGOs	Integer	No
		Twitter mentions	Twitter counts are determined by the number of registered users that tweet or retweet a post that links to a trackable scholarly product	Integer	No
		Patent mentions	Number of citations in patents across nine jurisdictions	Integer	No
		Peer review mentions	Number of Pubpeer comments for any paper with a DOI, PubMed ID, or ArXiv ID	Integer	No
		Weibo mentions	Count of mentions on Sina Weibo (no longer done, but historical data kept)	Integer	No
		Facebook mentions	Count of mentions in posts on a set or curated Facebook pages	Integer	No
		Wikipedia mentions	Count of mentions in the References section of Wikipedia pages	Integer	No
		Google+ mentions	Count of mentions on Google+ (no longer done, but historical data kept)	Integer	No
		LinkedIn mentions	Count of mentions on LinkedIn (no longer done, but historical data kept)	Integer	No
		Reddit mentions	Count of posts mentioning research on Reddit	Integer	No
		Pinterest mentions	Count of posts on Pinterest (no longer done, but historical data kept)	Integer	No
		F1000 mentions	Number of mentions in F1000 papers	Integer	No
		Q&A mentions	Number of mentions on Q&A (Stackoverflow)	Integer	No
		Video mentions	Number of mentions in description of Youtube videos	Integer	No
		Syllabi mentions	Count of mentions in the Open Syllabus Project	Integer	No
		Number of Mendeley readers	Count of Mendeley users that have added a research paper to a Mendeley library	Integer	No
		Number of Dimensions citations	Number of citations sourced from Digital Science (Dimension platform)	Integer	No
		Description	Written description of the paper	String	No
		Affiliations (GRID)	Affiliated University, Institution or other Organisation (from Global Research Identifier Database)	String(s) (comma separated)	No
		Publication Date	Date published	Integer (Data represented by 5 digit number based to Jan 1, 1900)	No
		DOI	Digital Object Identifier is a string of numbers, letters and symbols used to permanently identify an article or document and link to it on the web	String	Yes
		DOI URL	An URL (weblink) which is created by adding http://doi.org/ to the DOI	String	Yes
		Details Page URL	URL link to Altmetrics details page for paper	String	Yes
		Publisher	Journal publisher	String	No
		Subjects	Altmetric subject (of paper) classification	String	No
		Authors	Authors of paper, sourced from citation (note that for many authors, it may be abbreviated with et. al.)	String	No
		sharedit	Web link to paper through sharedit	String	Yes
		Publisher-preferred links	Preferred link specified by publisher	Empty	No

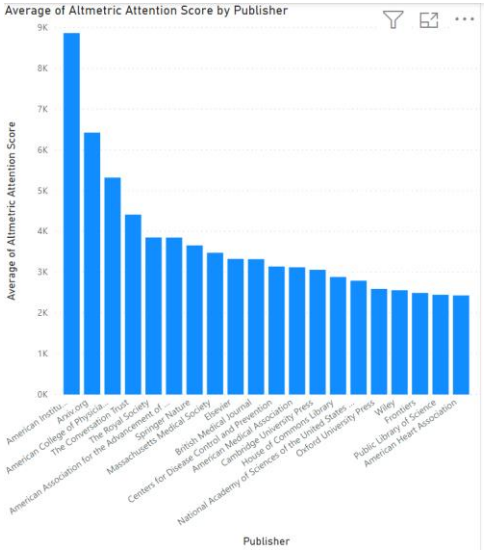

Altmetric Attention Score and F1000 mentions
F1000 mentions
Altmetric Attention Score

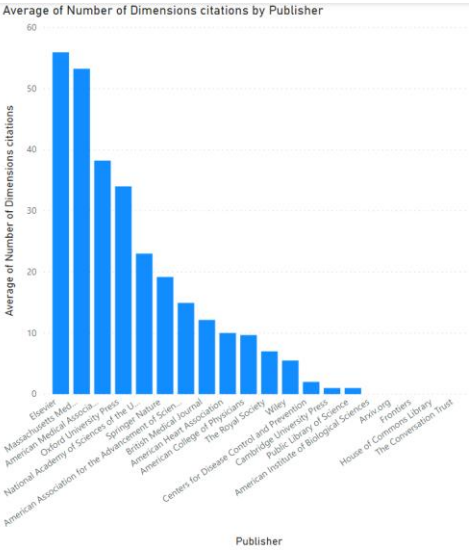

Altmetric Attention Score and Number of Mendeley readers
Number of Mendeley readers
Altmetric Attention Score
Altmetric Attention Score and News mentions
News mentions
Altmetric Attention Score

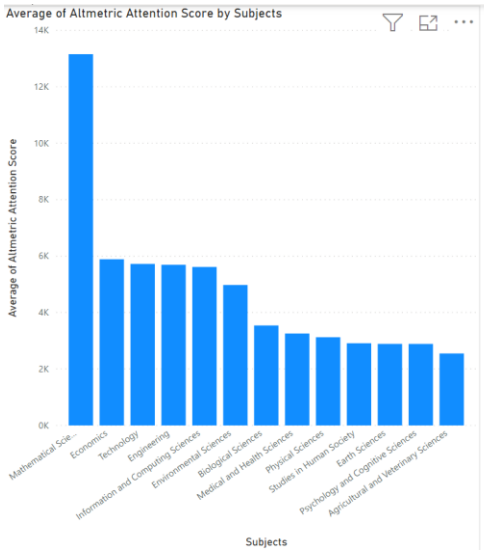

Altmetric Attention Score and Twitter mentions
Twitter mentions
Altmetric Attention Score
Number of Dimensions citations and Number of Mendeley readers
Number of Mendeley readers
Number of Dimensions citations

