NISO RP-25-201X-3

Altmetrics Data Quality Code of Conduct

For public review and comment from
February 25 – March 31, 2016

A Recommended Practice of the
National Information Standards Organization

DRAFT FOR PUBLIC COMMENT

About Recommended Practices

A NISO Recommended Practice is a recommended "best practice" or "guideline" for methods, materials, or practices in order to give guidance to the user. Such documents usually represent a leading edge, exceptional model, or proven industry practice. All elements of Recommended Practices are discretionary and may be used as stated or modified by the user to meet specific needs.

This recommended practice may be revised or withdrawn at any time. For current information on the status of this publication contact the NISO office or visit the NISO website (www.niso.org).

Published by
National Information Standards Organization (NISO)
3600 Clipper Mill Road
Suite 302
Baltimore, MD 21211
www.niso.org

ISBN: to be added at publication

Contents

1 Introduction 1

1.1 Purpose and Scope 1

1.2 Terms and Definitions 1

2 Recommendations 2

2.1 Transparency 2

2.2 Replicability 2

2.3 Accuracy 2

3 Annual Report 3

Appendix A: NISO Altmetrics Working Group C "Data Quality" Code of Conduct Self-Reporting Table 4

Appendix B : NISO Altmetrics Working Group C "Data Quality" Code of Conduct Self-Reporting Table: Samples 7

Bibliography 37

Foreword

About this Recommended Practice

Altmetrics are increasingly being used and discussed as an expansion of the tools available for measuring the scholarly impact of research in the knowledge environment. The NISO Alternative Assessment Metrics Project was begun in July 2013 with funding from the Alfred P. Sloan Foundation to address several areas of limitations and gaps that hinder the broader adoption of altmetrics. This document is one output from this project, intended to help organizations that wish to use altmetrics to ensure their consistent application across the community. “Working Group C” studied and discussed issues of data quality in the altmetrics realm, an essential aspect of evaluation before metrics can be used for research and practical purposes.

Additional working group outputs from this initiative in the areas of definitions, use cases, specific output types and use of persistent identifiers will be released soon for public comment.

NISO Business Information Topic Committee

The Business Information Topic Committee had the following members at the time it approved this Recommended Practice:

[to be added by NISO after approval]

NISO Altmetrics Working Group C Members

The following individuals served on the NISO Altmetrics Working Group C, which developed and approved this Recommended Practice:

Euan Adie
Altmetric

Scott Chamberlain
rOpenSci

Tilla Edmunds
Thomson Reuters

Martin Fenner
DataCite

Gregg Gordon
Social Science Research Network (SSRN)

Stefanie Haustein (co-chair)
Université de Montréal

Kornelia Junge
John Wiley & Sons, Ltd.

Stuart Maxwell
Scholarly iQ

Angelia Ormiston
Johns Hopkins University Press

Maria Stanton
American Theological Library Association (ATLA)

Greg Tananbaum (co-chair)
Scholarly Publishing and Academic Resources Coalition (SPARC)

Joe Wass
Crossref

Zhiwu Xie
Virginia Tech University Libraries

Zohreh Zahedi
Centre for Science and Technology Studies, University of Leiden

1 Introduction

1.1 Purpose and Scope

The Code of Conduct aims to improve the quality of altmetric data by increasing the transparency of data provision and aggregation as well as ensuring replicability and accuracy of online events used to generate altmetrics. It is not concerned with the meaning, validity, or interpretation of indicators derived from that data. Altmetrics are based on online events "derived from activity and engagement between diverse stakeholders and scholarly outputs in the research ecosystem," as defined in the forthcoming NISO Recommended Practice, *Altmetrics Definitions and Use Cases* (NISO-RP-25-201X-1).

1.2 Terms and Definitions

Term	Definition
altmetric data providers	Platforms that function as sources of online *events* used as altmetrics (e.g., Twitter, Mendeley, Facebook, F1000Prime, Github, SlideShare, Figshare). The working group is aware that not all altmetric data providers—Twitter and Facebook, for example—are part of the scholarly communication community.
altmetric data aggregators	Tools and platforms that aggregate and offer online *events* as well as derived *metrics* from altmetric data providers (e.g., Altmetric.com, Plum Analytics, PLOS ALM, ImpactStory, Crossref).
transparency	The degree to which information and details about the provided data are clear, well-documented, and open to all users (human and machine) for verification
replicability	The degree to which a set of data is consistent across providers and aggregators and over time
accuracy	The degree to which the collected data reflects the material it claims to describe

2 Recommendations

2.1 Transparency

Altmetric data providers are encouraged, and altmetric data aggregators are expected to be **transparent** by offering information about:

- how data are generated, collected, and curated (T1);
- how data are aggregated, and derived data generated (T2);
- when and how often data are updated (T3);
- how data can be accessed (T4);
- how data quality is monitored (T5).

2.2 Replicability

Altmetric data providers are encouraged, and altmetric data aggregators are expected to offer **replicable** data by ensuring that:

- the provided data is generated using the same methods over time (R1);
- changes in methods and their effects are documented (R2);
- changes in the data following corrections of errors are documented (R3);
- data provided to different users at the same time is identical or, if not, differences in access provided to different user groups are documented (R4);
- information is provided on whether and how data can be independently verified (R5).

2.3 Accuracy

Altmetric data providers are encouraged, and altmetric data aggregators are expected to offer **accurate** data by ensuring that:

- the data represents what it purports to reflect (A1);
- known errors are identified and corrected (A2);
- any limitations of the provided data are communicated (A3).

3 Annual Report

By following the Code of Conduct **altmetric data providers** and **altmetric data aggregators** agree to provide a publicly available annual report documenting in detail how they adhere to the recommendations above. The report should follow the standard format provided in the self-reporting table (see Appendix A) which complements the recommendations of the Code of Conduct and includes sample reports (see Appendix B) for a selection of altmetric data providers and aggregators.

