Crossref Event Data: Transparency First

Joe Wass, Crossref
jwass@crossref.org
https://orcid.org/0000-0002-0840-454X

September 15, 2016

Crossref will launch Crossref Event Data (CED) later in the year. The service will collect and distribute events that occur around scholarly publications on behalf of the community. It will capture a wide variety of events such as mentions of articles on blogs, social media and bookmarks.

http://eventdata.crossref.org

1 Events not Metrics

The role of Crossref is to collect and distribute data that spans the Scholarly Publishing space. Currently this includes assigning persistent identifiers (DOIs) and registering bibliographic metadata for publications. We serve as the central linking hub. We see the need for a neutral service for collecting altmetrics-type data around publications.

We consider individual events to be the clearest, most transparent and most easy-to-audit type of data to provide.

2 Motivation

Every type of event has a different context which is important for its interpretation. It is important to capture this. For example:

- Article references are added to and removed from Wikipedia. It is not meaningful to say "this article is cited in Wikipedia" without knowing when the reference was added, or if it was removed. Furthermore, the reason for a removal can inform interpretation: there may be a targeted edit war over a single reference or it may be untargeted vandalism.
- Articles are shared on Twitter using their landing pages. CED attempts to reverse the landing pages back into persistent identifiers (Crossref DOIs). It is important to record our assumptions about the context at the time the event was retrieved, for example the set of article domain names we were working from.
- A situation in which ten different tweets referencing an article may have different interpretation to a single tweet retweeted nine times. It is important to include the full data to expose this information.

This has led us to arrive at the following conclusions:

- It is easy to inadvertently misinterpret metrics or even events.
- The supporting evidence is the foundation of interpretation of the event.
- Services may consume our data to produce metrics or aggregations. We should enable transparency at all stages.
- By providing the fullest possible information we enable the data to be used for a wide and unpredictable range of purposes.
- Transparency should be built into the collection of events.
- Supporting evidence and contextual information for all data must be supplied as a first-order output.
- It is important to work to a set of common standards, for example the NISO Code of Conduct.

Crossref participated in the development and specification of the NISO "Altmetrics Recommended Practices on Data Metrics, Alternative Outputs, and Persistent Identifiers". Development of CED and the Data Quality recommendations happened in parallel and mutually informed one another. CED is one of the example data providers in the recommendations.

The resulting principles make CED transparent by design. For every event we collect, we will provide the underlying data that provides evidence for it.

3 Status

CED is under active development, but we are making available a sneak preview of the dataset and API. We consider input from the researcher and practitioner community to be of vital importance. We would like to discuss the dataset and the principles that underpin the service. We hope that by exploring the data we can ensure that the service is useful and relevant to researchers, and that we are able to gain insight into important contextual issues.

4 Available Data

The available data, which is at an experimental stage, will contain events from:

- Wikipedia (article references using DOIs as they are added and removed)
- Twitter (mentions of articles via DOIs or article landing pages)
- Facebook ('like' counts for a selection of articles)
- Reddit (mentions of articles in comments via DOIs or article landing pages)
- Blogs (mentions of articles in comments via DOIs or article landing pages)

The dataset will be available on an experimental API. We collect approximately 40,000 events per day.