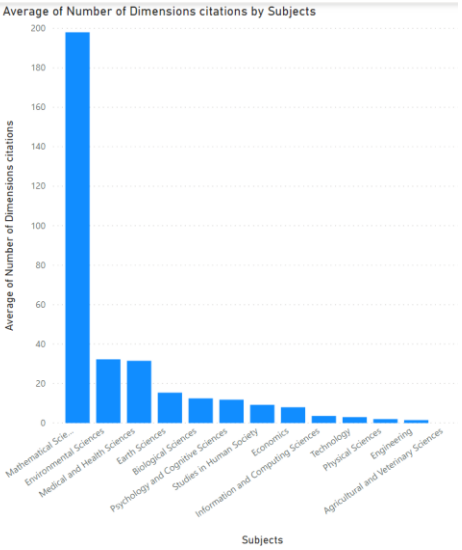

Average of Twitter mentions by Open_Access
Average of Twitter mentions
8K
7K
6K
5K
4K
3K
2K
1K
0K
OA
Free to read
Closed
Open_Access

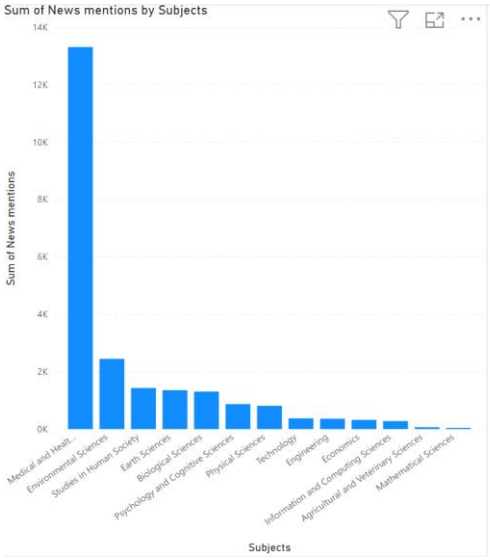

Average of Number of Dimensions citations by Open_Access
Average of Number of Dimensions citations
45
40
35
30
25
20
15
10
5
0
Free to read
Closed
OA
Open_Access

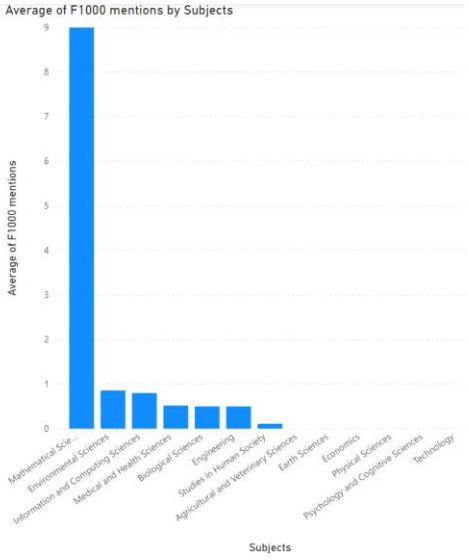

Average of Altmetric Attention Score by Publisher
Average of Altmetric Attention Score
9K
8K
7K
6K
5K
4K
3K
2K
1K
0K
Publisher

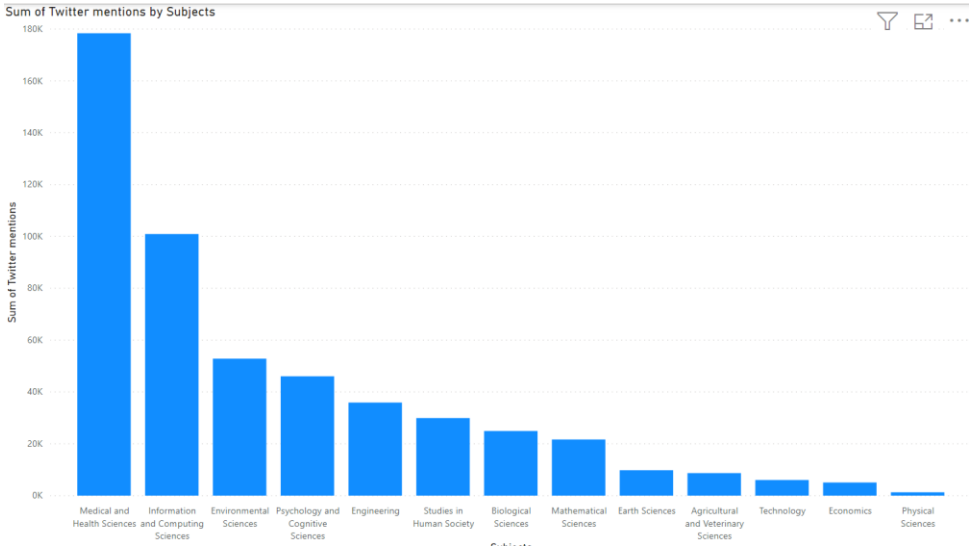

Average of Number of Dimensions citations by Publisher
Average of Number of Dimensions citations
60
50
40
30
20
10
0
Publisher


Average of Altmetric Attention Score by Subjects
Average of Altmetric Attention Score
14K
12K
10K
8K
6K
4K
2K
0K
Subjects


Average of Number of Dimensions citations by Subjects
Average of Number of Dimensions citations
200
180
160
140
120
100
80
60
40
20
0
Subjects


Sum of News mentions by Subjects
Sum of News mentions
14K
12K
10K
8K
6K
4K
2K
0K
Medical and Healt...
Environmental Sciences
Studies in Human Society
Earth Sciences
Biological Sciences
Psychology and Cognitive Sciences
Physical Sciences
Technology
Engineering
Economics
Information and Computing Sciences
Agricultural and Veterinary Sciences
Mathematical Sciences
Subjects


Average of F1000 mentions by Subjects
Average of F1000 mentions
9
8
7
6
5
4
3
2
1
0
Mathematical Scie...
Environmental Sciences
Information and Computing Sciences
Medical and Health Sciences
Biological Sciences
Engineering
Studies in Human Society
Agricultural and Veterinary Sciences
Earth Sciences
Economics
Physical Sciences
Psychology and Cognitive Sciences
Technology
Subjects


Sum of Twitter mentions by Subjects
Sum of Twitter mentions
180K
160K
140K
120K
100K
80K
60K
40K
20K
0K
Medical and Health Sciences
Information and Computing Sciences
Environmental Sciences
Psychology and Cognitive Sciences
Engineering
Studies in Human Society
Biological Sciences
Mathematical Sciences
Earth Sciences
Agricultural and Veterinary Sciences
Technology
Economics
Physical Sciences

Altmetrics 2018 Top 100 Data Review

Data Source

Altmetric Top 100 2019, *The 2018 Altmetric Top 100*, Altmetric, viewed and downloaded 27 March 2020,

<http://www.https://www.altmetric.com/top100/2018/>.