Appendix A
NISO Altmetrics Working Group C "Data Quality"
Code of Conduct Self-Reporting Table

This is the standard format for the self-reporting table to document compliance to the Code of Conduct (CoC) proposed by the NISO Altmetrics Working Group C: Data Quality.

Altmetric data providers are encouraged and altmetric data aggregators are expected to document the manner in which they follow each of the data quality recommendations listed in the CoC (T1-5, R1-5, A1-3). These items support particular CoC recommendations. Annual documentation must be provided publicly by filling out the "Aggregator / Provider Submission" for items #1-13 (see below) by aggregators and #1-11 and #13 by providers.

No field should be left blank. If a provider cannot submit the requested information, each element that cannot be provided should be stated. Annual updates of the self report need to be provided publicly by altmetric data providers and aggregators that claim CoC compliance. Reports from previous years should be archived to document CoC compliance over time.

The CoC self-reporting table includes examples of altmetric data aggregator and altmetric data provider submissions as identified by the NISO Altmetrics Working Group C: Data Quality. Examples include Altmetric.com, Crossref DET, and PLOS ALM (Public Library of Science Article-level Metrics) for altmetric data aggregators and Facebook, Mendeley, Twitter, and Wikipedia for altmetric data providers. These examples are subject to change. They are not necessarily complete but are meant to support altmetric data aggregators and providers when submitting their responses for each of the listed items.

Item	Description	Supports CoC Recommendation	Aggregator / Provider Submission	Last update of self-reporting table
#1	List all available data and metrics (providers and aggregators) and altmetric data providers from which data are collected (aggregators).	T1	*To be filled out by data aggregator / provider*	

#2	Provide a clear definition of each metric.	A1	*To be filled out by data aggregator / provider*	
#3	Describe the method(s) by which data are generated or collected and how data are maintained over time.	T1, T2, R1	*To be filled out by data aggregator / provider*	
#4	Describe all known limitations of the data.	A3	*To be filled out by data aggregator / provider*	
#5	Provide a documented audit trail of how and when data generation and collection methods change over time and list all known effects of these changes. Documentation should note whether changes were applied historically or only from change date forward.	R1, R2, R3	*To be filled out by data aggregator / provider*	
#6	Describe how data are aggregated.	T2	*To be filled out by data aggregator / provider*	
#7	Detail how often data are updated.	T3	*To be filled out by data aggregator / provider*	
#8	Describe how data can be accessed.	T4	*To be filled out by data aggregator / provider*	

#9	Confirm that data provided to different data aggregators and users at the same time are identical and, if not, how and why they differ.	R4	*To be filled out by data aggregator / provider*	
#10	Confirm that all retrieval methods lead to the same data and, if not, how and why they differ.	R4	*To be filled out by data aggregator / provider*	
#11	Describe the data-quality monitoring process.	T5, A2	*To be filled out by data aggregator / provider*	
#12	Provide a process by which data can be independently verified.	R5	*To be filled out by data aggregator*	
#13	Provide a process for reporting and correcting data or metrics that are suspected to be inaccurate.	A2	*To be filled out by data aggregator / provider*	

Appendix B
NISO Altmetrics Working Group C "Data Quality"
Code of Conduct Self-Reporting Table: Samples*

(This appendix is not part of the ANSI/NISO RP-25-201X-3 *Altmetrics Data Quality Code of Conduct*. It is included for information only. Note also that the following data were collected by the NISO Altmetrics Working Group C for the purposes of this Recommended Practice. They were not self-reported by the companies or organizations in question.)

NISO Altmetrics Working Group C "Data Quality" – Code of Conduct Self-Reporting Table

Example for data aggregator: Altmetric.com

Item	Description	Supports CoC Recommendation	**Aggregator / Provider Submission***	**Last update of self-reporting table****
#1	List all available data and metrics (providers and aggregators) and altmetric data providers from which data are collected (aggregators).	T1	Altmetric collects data from: Twitter, Facebook, Google+, policy documents, mainstream media, blogs, Mendeley, CiteULike, PubPeer, Publons, Reddit, Wikipedia, sites running Stack Exchange (Q&A), reviews on F1000, and YouTube. More details can be found on our Support page: http://bit.ly/1SXDl4j	2016/02/05

#2	Provide a clear definition of each metric.	A1	The Altmetric score of attention is a weighted algorithm providing an indicator of the amount of attention a particular piece of research output has received. Full details on how the score is calculated can be found here: http://www.altmetric.com/blog/scoreanddonut/ Altmetric tools also provide the raw mention counts by source, e.g., the number of posts we have seen about a specific research output on Google+. Raw counts can be viewed in the application, e.g., in the Altmetric Details Page, or exported for further analyses.	2016/02/05
#3	Describe the method(s) by which data are generated or collected and how data are maintained over time.	T1, T2, R1	Data are collected via a range of methods, largely via data provider APIs, third-party provider APIs, text mining and RSS feeds. More information on collection methods by source can be found on our Support page: http://bit.ly/1SXDl4j	2016/02/05
#4	Describe all known limitations of the data.	A3	Altmetric started tracking attention to research across sources in January 2012 and the data collected on articles published before this date is likely to be incomplete. In order to track attention to an output it must have a unique identifier that is supported in our system, e.g., Digital Object Identifier (DOI), arXiv ID, or International Standard Book Number (ISBN), and be hyperlinked or mentioned by journal, author, and date in order to be collected by our text-mining modules operating across news and policy sources. Links to original posts may break, or posts be deleted. We track public pages only, e.g., public Facebook posts, and cannot access private accounts.	2016/02/05

