Crossref Event Data: A new, open, underlying data set for altmetrics

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

September 8, 2017

1 A new type of infrastructure

Crossref Event Data is a new piece of infrastructure in the scholarly publishing landscape, currently in Beta. It collects and distributes links between scholarly literature and other research objects, for example data sets, blogs and social media. It comprises the underlying data that makes up altmetrics. We collect mentions of Crossref Registered Content (i.e. material with DOIs) from across the web and make it available in an open data set.

Crossref operate agents to track mentions of registered content across the web, Twitter, blogs, Wikipedia and other places. The platform is also open to other partners who can provide this style of link data. We also operate shared infrastructure with DataCite so we can collect and share data on each others' behalf.

In establishing a new layer of infrastructure we have introduced a new style of data: a sequence of Events, each of which records a link between two objects, for example a webpage and a DOI. This open data set represents altmetric-style activity as a sequence of discrete events that occur on the web. It does not constitute metrics itself, but can be used to create metrics, or to lead to much richer outcomes.

2 Sources and quirks

Event Data tracks a number of sources, including Twitter, Wordpress.com, Hypothes.is, various RSS Newsfeeds, StackExchange, Reddit and Wikipedia. Each one is monitored for activity, but the form that each input takes can vary considerably. A specialist piece of software, called an Agent, connects to the source and produces an intermediary data record in a standardized format, called an Evidence Record. This encapsulates a chunk of input in a way that can be interrogated for occurrences that may lead to Events. A software component called the Percolator does the interrogation and extraction of Events.

This intermediary step allows the specialist input to be represented in a common format, and for subsequent processing to start from a common point. It also provides a unique ID that can be used to trace all activity connecting an input to eventual Events.

In addition to the discrete sources, many inputs, for example Newsfeeds, Reddit, Wikipedia, and Wordpress.com, direct the Percolator to retrieve webpages. Each webpage may present its own challenges: web servers may be off-line, they may block access to the Event Data bot, including in an unpredictable pattern. They may request that the bot doesn't visit them via the robots.txt file standard.

Not everyone uses DOIs when they talk about articles, so one of the hurdles we have had to overcome is mapping other representations, such as article landing pages and PIIs, back to DOIs. We have also had to balance this requirement for processing data with the need for trustworthiness and transparency.

3 Challenges to finding a common representation for Events

Traditional Crossref metadata can be thought of as an evolving assertion store, maintained by members of Crossref, who are Publishers. Event Data is similar to other Crossref metadata, in that it contains links, but instead of receiving structured data, we have to retrieve it ourselves from unstructured sources.

Because the data is found in a range of places, with a range of different concerns, we can't simply represent the data as a evolving assertion store, nor should we. For example, if a web page contained a DOI link, and then that link was subsequently removed, we shouldn't simply delete that information. References on Wikipedia come and go, but individual versions are immutable. Furthermore, the absence of a DOI might be unintentional, for example, as a result of a web server outage.

The data model we use is instead an Event stream. Each Event corresponds to an observation that was made. Using this model, the same observation can be made repeatedly to show the presence of a link over time. References on Wikipedia are recorded against the specific page version.

4 Mitigation through transparency

The Event Data pipeline is designed to be as straightforward and easy to understand as possible. However, as it depends on external services at a number of stages of the pipeline, behaviour may be difficult to understand. For this reason every stage of the pipeline is as transparent as possible.

All the software is open source. Every significant action, including access to external services, decisions made and actions taken, is logged in an open Evidence Log. If the Percolator is blocked because of a robots.txt file, or the server is unavailable, or refuses access, that will be logged. Every input from an Agent is stored in an open Evidence Registry, and every Event contains a link back to its Evidence Record. This means that anyone who wishes to audit the process can see exactly how data was processed and the chain of custody.

5 Responsibilities when consuming the data

Representing this data as a stream of Events means that we can provide full transparency about how the data was generated, and which inputs led to which outputs. This enables users of this new layer of altmetrics data to trust, verify and audit it. It also allows downstream providers to demonstrate a provenance chain right from the original data through to the finished product.

Whilst this new platform will provide a layer of data which will lower the barrier of entry to those wishing to conduct altmetrics research and build tools, it also also confers certain responsibilities onto the consumer. One core commitment that Event Data makes is transparency and non-preferential treatment of material. This means that we don't make decisions, interpretation or judgments about the data. For example, what constitutes a duplicate Event may be interpreted differently by different people. Consumers of the data therefore need to think hard about how they use it.

Individual Events at first appear the same as the "altmetrics events" that are counted. However, because they seek to record 'underlying' data, they may provide more detail and redundancy. We also seek to describe occurrences as we see them, which means that if there are factors in the underlying data that should be accounted for, e.g. Twitter bots or citation rings, the consumer must be mindful of them.