General Data Commentary

- Data cut-off: November 2018
- Data coverage: 365 days to, but excluding, November 2018
- Key Data Measure: Altmetric Score (Top 100 most discussed research)
- **All data is dependent on Altmetric data acquisition and analytical capabilities**
- Altmetrics claims to manually check each piece of discussed research in its list and that it removes news articles
- Altmetrics claims to use a combination of manual and automatic check of the following fields in the data: manually check the following: Open Access status, Publisher, Subject area, Author affiliation and Publication date (first published online)
- Altmetrics claims that it filters for "inorganic (i.e. spam) attention".

Additional Data Commentary

- The columns: Handle.net IDs, ADS Bibcode, arXiv ID, RePEc ID, SSRN, URN, PubMedCentral ID, sharedit and Publisher-preferred links are either empty or largely empty.
- The columns: Patent Mentions, Peer Review Mentions, Weibo Mentions, Linkedin Mentions, Pinterest Mentions, Q@A Mentions and Syllabi Mentions either record only zero or have at most one value in a column.
- Both sets of columns described above should be dropped from the final analysis.
- Some fields available in 2018 have been excluded for 2019
- Open Access values for 2019 correspond to the Open Access values for 2018 but have name differences

- Where data is present in a column, it is consistent within the column.
- Data Provenance is known.
- Data sourcing organisation is well established and known.
- **Data is deemed to be of High Quality.**

Data sourced from Altmetrics:2018 Top 100

- No visualisations have been included for the 2018 Top 100 as the outcomes were similar to those of 2019

Altmetric Top 100-2018: Data Dictionary

Dataset	Owner	Variable Name	Definition	Format	Uniqueness
Altmetric Top 100-2018	Altmetrics	Journal ISSNs	International Standard Serial Number: A journal can have more than one ISSN - for example, one for print and one for on-line publication. Some publications do not have an ISSN.	String(s) (empty or comma separated)	No
		Handle.net IDs	Corporation for National Research Initiatives persistent identifier for information resources	String (Empty if not assigned/known)	Yes
		ADS Bibcode	Astrophysics data system bibliographic code	String (Empty if not assigned/known)	Yes
		arXiv ID	Article identifier scheme used by arXiv - a free distribution service and an open archive for scholarly articles maintained by Cornell University	Float (Empty if not assigned/known)	Yes
		RePEc ID	Research Papers In Economics permanent identifier attributed to a person	No values assigned/known	No
		SSRN	Social Science Research Network	Integer	Yes
		URN	Uniform Resource Name	No values assigned/known	No
		PubMed ID	PubMed identifier is a unique integer value (PubMed is an index)	Integer (empty if not assigned/known)	Yes
		PubMedCentral ID	PubMedCentral (full text archive) Identifier	String (Empty if not assigned/known)	Yes
		News mentions	The number of mainstream online news and magazine outlets that reference a research output	Integer	No
		Blog mentions	The number of times a scholarly output has been linked to from a blog	Integer	No
		Policy mentions	The number of times a research output has been cited in policy documents from government bodies or NGOs	Integer	No
		Twitter mentions	Twitter counts are determined by the number of registered users that tweet or retweet a post that links to a trackable scholarly product	Integer	No
		Patent mentions	Number of citations in patents across nine jurisdictions	Integer	No
		Peer review mentions	Number of Pubpeer comments for any paper with a DOI, PubMed ID, or ArXiv ID	Integer	No
		Weibo mentions	Count of mentions on Sina Weibo	Integer	No
		Facebook mentions	Count of mentions in posts on a set or curated Facebook pages	Integer	No
		Wikipedia mentions	Count of mentions in the References section of Wikipedia pages	Integer	No
		Google+ mentions	Count of mentions on Google+	Integer	No
		LinkedIn mentions	Count of mentions on LinkedIn	Integer	No
		Reddit mentions	Count of posts mentioning research on Reddit	Integer	No
		Pinterest mentions	Count of posts on Pinterest	Integer	No
		F1000 mentions	Number of mentions in F1000 papers	Integer	No
		Q&A mentions	Number of mentions on Q&A (Stackoverflow)	Integer	No
		Video mentions	Number of mentions in description of Youtube videos	Integer	No
		Syllabi mentions	Count of mentions in the Open Syllabus Project	Integer	No
		Number of Mendeley readers	Count of Mendeley users that have added a research paper to a Mendeley library	Integer	No
		Number of Dimensions citations	Number of citations sourced from Digital Science (Dimension platform)	Integer	No
		Rank	Ranking of paper in list. Paper with highest Altmetric score has a value of 1.	Integer	Yes
		Altmetric Attention Score	A values based on weights assigned to social media memtions of an article or similar published document, including peer reviews, Wikipedia citations, discussions on research blogs, mainstream media coverage, bookmarks, and mentions on social networks such as Twitter.	Integer	No
		Title	Title of a published paper or other similar document (e.g. editorial)	String	Yes
		Description	Written description of the paper	String	No
		Journal/Collection Title	Title of the Journal in which the paper was published	String	No
		Publisher	Journal publisher	String	No
		Subjects	Altmetric subject (of paper) classification	String	No
		Subjects (FoR)	Numeric scheme for classification subject(s) of paper	String(s) (comma separated)	No
		Affiliations (GRID)	Affiliated University, Institution or other Organisation (from Global Research Identifier Database)	String(s) (comma separated)	No
		Publication Date	Date published	String (dd/mm/yyyy)	No
		DOI	Digital Object Identifier is a string of numbers, letters and symbols used to permanently identify an article or document and link to it on the web	String	Yes
		DOI URL	An URL (weblink) which is created by adding http://doi.org/ to the DOI	String	Yes
		Details Page URL	URL link to Altmetrics details page for paper	String	Yes
		Badge URL	Link to Altmetric "donut" image for paper	String	Yes
		Open_Access	Level of public accessibility	String (empty or one of three possible values: OA, Not OA, Free to read)	No
		Funders	Supporting Institution/Agnecy	String	No
		Unique countries	Countries of residence for contributing authors	String	No
		Authors	Authors of paper, sourced from citation	String	No
		Publisher-preferred links	Authors of paper, sourced from citation (note that for many authors, it may be abbreviated with et. al.)	String	No