#5	Provide a documented audit trail of how and when data generation and collection methods change over time and list all known effects of these changes. Documentation should note whether changes were applied historically or only from change date forward.	R1, R2, R3	Altmetric does not have an audit trail before January 2016.	2016/02/05
#6	Describe how data are aggregated.	T2	Online events about research outputs are aggregated and mapped by their external persistent identifiers, e.g., DOI, Handle, PubMed Identifier (PMID), arXiv ID.	2016/02/05
#7	Detail how often data are updated.	T3	Update frequency differs across data sources—from real-time to daily. More details on update frequency by source can be found on our Support page: http://bit.ly/1SXDl4j	2016/02/05
#8	Describe how data can be accessed.	T4	Altmetric provides access to the data via end-user interfaces, the Altmetric Application Programming Interface (API), or by providing a snapshot of the data set made available upon request to organizations or individuals for research purposes. Our API documentation is open and available here: http://api.altmetric.com	2016/02/05
#9	Confirm that data provided to different data aggregators and users at the same time are identical and, if not, how and why they differ.	R4	All Altmetric applications are based on the same database. Users access the same data across each tool, except where data are cached and restricted according to access level. Access level varies across products. Explorer for Publishers, Explorer for Institutions, Explorer for Funders, Altmetric Badges, and the Altmetric Commercial API require a subscription to access all data. The Altmetric Bookmarklet, Institutional Repository Badges, Explorer for Academic Librarians, and the Researcher API are free tools that provide access to all mentions. More details can be found on our Products page: http://www.altmetric.com/products/.	2016/02/05

			The article report pages seen within the Altmetric Explorer product or when the Altmetric Badges are clicked on are cached for 60 minutes by the content delivery network we use. Therefore, it is possible that a change to an output that appears in the API results immediately will not be reflected in the relevant article report page for up to an hour. The article report pages seen within the Altmetric Explorer product or when the Altmetric Badges are clicked on are cached for 60 minutes by the content delivery network we use. Therefore, it is possible that a change to an output that appears in the API results immediately will not be reflected in the relevant article report page for up to an hour.	
#10	Confirm that all retrieval methods lead to the same data and, if not, how and why they differ.	R4	Different retrieval methods will lead to the same data as all Altmetric applications use the same underlying database and API. However, the article report pages seen within the Altmetric Explorer product or when the Altmetric Badges are clicked on are cached for 60 minutes by the content delivery network we use (Fastly). Therefore, it is possible that a change to an output that appears in the API results immediately will not be reflected in the relevant article-report page for up to an hour.	2016/02/05
#11	Describe the data-quality monitoring process.	T5, A2	Data quality is monitored in a range of ways: by manually curating sources; monitoring potential gaming and spammy posts; setting thresholds to automatically flag suspicious activity, such as rate of change in attention for an output; creating suspicious-person profiles; and manually monitoring Altmetric staff's alerts and reported issues. Regular data clean-up tasks are also run, e.g., cross-referring data accuracy against external sources such as Crossref.	2016/02/05
#12	Provide a process by which	R5	See item #8—the tools and services provided by Altmetric use the	2016/02/05

	data can be independently verified (aggregators only).		API documented at http://api.altmetric.com	
#13	Provide a process for reporting and correcting data or metrics that are suspected to be inaccurate.	A2	Suspected inaccurate metrics or data can be reported to support@altmetric.com and via our Support portal: help.altmetric.com. Missed mentions can be reported via an online form: www.surveymonkey.com/s/missedmentions. All Altmetric Details Pages include a "What is this page?" message to provide opportunities for reporting data errors and linking to the Missed Mentions form. The page also provides an introduction to Altmetric data.	2016/02/05

NISO Altmetrics Working Group C "Data Quality" – Code of Conduct Self-Reporting Table

Example for data provider: Crossref DOI Event Tracking (DET)

Crossref DET (name to be confirmed) is a new service by Crossref that will launch during 2016. Openness is at the core of the design of DET. Crossref is working towards abiding by the Altmetrics Data Quality Code of Conduct as it moves toward the launch of DET.

Item	Description	Supports CoC Recommendation	Aggregator / Provider Submission*	Last update of self-reporting table**
#1	List all available data and metrics (providers and aggregators) and altmetric data providers from which data are collected (aggregators).	T1	DET is a platform for collecting event data. The data are gathered through a combination of actively collecting data from non-scholarly sources and allowing scholarly sources to send data. It focuses on events ("these things happened") not aggregations ("this many things happened") or metrics ("you got this score"). At launch Crossref DET will include: • **Links from Crossref DOIs to DataCite DOIs**. These are dataset citations made by publishers that indicate when the metadata for an article cites a dataset via Crossref. • **Links from DataCite DOIs to Crossref DOIs**. These are article citations made by dataset publishers that indicate in the metadata for a dataset that the dataset is linked to a Crossref DOI, via DataCite. • **Twitter DOI mentions**. These are tweets that mention an article or dataset by its DOI, or via the landing page of the DOI. It applies to DOIs that belong to Crossref and DataCite. The data are supplied by Twitter and filtered by Crossref DET. • **Wikipedia DOI citations and uncitations**. These are edits to Wikipedia pages that mention a DOI directly, or edits that remove such mentions. The data are supplied by Wikipedia and filtered by Crossref DET. • **Data supplied by other providers**. We allow data providers to supply us with individual events concerning DOIs. We are	2016/02/05

			working with a prominent player in the scholarly space. Every event, such as “this DOI was annotated” is recorded. The data are sent directly from the provider. • **Facebook**. Number of “shares,” “likes” and “comments” for a given DOI, as retrieved from the Facebook API.	
#2	Provide a clear definition of each metric.	A1	Crossref DET reports raw events, not metrics. The following events are provided: • **Links from Crossref DOIs to DataCite DOIs**. Crossref is the central linking hub for scholarly communications. Publishers deposit metadata about articles as they are published. This includes links to datasets via DataCite. • **Links from DataCite DOIs to Crossref DOIs**. Researchers deposit scholarly research objects for citation to DataCite. Researchers deposit datasets and provide links to scholarly works via Crossref DOIs. • **Twitter DOI mentions**. People discuss scholarly works via their DOIs, or the landing pages to which those DOIs resolve. Crossref works with the Twitter data source, filtering Crossref and DataCite DOIs and corresponding landing pages. • **Wikipedia DOI citations and uncitations**. Wikipedia pages are edited on a constant basis. A page can reference a DOI, and an edit to a page can introduce or remove a link to a DOI. Crossref tracks when these events happen and records when a DOI is added or removed from a page, the DOI, and the page and revision numbers. • **Data supplied by other providers**. Providers are able to push events, such as a DOI is annotation or download, into the DET service. The content of the event is dependent on the type of source. DET will make the event available verbatim. Events are supplied by the party that generated them. • **Facebook**. Facebook Graph API allows DET to query for every DOI it knows about and record how many times a DOI was shared, liked, and commented on. Each time this data	2016/02/05

