Request for Information: Event Data Source Gap Analysis

Project

Crossref runs Event Data to capture and share online mentions of research outputs registered with Crossref. Our aim is to provide context to published research and connect diverse parts of scholarly dialogue. We monitor a range of online sources for mentions of DOIs or their corresponding URLs, and make the resulting events public through an API. Event Data is supplemented by logs containing details about the observations made and the decision-making processes leading to specific events. Since 2017, we have collected nearly 740 million events.

General details about Event Data are available on Crossref's education pages and there is a comprehensive user guide (including a QuickStart page for the API).

Background

An individual event consists of a metadata triple:

- *Subject*: An online mention of a DOI, such as blog post, tweet, news article, or bibliography. We know that not everyone references the DOI and so we also attempt to collect events that use the URL the DOI resolves to.
- *Object*: A work with a Crossref DOI.
- *Relationship:* A single word or short phrase describing how the subject and object are linked, for example: cites, is cited by, mentions, reviews.

Events include additional metadata such as timestamps, the source name, and a link to an evidence log with information about how the event was discovered. Crossref events can be retrieved via an API (see for example http://api.eventdata.crossref.org/v1/events).

Our Event Data service was launched in 2017 with 12 sources:

Agent/data source	**Event type**
Crossref metadata	Links to DataCite registered content
DataCite metadata	Links to Crossref registered content
F1000Prime	Recommendations of research publications
Hypothes.is	Annotations in Hypothes.is
The Lens (Cambia)	Citations in patents

Newsfeed	Discussed in blogs and media
Reddit	Discussed on Reddit
Reddit Links	Discussed on sites linked to in subreddits
Stack Exchange Network	Discussed on StackExchange sites
Twitter	Mentions in tweets
Wikipedia	References on Wikipedia pages
Wordpress.com	Discussed on Wordpress.com sites

The Lens is no longer active as a source and no new sources have been added since Event Data was launched, although we have recently started to cover relationship metadata (which describes links between different metadata records) and are actively considering adding all references deposited by members in metadata records to event data.

The current sources cover various data types and demonstrate that we are able to create events with a diverse range of subjects and relationships. They were chosen on the basis of our knowledge of academic publishing, alongside identifying sites containing links to DOIs. Crossref has collated a list of further possible sources that have not yet been added.

In choosing sources for Event Data, we focused on a range that would demonstrate collection of a diversity of different data types. This was important as a proof-of-concept and to explore the possibilities for further expansion of the product in the future. Our aim has always been to work with the community to assess the value of new sources. The downside of a broad range of event types is that many use cases are only partially fulfilled. For example, while we capture social media events from Twitter and Reddit, we do not capture data from Facebook, Instagram, LinkedIn, TikTok, WeChat, and so on; we collect citations between works registered by Crossref and DataCite, but do not cover other DOI registration agencies, which are increasingly significant in countries such as Korea and Japan. We would like future development to be led by fulfilling the most important use cases.

Source Gap Analysis

In order to identify potential future sources, we would like to systematically survey where research is mentioned online. We are therefore commissioning a gap analysis to estimate the current coverage of Event Data and identify which new sources will have the most impact.

The main use cases we would like to fulfill are:

Data citation: Publishers deliver details of data use and reuse via citations in metadata registered with Crossref and DataCite. We can represent each citation as an event and in some cases differentiate between use as supplementary data or citation.

The research nexus: A generalisation of the data citation use case. There are diverse types of scholarly output that are deposited with Crossref, but link to other metadata records that are not part of Crossref's database. Examples include software, genbank records, protocols, pre-registered research plans, funding sources, and peer review reports. Linking these elements will help efforts towards reproducibility in research and findability of content. These research objects can also be linked to people and organisations via identifiers including ORCID and ROR.

Data sources for publishing platforms: Events can be presented by publishers alongside their research outputs in order to provide context to published works. There is significant overlap with altmetric platforms for this case, however we do not seek to reproduce the curation of their sources, or analysis, reports, and dashboards. We also have doubts that we will be able to approach the comprehensive coverage of commercial services. There are many potentially useful sources in this area, however to date we have not systematically surveyed them.