Plum Analytics: PlumX Top 100

Data Source

The PlumX Altmetrics & Sci-Hub Dataset Update, *The PlumX Altmetrics & Sci-Hub Dataset Update*, Plum Analytics, viewed and downloaded 27 March 2020,

<https://www.plumanalytics.com/plumx-altmetrics-sci-hub-downloads-dataset/>.

General Data Commentary

- Data sourced from: https://figshare.com/articles/PlumX_Altmetrics_Sci_Hub_Downloads/3380671
- The dataset is a compilation of PlumX metrics with the top 100 most-downloaded DOIs in the Sci-Hub dataset and was used by Elbakyan and Bohannonin their paper (2016) "Who's downloading pirated papers? Everyone".
- There is no key data mention – the paper focusses on social media mentions across a range of social platforms
- There is one measure for notional academic use - Usage
- **All data is dependent on the authors (Elbakyan and Bohannonin) data acquisition and analytical capabilities in combination with Plum Analytics**

Additional Data Commentary

- The ISBN column is largely empty.

- Where data is present in a column, it is consistent within the column.
- Data provenance is known – sourced by Elbakyan and Bohannonin from Plum Analytics
- Original data sourcing organisation is well established and known.
- **Data is deemed to be of High Quality.**

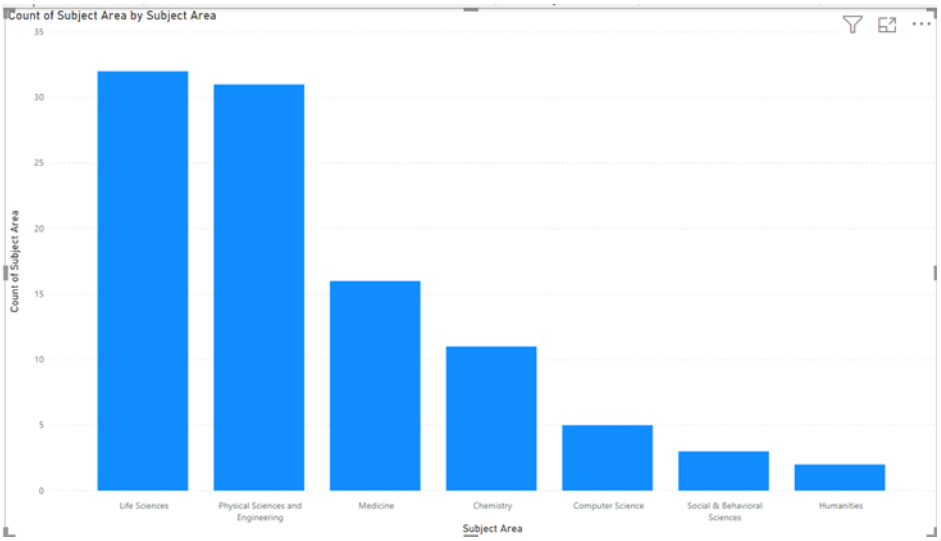
Data commentary from PlumX Altmetrics & Sci-Hub Dataset Update

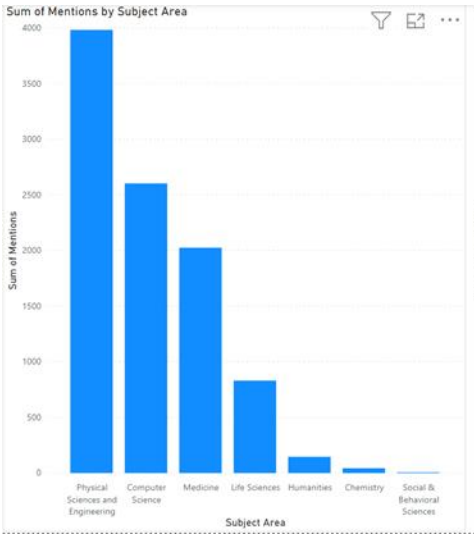
- Strong bias by subject matter by sum and average of measures
- Life Sciences and Physical Sciences & Engineering dominate the PlumX Top 100 papers
- Averages in some measures not strongly differentiated
- Filters were applied for type pf paper published and year of publishing
- No correlation apparent between Usage and Mentions – Usage is a substitute for academic reading and Mentions is a measure of social media attention

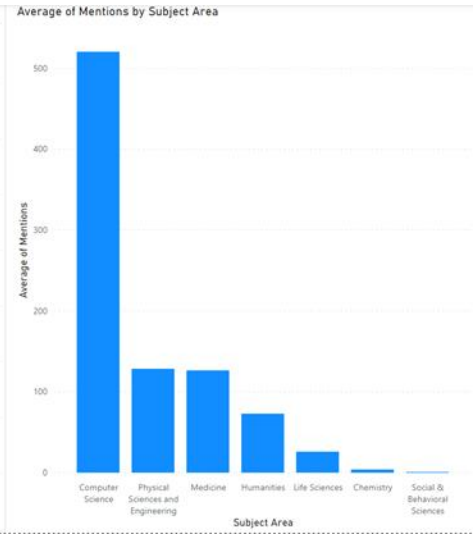
PlumX Altmetrics & Sci-Hub Dataset Update: Data Dictionary*