			are collected is treated as an event.	
#3	Describe the method(s) by which data are generated or collected and how data are maintained over time.	T1, T2, R1	• **Links from Crossref DOIs to DataCite DOIs**. CrossRef identifies deposits and updates DET when it sees a DataCite DOI cited. This will happen in bulk for historical data, and will then be completed live as new deposits are made. • **Links from DataCite DOIs to Crossref DOIs**. DataCite identifies deposits and updates DET when it sees a Crossref DOI cited. This will happen in bulk for historical data, and then will be done live as new deposits are made. • **Twitter DOI mentions**. Crossref DET subscribes to the Twitter firehose, filtering it by Crossref and DataCite DOIs and those domains that DOIs resolve to. It stores all tweets that mention DOIs. For tweets that mention article or dataset landing pages, DET will attempt to identify the corresponding DOI and record that link (including both the DOI and the landing page URL). Not all landing pages URLs can be mapped to DOIs, but if a new technique enables a previously unknown mapping for a historical tweet, this event will be raised. The firehose is a live stream. • **Wikipedia DOI citations and uncitations**. Crossref DET subscribes to the Wikipedia live stream of edits. For every edit that is made to any Wikipedia article, DET will analyze the content of the edit and look for DOIs having been added or removed. An event will be recorded for either the adding or removal of a DOI in a Wikipedia page. The edit stream is live and produces a live stream of events. • **Data provided by other providers**. Crossref DET provides a "Push API" that enables data sources to push data into DET. Providers can push data in batches or live. This is a generic capability, but allows for significant players in the scholarly space to publish DOI event data. • **Facebook**: The Facebook API is queried for every DOI that belongs to Crossref or DataCite. The results are stored directly. The Facebook API is queried periodically. There are no guarantees about how often the Facebook API is queried	2016/02/05

			as this depends on practical issues of scalability.	
#4	Describe all known limitations of the data.	A3	• **Links from Crossref DOIs to DataCite DOIs**. Publishers must provide data. Crossref has around 5,000 publisher members and there are some variabilities among them. • **Links from DataCite DOIs to Crossref DOIs**. Researchers must provide data to DataCite. • **Twitter DOI mentions**. All DOIs in tweets can be reliably identified. In the case of landing pages, Crossref DET will make a best effort to resolve the landing pages, but there is no 100 percent reliable way to do this. • **Wikipedia DOI citations and uncitations**. The Wikipedia live stream or supporting infrastructure may become unavailable. If this happens, those events will be missed. • **Data provided by other providers**. The content of pushed data are the responsibility of those pushing the data. However, as they are the source, the data they do push can be considered to be canonical and of the best available quality. • **Facebook**. As Crossref DET will be querying the Facebook API for a large number of DOIs, the period between updates is entirely dependent on practical scaling issues. DET may prioritize fetching data for DOIs that are more likely to have activity.	2016/02/05
#5	Provide a documented audit trail of how and when data generation and collection methods change over time and list all known effects of these changes. Documentation should note whether changes were applied historically or only from change date forward.	R1, R2, R3	Events data are passed directly through. We provide no metrics. All events have a timestamps for when they occurred and when they were generated or collected. Thus the infrastructure used to generate and collect events can be matched to the timestamp. The Lagotto software is open source, so date stamps can be correlated to the version of the software that was running.	2016/02/05

#6	Describe how data are aggregated.	T2	Events are stored individually and returned individually. DET will collect data and make it available without aggregation.	2016/02/05
#7	Detail how often data are updated.	T3	DET provides an API to allow users to get data at any point. Data will be made available on the API as soon as possible after it is inserted into DET. • **Links from Crossref DOIs to DataCite DOIs**. Every time DOI metadata is deposited with Crossref the related events occur and are pushed into DET, effectively creating a live stream. • **Links from DataCite DOIs to Crossref DOIs**. Every time DOI metadata is deposited with DataCite the related events occur and are pushed into DET, effectively creating a live stream. • **Twitter DOI mentions**. A live stream. • **Wikipedia DOI citations and uncitations**. A live stream. • **Data from other providers**. Depending upon the providers, these can be received as a live stream or sent in batches. • **Facebook**. The update of Facebook events is yet to be determined.	2016/02/05
#8	Describe how data can be accessed.	T4	All data will be freely available via the DET API. The raw data will be the primary way of interacting with DET. For a fee, we will also provide an SLA (service-level agreement) that will guarantee consistency of service (guaranteed response times to API calls). The data will be identical to the free version, however.	2016/02/05
#9	Confirm that data provided to different data aggregators and users at the same time are identical and, if not, how and why they differ.	R4	DET provides an API, which will allow users to make queries against DOIs to retrieve events. DET also provides an SLA version of the API. This will have identical data, but we make guarantees of response times. There will be a single API for all data, which is open. Using the SLA version of the API provides identical data.	2016/02/05

#10	Confirm that all retrieval methods lead to the same data and, if not, how and why they differ.	R4	Different retrieval methods will lead to the same data as all Altmetric applications use the same underlying database and API. However, the article report pages seen within the Altmetric Explorer product or when the Altmetric Badges are clicked on are cached for 60 minutes by the content delivery network we use (Fastly). Therefore, it is possible that a change to an output that appears in the API results immediately will not be reflected in the relevant article-report page for up to an hour.	2016/02/05
#11	Describe the data-quality monitoring process.	T5, A2	The main failure mode will be service interruptions, meaning data sources becoming unavailable. These will be monitored per source to ensure that there is a constant stream of data. For DET, quality means consistency not, e.g., detection of gaming.	2016/02/05
#12	Provide a process by which data can be independently verified (aggregators only).	R5	All data will be freely available. The source code of the software used to generate the data will also be freely available.	2016/02/05
#13	Provide a process for reporting and correcting data or metrics that are suspected to be inaccurate.	A2	Crossref support will be able to handle requests. We can attempt to reprocess raw data to re-generate events. We can back-fill missing events with appropriate date-stamps. As we are not aggregating events into metrics or scores, we will not provide scores which might later need adjustment.	2016/02/05