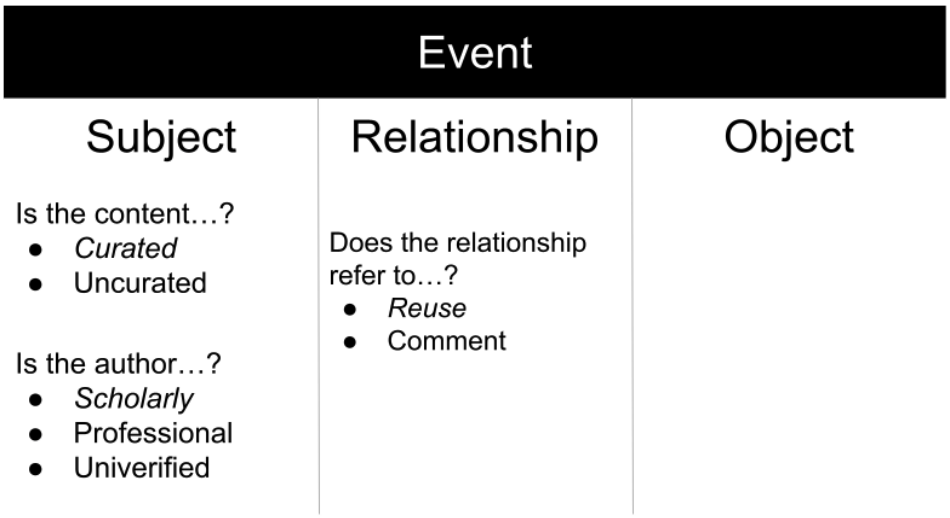
Impact assessments: A number of different types of organisations perform impact assessments, including institutions, funders, publishers, and consultants. Many are looking for new ways of measuring the impact of scholars and their scholarly outputs, other than tools like journal impact factor and citation counts. Events can be used to compliment these tools as input

for impact analysis. Outputs such as policy papers and patents are of relevance in these kinds of assessments.

Event Classification

To aid the analysis, we propose that events can be classified as:

- Reuse, or comment.
- Curated content or non-curated content.
- Scholarly, professional (including practitioners and journalists), or unverified authorship.



Sources containing curated scholarly subject content which is a reuse of the object are likely to be more valuable for the research nexus use case. Impact assessments may also be interested in non-curated content from professional authors (such as policy or technical documents). Social media is often valued by publishers for its immediacy and would be classed as comment, non-curated and with unverified authorship.

Not all sources are compatible with Event Data. We make data publicly available and do not track or monitor usage, which may conflict with the terms of use for some sources. Sometimes we can work around constraints, for example for Twitter data we are committed to periodically checking for deleted tweets and removing the corresponding events.

In other cases there may not be an API or equivalent method by which data can be collected by Event Data (one prominent example here is LinkedIn, for which no such method is available). In the analysis, we seek an assessment of the availability and compatibility of the most significant sources, noting where it may impact use cases.

Outcomes of gap analysis

From an analysis of events and sources, we would like to know:

- How can we assess the importance of an event source to a use case?
- Can the gap between current and total coverage be quantified in terms of number of sources and/or number of events?
- To what extent do we currently meet the use cases above, and which major sources of events are not yet included?
- For the most significant missing sources, is there a viable technical solution for gathering events (such as an API) and are there any associated costs with accessing the source?

A consultation with Event Data's current and potential future users is beyond the scope of this project, however the outcomes of this analysis will be presented to Event Data stakeholders in order to build a roadmap for new event sources.

Call for proposals

Deadline for responses: 11 July 2021
Contact for further questions and to submit: Martyn Rittman, mrittman@crossref.org
Anticipated start date: September 2021
Anticipated end date: October 2021

We seek written proposals for carrying out a gap analysis of Event Data sources. Proposals must include:

- Your overall approach to performing the gap analysis.
- A description of any previous relevant experience.
- Time and cost estimation to complete the work.

Additional Resources

- General information about Event Data https://www.crossref.org/documentation/event-data/
- Event Data user guide https://www.eventdata.crossref.org/guide/index.html
- General information about Crossref https://www.crossref.org/about/