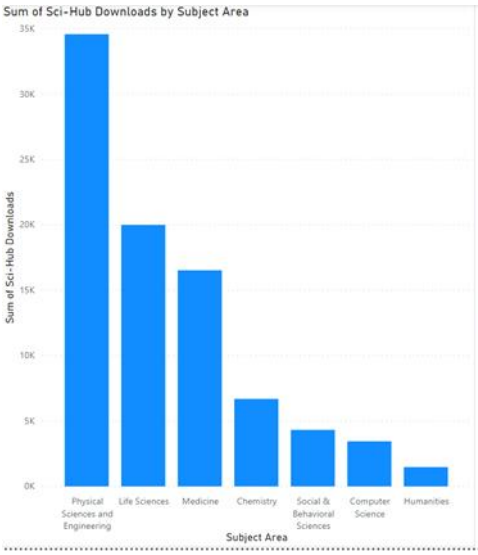
Dataset	Owner	Variable Name	Definition	Format	Uniqueness
The PlumX Altmetrics & Sci-Hub Dataset (Top 100, 2016)	Plum Analytics	DOI	DOI is the unique identifier assigned to each paper on Sci-Hub	String	Yes
		Title	Title of the research paper	String	Yes
		Sci-Hub Downloads	These values were calculated by summing the number of rows for each DOI in the Sci-Hub dataset. The top 100 download counts were then identified and added to https://plu.mx/scihub100.	Integer	No
		Publisher Cost	This is the cost to download a PDF from the publisher's site (as of May 10, 2016).	Float	No
		SciHub x Pub Cost	The publisher cost was multiplied with the Sci-Hub Downloads count to estimate the dollar amount that these downloads may be costing publishers.	Float	No
		Subject Area	Each DOI was investigated and sorted into a general subject area by Noella Natalino, MLIS, Product/Content Manager for Plum Analytics.	String	No
		ISBN	International Standard Book Number	Integer (empty if not assigned/known)	Yes
		PMID	PubMed identifier is a unique integer value (PubMed is an index). 72 of the top 100 DOIs are available via PubMed.	Integer (empty if not assigned/known)	Yes
		Year	The publication year of the research paper	Integer (yyyy)	No
		Type	The type of research paper: Article (62), Review (25), Letter (6), Book Chapter (4), Conference Paper (2), Book (1)	String	No
		Captures	Captures track when end users bookmark, favorite, or save an item for future use. Downloaded May 10, 2016. Visit https://plu.mx/scihub100 for the most recent counts.	Integer	No
		Social Media	Social media metrics are the +1s, likes, shares, and tweets about research. Downloaded May 10, 2016. Visit https://plu.mx/scihub100 for the most recent counts.	Integer	No
		Mentions	Mentions are the blog posts, comments, reviews, and wikipedia links about research. Downloaded May 10, 2016. Visit https://plu.mx/scihub100 for the most recent counts.	Integer	No
		Usage	Usage metrics are the # of clicks, downloads, views, and library holdings for research. Downloaded May 10, 2016. Visit https://plu.mx/scihub100 for the most recent counts.	Integer	No

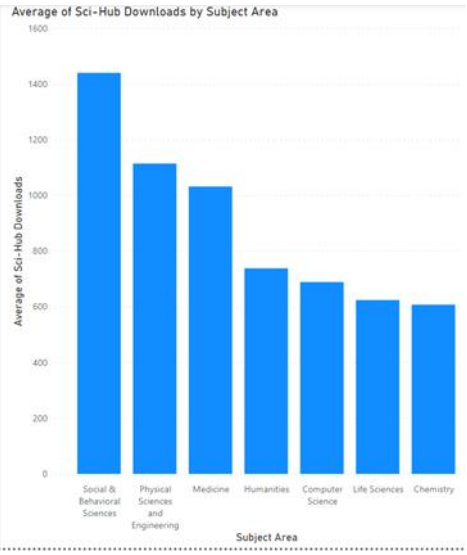
*This Data Dictionary has been adapted from the data dictionary supplied with the dataset.

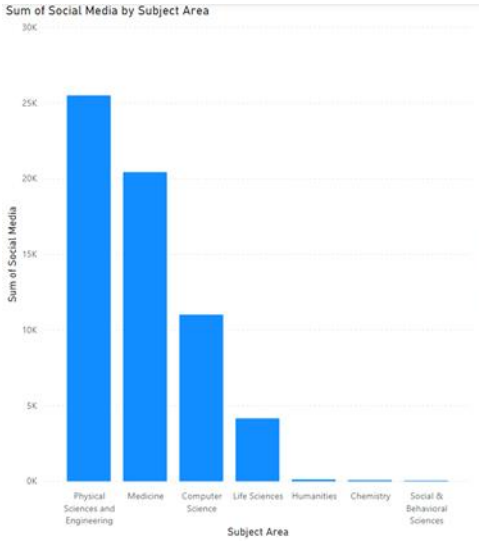

Count of Subject Area by Subject Area
Count of Subject Area
35
30
25
20
15
10
5
0
Life Sciences
Physical Sciences and Engineering
Medicine
Chemistry
Computer Science
Social & Behavioral Sciences
Humanities
Subject Area