NISO Altmetrics Working Group C "Data Quality" – Code of Conduct Self-Reporting Table

Example for data aggregator: PLOS (Public Library of Science) Article Level Metrics (ALM)

Item	Description	Supports CoC Recommendation	Aggregator / Provider Submission*	Last update of self-reporting table**
#1	List all available data and metrics (providers and aggregators) and altmetric data providers from which data are collected (aggregators).	T1	PLOS collects metrics data from the following data providers: • **Citations**: Web of Science, Scopus, Crossref, PubMed, Europe PMC, DataCite • **Altmetrics**: Twitter, Facebook, Reddit, Mendeley, CiteULike, F1000Prime, ScienceSeeker, ResearchBlogging, Wordpress.com, Wikipedia, ORCID, and PLOS Comments • **Usage Stats**: PLOS, PubMed Central, Figshare	2016/02/05
#2	Provide a clear definition of each metric.	A1	• **Web of Science**: Citation counts from the Web of Science database • **Scopus**: Citation counts from the Scopus database • **Crossref**: Citation counts from the Crossref citedBy service for members • **PubMed**: Citation counts from full-text articles in PubMed Central • **Europe PMC**: Citation counts from full text articles in PubMed Central • **DataCite**: Number of references as relatedIdentifier in DataCite metadata • **Twitter**: Number of tweets containing the DOI or journal-landing-page URL of the article • **Facebook**: Number of shares, likes, and comments for the journal-landing-page URL for the article, including private activity • **Reddit**: Reddit score and number of comments associated	2016/02/05

			with the DOI or journal-landing-page URL for the article • **Mendeley**: Number of individual-user and group-readership counts • **CiteULike**: Number of bookmarks • **F1000Prime**: F1000 score and article classification • **ScienceSeeker**: Number of blog posts • **ResearchBlogging**: Number of blog posts • **Wordpress.com**: Number of blog posts • **Wikipedia**: Number of Wikipedia pages in 20 most popular Wikipedia sites worldwide, subdivided by language • **ORCID**: Number of ORCID records • **PLOS comments**: Number of comments on the PLOS article page • **PLOS Usage stats**: COUNTER usage stats for HTML page views and PDF downloads from the PLOS website • **PubMed Central Usage stats**: Usage stats for HTML abstract, full-text page views, and PDF downloads from PubMed Central • **Figshare**: Usage stats for PLOS supplementary information hosted by Figshare	
#3	Describe the method(s) by which data are generated or collected and how data are maintained over time.	T1, T2, R1	Data are collected via public or private APIs. For F1000Prime and PubMed Central, usage data are downloaded as bulk files on a weekly or monthly basis, respectively.	2016/02/05
#4	Describe all known limitations of the data.	A3	The PLOS ALM service was started in 2009, with data providers added over time. No data for Twitter are available before the service launched in June 2012 because of limitations of the Twitter public APIs in providing historic data. For some services (e.g., Web of Science, Scopus, Mendeley, Facebook) only counts are available.	2016/02/05

#5	Provide a documented audit trail of how and when data generation and collection methods change over time and list all known effects of these changes. Documentation should note whether changes were applied historically or only from change date forward.	R1, R2, R3	No audit trail is available for PLOS ALM data. Changes in the open-source software that runs ALM, which can potentially affect how data are collected, are documented at https://github.com/lagotto/lagotto/releases.	2016/02/05
#6	Describe how data are aggregated.	T2	Data are aggregated by persistent identifier (DOI and PMID), and by month and day for the first 30 days after publication.	2016/02/05
#7	Detail how often data are updated.	T3	PLOS usage statistics are collected daily, PubMed Central usage stats are collected monthly, and F1000Prime data are collected weekly. Twitter data are collected every six hours the first week after publication. All other data are collected based on article age, with daily data collection during the first month after publication, followed by weekly data collection during the first year after publication, and monthly after the first year.	2016/02/05
#8	Describe how data can be accessed.	T4	Data are made available via open API (http://alm.plos.org/api, no registration), in the metrics tab available for every PLOS article, via ALM Reports (http://almreports.plos.org), and as CSV file downloadable monthly via the Zenodo data repository (e.g., http://doi.org/10.5281/ZENODO.44558 from January 2016 onwards).	2016/02/05
#9	Confirm that data provided to different data aggregators and users at the same time are identical and, if not, how and why they differ.	R4	Data provided to different aggregators and users is identical. The only exception is Web of Science data, which are only available to PLOS services because of license restrictions.	2016/02/05

#10	Confirm that all retrieval methods lead to the same data and, if not, how and why they differ.	R4	Data provided via different retrieval methods is identical. The only exception is Web of Science data, which are not available via API and monthly CSV file because of license restrictions.	2016/02/05
#11	Describe the data-quality monitoring process.	T5, A2	Data quality of newly collected data is monitored via an automated process that runs every 24 hours and looks for outliers (unusual spikes in activity, etc.). Data quality is also monitored manually by PLOS staff, taking into account input from external users.	2016/02/05
#12	Provide a process by which data can be independently verified (aggregators only).	R5	The PLOS ALM service runs using open-source software (https://github.com/lagotto/lagotto), which can be installed to collect data and compare them to the PLOS data. Data can also be independently verified by obtaining them directly from data providers (e.g., Mendeley, Facebook, Wikipedia, etc.).	2016/02/05
#13	Provide a process for reporting and correcting data or metrics that are suspected to be inaccurate.	A2	Data or metrics that are suspected to be inaccurate can be reported to PLOS staff via a feedback form at (http://www.plosone.org/feedback/new).	2016/02/05

NISO Altmetrics Working Group C "Data Quality" – Code of Conduct Self-Reporting Table

Example for data aggregator: Facebook

Item	Description	Supports CoC Recommendation	Aggregator / Provider Submission*	Last update of self-reporting table**
#1	List all available data and metrics (providers and aggregators) and altmetric data providers from which data are collected (aggregators).	T1	Facebook provides different online-event counts for a specific URL. These counts comprise "shares," "likes," and "comments". Aggregates are provided for the each of these social shares based on the total number of Facebook users who have shared, liked, or commented on a particular URL, respectively. Shares, likes, and comments that are public (i.e., are not restricted to specific user groups) contain further information such as the user name and time of event. Available data are further described in the Graph API documentation: https://developers.facebook.com/docs/graph-api.	2016/02/05
#2	Provide a clear definition of each metric.	A1	Facebook provides the following event counts: • **Shares** represent the number of times a particular URL has been shared by Facebook users on their own or other users' Facebook walls. Shares are thus posts that include a URL. Shares that are made available publicly (i.e., those for which access is not restricted to a certain user group) include the information about by whom and when the URL was shared. Each user can share the same URL multiple times; aggregated share counts thus do not necessarily reflect the number of unique users who have shared that URL. • **Likes** represent the number of times a particular post, share or comment has been "liked" (i.e., as indicated by a click on the Facebook "like button") by Facebook users. Each Facebook user can only like each post or comment once, but can "unlike" the same post, which removes the	2016/02/05

			particular like. Therefore, each like count represents the sum of users that have liked a URL at a particular moment in time. • **Comments** represent the number of times Facebook users have commented on their own or others' posts, shares, or comments. Each user can comment on the same post, share, or comment multiple times; aggregated comment counts do thus not necessarily reflect the number of unique users who have commented on a particular URL.	
#3	Describe the method(s) by which data are generated or collected and how data are maintained over time.	T1, T2, R1	The Graph API is well documented, but information about how the counts are generated is not available. No information about users is provided.	2016/02/05
#4	Describe all known limitations of the data.	A3	For pages that are not freely accessible—e.g., when a publisher requires cookies or a manual selection of options—Facebook is not able to properly determine the canonical URL and does thus not provide the correct online event counts. Facebook events are only available via the Graph API, further information regarding the limitation of the provided data are not available.	2016/02/05
#5	Provide a documented audit trail of how and when data generation and collection methods change over time and list all known effects of these changes. Documentation should note whether changes were applied historically or only from change date forward.	R1, R2, R3	Facebook regularly updates its API, sometimes including backwards-incompatible changes to how share, like, and comment counts are generated. API changes are versioned and documented publicly at https://developers.facebook.com/docs/apps/changelog and https://developers.facebook.com/blog/. The latest API is v.2.5, released October 7, 2015.	2016/02/05

#6	Describe how data are aggregated.	T2	The Graph API is well documented, but information about how the counts are aggregated is not available. https://developers.facebook.com/docs/sharing/webmasters/crawler.	2016/02/05
#7	Detail how often data are updated.	T3	In the Graph API, Facebook provides a timestamp that documents when this information was last updated.	2016/02/05
#8	Describe how data can be accessed.	T4	The Graph API is openly available. Users need to register for an API key for higher rate-limits.	2016/02/05
#9	Confirm that data provided to different data aggregators and users at the same time are identical and, if not, how and why they differ.	R4	As far as is known, all users get the same data from the Graph API.	2016/02/05
#10	Confirm that all retrieval methods lead to the same data and, if not, how and why they differ.	R4	Facebook has permission levels. The application retrieving the data must have the open key. Users can make their accounts public or private and can change the privacy setting of single posts from public, to restricted to certain user groups, to private and vice versa. Facebook data retrieved via the API represent a certain moment in time. If data posted at time A are changed at time B, results retrieved at A will differ from those retrieved with the same retrieval method at B. Changes in the API may change query results.	2016/02/05
#11	Describe the data-quality monitoring process.	T5, A2	Facebook has a built-in control at multiple entry points to attempt accuracy. However, further information about the data-quality monitoring process is not available.	2016/02/05

#13	Provide a process for reporting and correcting data or metrics that are suspected to be inaccurate.	A2	Users can submit a request to the Facebook developers' bug site. However, there is insufficient information about what actions Facebook will take in response to the request, unless an API retrieval change is needed. It does not appear that Facebook will adjust the data, but rather just correct the API.	2016/02/05

NISO Altmetrics Working Group C "Data Quality" – Code of Conduct Self-Reporting Table

Example for data aggregator: Mendeley

Item	Description	Supports CoC Recommendation	Aggregator / Provider Submission*	Last update of self-reporting table**
#1	List all available data and metrics (providers and aggregators) and altmetric data providers from which data are collected (aggregators).	T1	Mendeley offers total readership statistics per scholarly document added by Mendeley users to their private libraries. These statistics include academics status (students, professors, librarians, etc.), disciplines (sub disciplines) and countries of the Mendeley users, which can be selected by users from a list provided by Mendeley. Some of this demographic information is mandatory (e.g., discipline), while some is optional (e.g., country). This influences the extent to which this data are available for Mendeley readership counts. Mendeley offers a free open API for collecting the readership metrics including aggregated demographic information in a very fast way. The API is well documented: https://api.mendeley.com/apidocs.	2016/02/05
#2	Provide a clear definition of each metric.	A1	A readership count of a document reflects the number of Mendeley users that have added it to their libraries at a given point in time. However, the act of bookmarking/saving in Mendeley does not directly reflect reading the document; no clear definition of readership is available.	2016/02/05