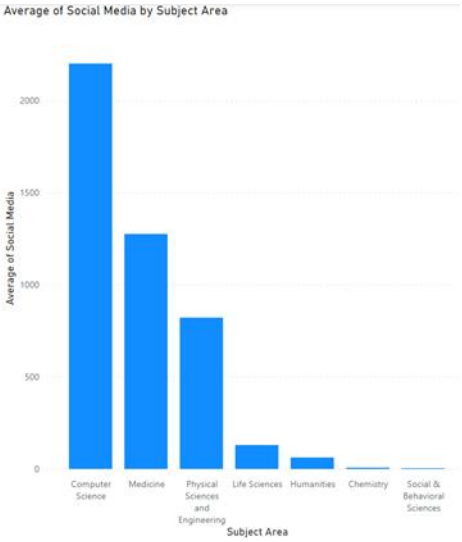

Sum of Mentions by Subject Area
Sum of Mentions
4000
3500
3000
2500
2000
1500
1000
500
0
Physical Sciences and Engineering
Computer Science
Medicine
Life Sciences
Humanities
Chemistry
Social & Behavioral Sciences
Subject Area

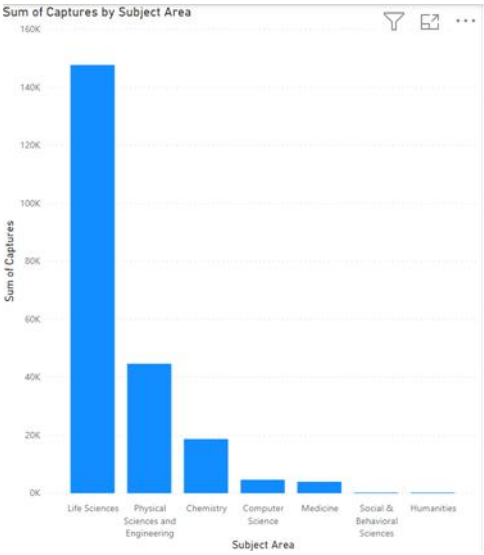

Average of Mentions by Subject Area
Average of Mentions
500
400
300
200
100
0
Computer Science
Physical Sciences and Engineering
Medicine
Humanities
Life Sciences
Chemistry
Social & Behavioral Sciences
Subject Area

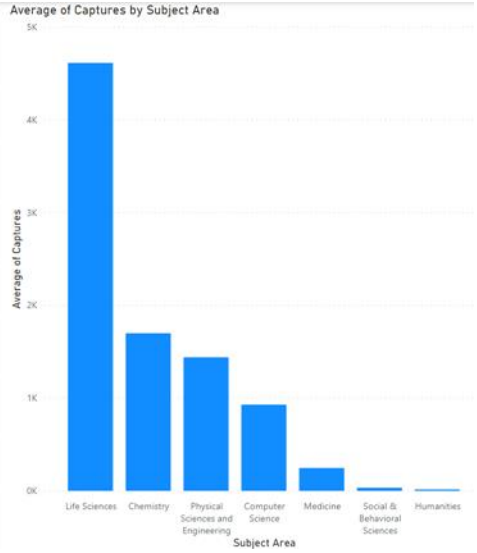

Sum of Sci-Hub Downloads by Subject Area
Sum of Sci-Hub Downloads
35K
30K
25K
20K
15K
10K
5K
0K
Physical Sciences and Engineering
Life Sciences
Medicine
Chemistry
Social & Behavioral Sciences
Computer Science
Humanities
Subject Area

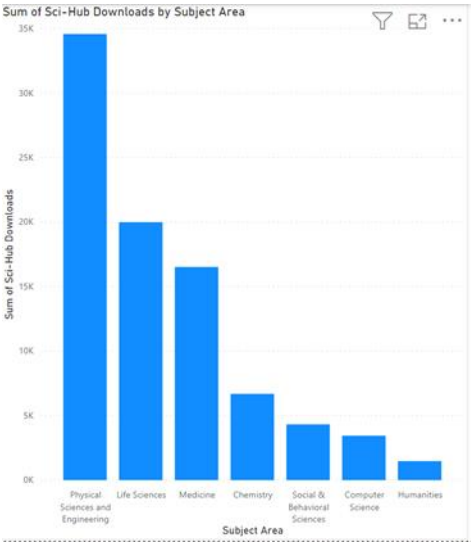

Average of Sci-Hub Downloads by Subject Area
Average of Sci-Hub Downloads
1600
1400
1200
1000
800
600
400
200
0
Social & Behavioral Sciences
Physical Sciences and Engineering
Medicine
Humanities
Computer Science
Life Sciences
Chemistry
Subject Area

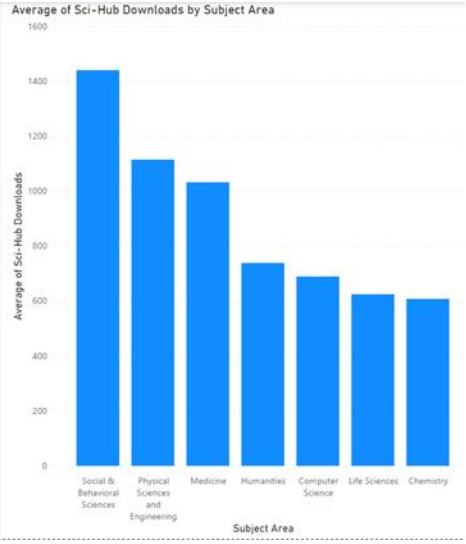

Sum of Social Media by Subject Area
30K
25K
20K
15K
10K
5K
0K
Sum of Social Media
Physical Sciences and Engineering
Medicine
Computer Science
Life Sciences
Humanities
Chemistry
Social & Behavioral Sciences
Subject Area