#3	Describe the method(s) by which data are generated or collected and how data are maintained over time.	T1, T2, R1	Information about how readership is generated is not available.	2016/02/05
#4	Describe all known limitations of the data.	A3	The API requires an API key and uses rate limits. Readership data are anonymous: it does not include the information about owners of the private libraries, so that it is not possible to verify whether the readership count actually reflects the number of Mendeley users of a document. Some publications are saved in Mendeley but their readership counts are not available; for these, the message "readership counts are being calculated" is provided. Although selecting an academic status and discipline are obligatory when creating an account in Mendeley, some publications with total readerships statistics do not have any information about the users' academic status. The update of academic status lags behind the update of total readership counts and can cause discrepancies between the readership counts per academic status retrieved via the Mendeley online catalog and the API. There are duplicates in the catalog; for example, one document may appear three times in the Mendeley catalog with different readership counts for each entry. Information highlighting these limitations or any known errors is not provided. It is unclear whether errors are systematically identified and corrected.	2016/02/05

#5	Provide a documented audit trail of how and when data generation and collection methods change over time and list all known effects of these changes. Documentation should note whether changes were applied historically or only from change date forward.	R1, R2, R3	Information regarding changes of generating and calculating readership counts over time is available. API changes are documented at https://api.mendeley.com/apidocs.	2016/02/05
#6	Describe how data are aggregated.	T2	No information is available regarding how data are aggregated and how entries with identical identifiers (DOI, PubMed ID, arXiv ID, etc.), but differences in metadata, are handled. It is not clear how duplicates are handled and how and when their readership counts might be aggregated.	2016/02/05
#7	Detail how often data are updated.	T3	Readership counts may increase or decrease over time, based on users adding or removing documents from their libraries. Readership counts do not include timestamps, so it is not clear when and how often data are updated. An exception, however, is the monthly readership count that is provided for a Mendeley user's own papers (i.e., those he or she has authored); for these papers monthly historical readership data are provided for the last 12 months. No information is available on the frequency of updates and how long it takes until a user adding or removing a document to their Mendeley library is reflected in the readership count.	2016/02/05
#8	Describe how data can be accessed.	T4	Data can be accessed via the Mendeley catalog (https://www.mendeley.com/research-papers) or the open API (https://api.mendeley.com/apidocs). The API includes detailed information about how to use the API for data extraction: http://dev.mendeley.com/methods/?shell#introduction. However, not all data listed in the documentation (e.g., date	2016/02/05

			created) are available via the public API.	
#9	Confirm that data provided to different data aggregators and users at the same time are identical and, if not, how and why they differ.	R4	All users get the same data from the Mendeley API.	2016/02/05
#10	Confirm that all retrieval methods lead to the same data and, if not, how and why they differ.	R4	Mendeley readership counts retrieved through the web catalog and the API for the same document at the same time may differ because total readership counts and readership counts per academic status and discipline are not calculated simultaneously. Using different metadata (e.g., DOI, PMID, document title etc.) and different retrieval methods (web catalog vs. API) may result in different readership counts for the same document.	2016/02/05
#11	Describe the data-quality monitoring process.	T5, A2	No information is provided regarding the data-quality monitoring process and internal checks and control.	2016/02/05
#13	Provide a process for reporting and correcting data or metrics that are suspected to be inaccurate.	A2	Mendeley offers a support portal (http://support.mendeley.com) for questions and reporting problems using Mendeley and a feedback forum (https://feedback.mendeley.com) for suggestions for improvements.	2016/02/05

NISO Altmetrics Working Group C "Data Quality" – Code of Conduct Self-Reporting Table

Example for data aggregator: Twitter

Item	Description	Supports CoC Recommendation	Aggregator / Provider Submission*	Last update of self-reporting table**
#1	List all available data and metrics (providers and aggregators) and altmetric data providers from which data are collected (aggregators).	T1	Twitter provides data through both its web interface (http://www.twitter.com) and its APIs. The API specifications are documented here: https://dev.twitter.com/overview/documentation. Twitter explicitly provides information on four main types of objects: Tweets, Users, Entities, and Places. Each type of object has many metadata fields and each field has specific meanings. Some of this available data may be used as metrics: • **followers_count**: The number of followers a particular user currently has. • **favorite_count**: Indicates approximately how many times a particular tweet has been "liked" by Twitter users. • **retweet_count**: Number of times a particular tweet has been retweeted. Some metrics may also be deduced from the API calls, for example, the total number of items returned from a search API query, such as the number of tweets mentioning a DOI.	2016/02/05
#2	Provide a clear definition of each metric.	A1	No detailed information is provided to provide a clear definition of each metric.	2016/02/05

#3	Describe the method(s) by which data are generated or collected and how data are maintained over time.	T1, T2, R1	Some known limitations of Twitter metrics include: • Twitter data consumers should tolerate the addition of new fields and variance in ordering of fields with ease. Not all fields appear in all contexts. It is generally safe to consider a nulled field, an empty set, and the absence of a field as the same thing. • Tweets found in search results vary somewhat in structure from other API results. • Twitter's search service and, by extension, the Search API is not meant to be an exhaustive source of tweets. Not all tweets will be indexed or made available via the search interface. • The Twitter Search API is part of Twitter's REST (Representational State Transfer) API. It allows queries against the indices of recent or popular tweets and behaves similarly to, but not exactly like, the Search feature available in Twitter mobile or web clients, such as Twitter.com search. The Twitter Search API searches against a sampling of recent tweets published in the past seven days (as indicated by the API documentation as of Feb 1, 2016).	2016/02/05
#4	Describe all known limitations of the data.	A3		2016/02/05
#5	Provide a documented audit trail of how and when data generation and collection methods change over time and list all known effects of these changes. Documentation should note whether changes were applied historically or only from change date forward.	R1, R2, R3	The Twitter API is versioned, although an audit trail does not appear to exist.	2016/02/05