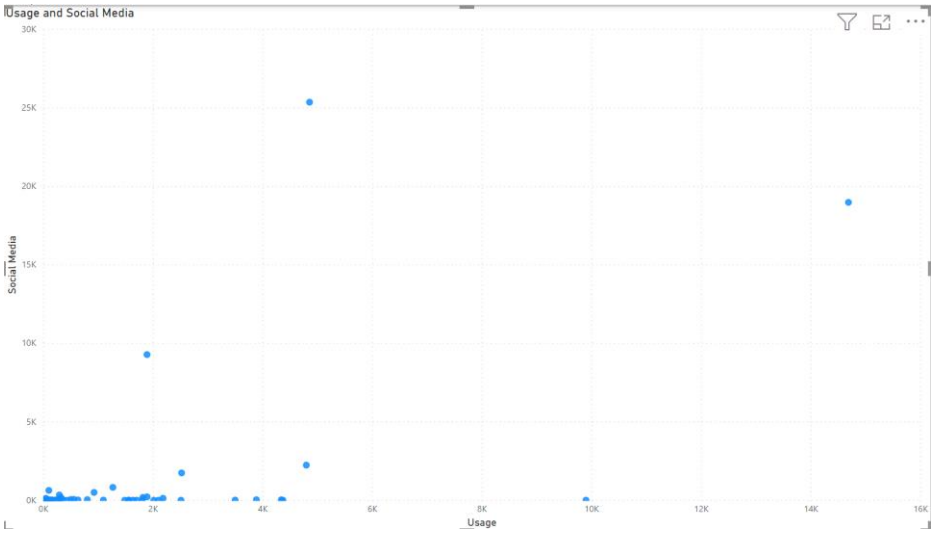

Average of Social Media by Subject Area
2000
1500
1000
500
0
Average of Social Media
Computer Science
Medicine
Physical Sciences and Engineering
Life Sciences
Humanities
Chemistry
Social & Behavioral Sciences
Subject Area


Sum of Captures by Subject Area
160K
140K
120K
100K
80K
60K
40K
20K
0K
Sum of Captures
Life Sciences
Physical Sciences and Engineering
Chemistry
Computer Science
Medicine
Social & Behavioral Sciences
Humanities
Subject Area


Average of Captures by Subject Area
5K
4K
3K
2K
1K
0K
Average of Captures
Life Sciences
Chemistry
Physical Sciences and Engineering
Computer Science
Medicine
Social & Behavioral Sciences
Humanities
Subject Area


Sum of Sci-Hub Downloads by Subject Area
35K
30K
25K
20K
15K
10K
5K
0K
Sum of Sci-Hub Downloads
Physical Sciences and Engineering
Life Sciences
Medicine
Chemistry
Social & Behavioral Sciences
Computer Science
Humanities
Subject Area


Average of Sci-Hub Downloads by Subject Area
1600
1400
1200
1000
800
600
400
200
0
Average of Sci-Hub Downloads
Social & Behavioral Sciences
Physical Sciences and Engineering
Medicine
Humanities
Computer Science
Life Sciences
Chemistry
Subject Area


Usage and Social Media
30K
25K
20K
15K
10K
5K
0K
Social Media
0K
2K
4K
6K
8K
10K
12K
14K
16K
Usage

Data sourced from Bormann and Haunschild paper "Do altmetrics correlate with the quality of papers? A large-scale empirical study based on F1000Prime data"

Data Source

Bornmann, L & Haunschild, R 2018, 'Do altmetrics correlate with the quality of papers? A large-scale empirical study based on F1000Prime data.', *PLoS ONE*, vol. 13, no. 5, p. e0197133.

General Data Commentary

- Data for the paper sourced from: https://figshare.com/articles/Do_altmetrics_correlate_with_the_quality_of_papers_-_underlying_data/6120158
- The dataset is a compilation of PlumX metrics with the top 100 most-downloaded DOIs in the Sci-Hub dataset and was used by Elbakyan and Bohannonin their paper (2016) "Who's downloading pirated papers? Everyone".
- Data set comprises ~32000 rows of data based on research papers published between 2011 and 2013
- All rows contain complete tuples
- **All data is dependent on the authors (Bornmann and Haunschild) data acquisition and analytical capabilities in combination with their sources**

Additional Data Commentary

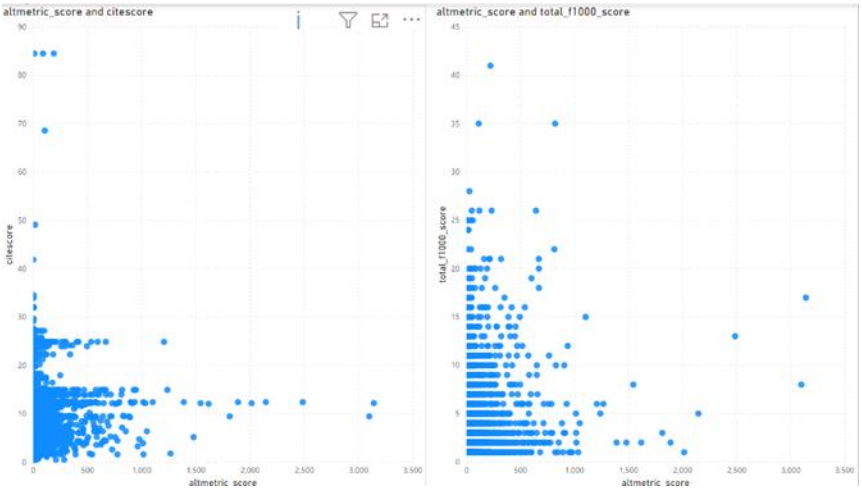
- Data is consistent within the columns.
- Data provenance is known – sourced by Bornmann and Haunschild from Altmetric, Web of Science, Scopus, Clarivate Analytics and F1000.
- The authors engaged in data cleansing and matching to ensure that all fields were filled.
- The authors are both well-established researchers in the field of scientometrics.
- **Data is deemed to be of High Quality.**

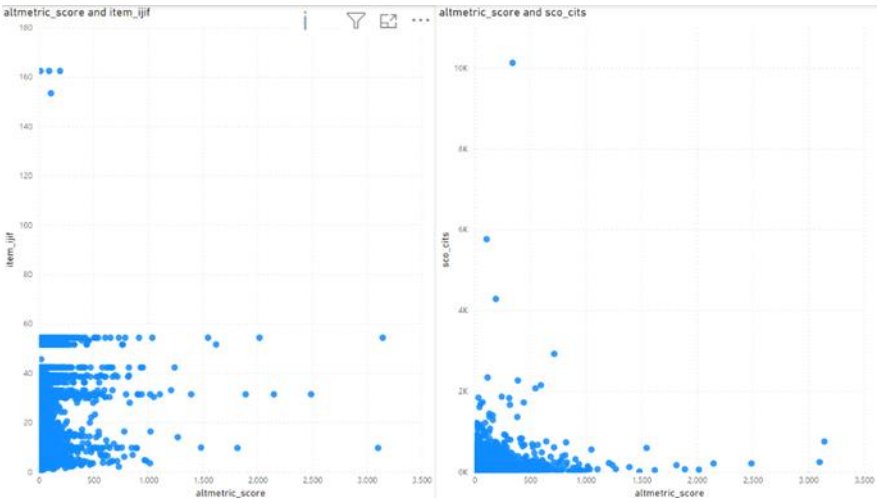
Data commentary from "Do altmetrics correlate with the quality of papers? A large-scale empirical study based on F1000Prime data"

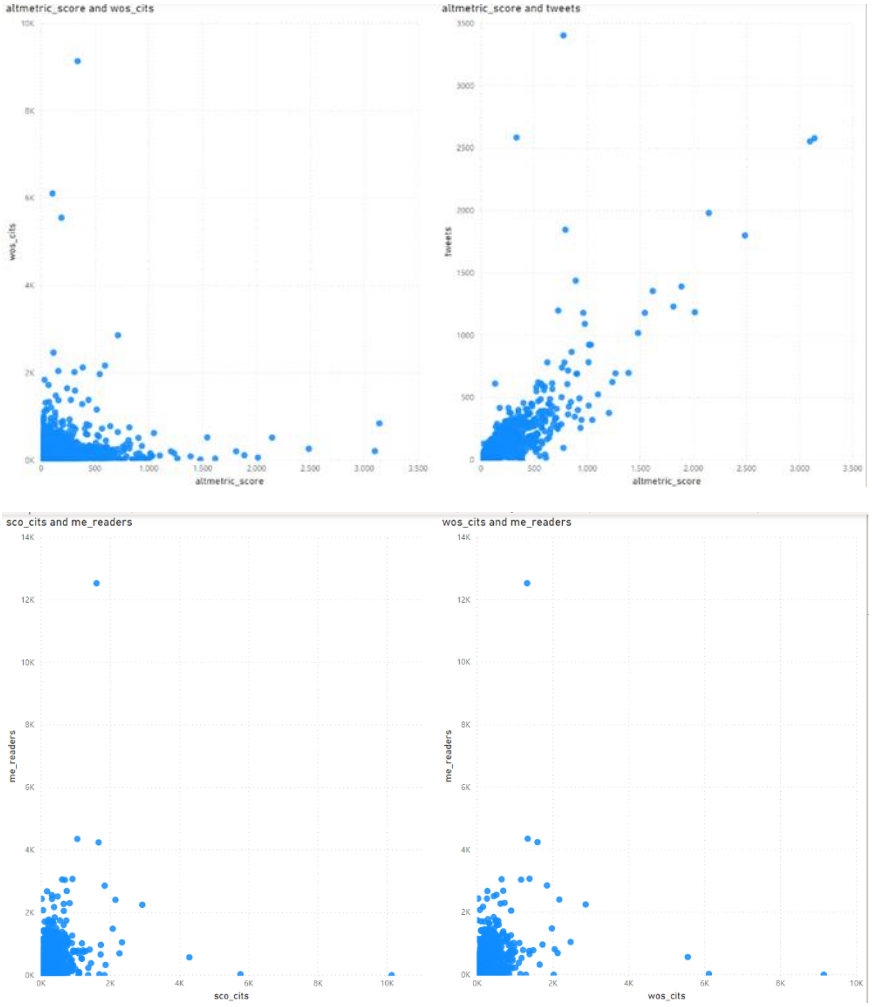
- Low correlation between Altmetric score and citation measures
- High correlation (relatively) between Altmetric score and tweets
- Evidence of correlation between Mendeley readership and WoS and Sco measures is limited – Bornmann and Huanschild concluded (using factor analysis) that Mendeley readership may be of value as a measure of quality
- Bornmann and Huanschild concluded (using factor analysis) that social media measures may not be of value as a measure of quality

“Do altmetrics correlate with the quality of papers? A large-scale empirical study based on F1000Prime data”: Data Dictionary*

Dataset	Owner	Variable Name	Definition	Format	Uniqueness
Supporting data for the Paper "Do altmetrics correlate with the quality of papers? A large-scale empirical study based on F1000Prime data"	Lutz Bornmann and Robin Haunschild (compiled from various, but trusted, sources)	pubyear	Year of publication for paper: Papers from 2011-2013 in F1000; Papers with a DOI excluded. Papers not referenced by Altmetrics excluded.	Integer (yyyy)	No
		wos_cits_3	Lagged (over 3 years) and smoothed Web of Science citation counts	Integer	No
		wos_cits	Web of Science citation count for paper	Integer	No
		sco_cits_3	Lagged (over 3 years) and smoothed Scopus citation counts	Integer	No
		sco_cits	Scopus citation count for paper	Integer	No
		tweets	Altmetrics tweet counts for paper	Integer	No
		me_readers	Altmetrics Mendeley counts for paper	Integer	No
		altmetric_score	Altmetrics score for paper	Float	No
		citescore	CiteScore citation score	Float	No
		item_ijif	Clarivate Analytics Journal Impact Factor	Float	No
		total_f1000_score	Sum of F1000 evaluation scores	Integer	No






altmetric_score and wos_cits
wos_cits
altmetric_score
altmetric_score and tweets
tweets
altmetric_score
sco_cits and me_readers
me_readers
sco_cits
wos_cits and me_readers
me_readers
wos_cits