#6	Describe how data are aggregated.	T2	Twitter provides information on events based on different API calls. Aggregation of Twitter metrics depends on the API calls. Users or altmetric data aggregators decide whether and how to aggregate Twitter metrics such as the number of tweets and retweets of a document.	2016/02/05
#7	Detail how often data are updated.	T3	It is generally expected that the Twitter data are updated in real time, but what real time means is unknown.	2016/02/05
#8	Describe how data can be accessed.	T4	The Twitter API documentation provides information on access. OAuth is required for accessing the REST API, and subject to rate limit. The Public Streaming API provides a sample of all tweets. Access to the Twitter Firehose, the full tweets stream, requires special permission.	2016/02/05
#9	Confirm that data provided to different data aggregators and users at the same time are identical and, if not, how and why they differ.	R4	It is not guaranteed that all users get the same data. It has been shown that timeline data has random omissions on recent tweets for different users, and the Search API is not meant to be complete but provides access to a sample of recent Tweets published in the past seven days (see #3).	2016/02/05
#10	Confirm that all retrieval methods lead to the same data and, if not, how and why they differ.	R4	It is not guaranteed that different retrieval methods result in the same data. It has been shown that followers_count, favorite_count, and retweet_count do not immediately reflect recent changes.	2016/02/05

#11	Describe the data-quality monitoring process.	T5, A2	A web service provides information about the API operational health status in the most recent week, e.g., “operating normally,” “has performance issues,” or “encounter interruptions”: https://dev.twitter.com/overview/status.	2016/02/05
#13	Provide a process for reporting and correcting data or metrics that are suspected to be inaccurate.	A2	No information is available on how inaccurate data or metrics can be corrected.	2016/02/05

NISO Altmetrics Working Group C "Data Quality" – Code of Conduct Self-Reporting Table

Example for data aggregator: Wikipedia

Item	Description	Supports CoC Recommendation	Aggregator / Provider Submission*	Last update of self-reporting table**
#1	List all available data and metrics (providers and aggregators) and altmetric data providers from which data are collected (aggregators).	T1	The core metric one can derive from Wikipedia is mentions of DOIs in Wikipedia articles. Another metric one could use for altmetrics is page views, but it seems that most aggregators only use number of mentions of, for example, a DOI, and not how many views occur on a page where a DOI is mentioned. Wikipedia does not provide DOI mentions per article; this data needs to be harvested from Wikipedia content.	2016/02/05
#2	Provide a clear definition of each metric.	A1	Data refers to Wikipedia content (its pages). This data are collected as users edit pages. It is unclear how soon this data are available via the API: https://www.mediawiki.org/wiki/API:Main_page.	2016/02/05

#3	Describe the method(s) by which data are generated or collected and how data are maintained over time.	T1, T2, R1	Wikipedia provides API access to all its content and records changes when users edit pages. In the context of altmetrics, Wikipedia data are aggregated by many aggregators (e.g., Altmetric, Crossref DET, ImpactStory, Lagotto), which extract information about Wikipedia pages that mention scholarly document identifiers such as DOIs. Aggregation is not performed by Wikipedia but by data aggregators or users. For example, the Lagotto instance for PLOS articles reports the Wikipedia mentions by aggregating all DOI mentions in the top 25 Wikipedia language sites.	2016/02/05
#4	Describe all known limitations of the data.	A3	The limitations of provided data are unknown.	2016/02/05
#5	Provide a documented audit trail of how and when data generation and collection methods change over time and list all known effects of these changes. Documentation should note whether changes were applied historically or only from change date forward.	R1, R2, R3	Content on Wikipedia can change through time as article pages are edited. This may pose a problem for consistency as a data request at time X may give a different result than at X + 1 year. Because of the above, Wikipedia is one of the data providers where metrics may actually go down, something that we (almost) never see for citations or downloads.	2016/02/05
#6	Describe how data are aggregated.	T2	Wikipedia provides information on events based upon changes to Wikipedia pages. Aggregation of Wikipedia metrics depends on the API calls. Users or altmetric data aggregators decide whether and how to aggregate Wikipedia metrics, such as the number of times a document is mentioned, using different identifiers (e.g., DOI, URL, PMID) or in Wikipedia articles in different languages.	2016/02/05
#7	Detail how often data are updated.	T3	It is unclear how soon after a change to a Wikipedia page is made the data on the changes is available via the API.	2016/02/05

#8	Describe how data can be accessed.	T4	Wikipedia data can be accessed via the API documented at https://www.mediawiki.org/wiki/API:Main_pageIn addition, bulk downloads can be fetched at https://dumps.wikimedia.org/.	2016/02/05
#9	Confirm that data provided to different data aggregators and users at the same time are identical and, if not, how and why they differ.	R4	Data provided through via the API at the same time is identical for all users.	2016/02/05
#10	Confirm that all retrieval methods lead to the same data and, if not, how and why they differ.	R4	It is assumed that different retrieval methods lead to the same results.	2016/02/05
#11	Describe the data-quality monitoring process.	T5, A2	No information is provided regarding the data-quality monitoring process.	2016/02/05
#13	Provide a process for reporting and correcting data or metrics that are suspected to be inaccurate.	A2	The core metric one can derive from Wikipedia is mentions of DOIs in Wikipedia articles. Another metric one could use for altmetrics is page views, but it seems that most aggregators only use number of mentions of, for example, a DOI, and not how many views occur on a page where a DOI is mentioned. Wikipedia does not provide DOI mentions per article; this data needs to be harvested from Wikipedia content.	2016/02/05

Bibliography

(This appendix is not part of the NISO RP-25-201X-3, *Altmetrics Data Quality Code of Conduct*. It is included for information only.)

Cai, L., & Zhu, Y. “The challenges of data quality and data quality assessment in the big data era.” *Data Science Journal* 14, no. 2. (2015). DOI: 10.5334/dsj-2015-002

Immonen, A., Paakkonen, P., and Ovaska, E. “Evaluating the quality of social media data in big data architecture.” *IEEE Access* 3 (2015): 2028-2043. DOI: 10.1109/ACCESS.2015.2490723

The International Association for Information and Data Quality (IAIDQ). “Glossary,” last updated July 19, 2015, from: http://iaidq.org/main/glossary.shtml

National Information Standards Organization. (2016). *Altmetrics Definitions and Use Cases* (NISO-RP-25-201X-1).