Crossref Event Data: Transparency First

Joe Wass
Tech Lead, Event Data
@joewass

#wikicite17
Wikipedia is similar to journal publishing
Wikipedia is different to journal publishing
Crossref
DOIs, Metadata, Links
Altmetrics
We’re not interested in metrics
Linking infrastructure
Wikipedia experiments
DOI Resolution logs & referrals
DOI Resolution logs & referrals
Enter Wikipedia

--

HTTPS and Wikipedia

This is a joint blog post with Dario Tandavelli, coming from WikiCita 2018.

In 2016 we were taking our first steps along the path that would lead us to Crossref Event Data. At this time I started looking into the DOIs resolution logs to see if we could get any interesting information out of them. This project, which became Chromograph, showed which domains were driving traffic to Crossref DOIs.

You can read about the latest results from this analysis in the "Where do DOI Clicks Come From?" blog post.

Having this data tells us, among other things:

- where people are using DOIs in unexpected places
- where people are using DOIs in unexpected years
- where we know people were using DOIs but the links are more popular than we realised

By the time the ALM Workshop 2014 rolled around there was some preliminary data and we realised that Wikipedia came into the mix. As there are many DOIs in Wikipedia and people click them!

I met with Dario Tandavelli, lead of research at the Wikimedia Foundation, and shared the data. Dario — who co-authored in 2010 the Alliance Manifesto — has been interested in understanding how scholarly publications are used in Wikipedia. Over the past years Wikipedia contributors have made extensive use of references to the scientific literature using DOIs, and by doing so they have created a resource that represents today in many ways the "third matter to all research". There is growing interest in the community understanding how DOIs are being used in Wikipedia and its non-traditional scholarship.

During our discussions the subject of Wikipedia’s gradual transition to HTTPS was raised. We anticipated that this change would affect our data gathering.

Changes

When you’re reading a webpage and click on a link to another page, your web browser will usually hit the server of that second page. The last page you were on. This forms the basis of trackers like Google Analytics.

In HTTPS a small part, from the user’s perspective, is usually sent to the server. This is necessary for security reasons, to ensure that the data is not intercepted. However, this may affect the way that DOIs are used in Wikipedia.

DDO hyperlinks are just like any other hyperlink, and are really HTTP/1.0.
Early prototype

• Cocytus by Anthony Di Franco & Max Klein

• Adds and removes of DOIs from articles
Zooming back out
Events
Events

{
"obj_id": "https://doi.org/10.1080/20477724.2015.1133107",
"occurred_at": "2016-08-10T15:54:29Z",
"subj_id": "https://reddit.com/r/science/comments/4x2t19/occurrence_of_human_pathogenic_clostridium/",
"total": 1,
"id": "a8df79fc-5a1f-46df-96d6-9fb1f422aca7",
"subj": {
"pid": "https://reddit.com/r/science/comments/4x2t19/occurrence_of_human_pathogenic_clostridium/",
"author": {
"literal": "https://reddit.com/u/PHealthy"
},
"title": "Occurrence of human pathogenic Clostridium botulinum among healthy dairy animals: an emerging public health hazard",
"issued": "2016-08-10T15:54:29.000Z",
"URL": "https://reddit.com/r/science/comments/4x2t19/occurrence_of_human_pathogenic_clostridium/",
"type": "post"
},
"message_action": "create",
"source_id": "reddit",
"timestamp": "2016-08-11T11:26:47Z",
"relation_type_id": "discusses"
}
Principles
Trustworthy infrastructure
No interpretation
No metrics
Transparency
Evidence Trail
Open source
Open data
From prototype to production
Common model
Diffs require interpretation
Reliable linked data set
Representing Wikipedia Events

- Treat every page version as an entity
- Full links with version page and canonical page
Wikipedia Event - is_version_of

{
    "license": "https://creativecommons.org/publicdomain/zero/1.0/",
    "obj_id": "https://fr.wikipedia.org/wiki/Pierre_de_Bonzi",
    "occurred_at": "2017-04-15T23:53:50Z",
    "id": "0000524c-d000-4011-b35f-0709bd4b34ca",
    "terms": "https://doi.org/10.13003/CED-terms-of-use",
    "action": "add",
    "source_id": "wikipedia",
    "timestamp": "2017-04-15T23:54:36Z",
    "relation_type_id": "is_version_of"
}
Wikipedia Event - replaces

{
    "obj_id": "https://en.wikipedia.org/w/index.php?
title=Varg_Vikernes&oldid=780351110",
    "occurred_at": "2017-05-14T22:37:56Z",
    "subj_id": "https://en.wikipedia.org/w/index.php?
title=Varg_Vikernes&oldid=780410405",
    "id": "0000549c-ad9e-412c-ac13-4934646db3dc",
    "terms": "https://doi.org/10.13003/CED-terms-of-use",
    "action": "add",
    "source_id": "wikipedia",
    "timestamp": "2017-05-14T22:39:39Z",
    "relation_type_id": "replaces"
}
Wikipedia Event - references

{"obj_id": "https://doi.org/10.1007/s00253-011-3689-1",
"occurred_at": "2017-05-17T16:18:29Z",
"id": "000053b8-4eaf-460f-80fd-84d8bd0ff640",
"subj": {
  "url": "https://en.wikipedia.org/wiki/Gliotoxin",
  "title": "Gliotoxin",
},
"source_id": "wikipedia",
"obj": {
  "pid": "https://doi.org/10.1007/s00253-011-3689-1",
  "url": "https://doi.org/10.1007/s00253-011-3689-1"
},
"timestamp": "2017-05-17T16:19:23Z",
"relation_type_id": "references"}
Representing Wikipedia Events
Bridging the Evidence Gap
What we got
What we did to it
How we produced the Events
Evidence Record

• Linked from most Events
• Input links and text
• potential candidate matches
• successful matches
• discarded matches
• robots.txt exclusions
• sites unavailable
• URLs matched to DOIs etc
• …
No black box

Artifacts → Evidence Records → Events
Wikipedia → Evidence Records → Events
Event Data Data Flows
Extensible: Wikipedia Evidence Record

- RESTBase URL
- Edits to pages that had no references in them
- Don’t include text - high volume and relatively low signal, versions archived in Wikipedia
Wikipedia Event - references

https://evidence.eventdata.crossref.org/evidence/20170514-wikipedia-ce59b8d6-c83e-4d7c-9fec-35a174da1d17
Wikipedia Event with Evidence Link

```
{
  "license": "https://creativecommons.org/publicdomain/zero/1.0/",
  "obj_id": "https://doi.org/10.1007/s00253-011-3689-1",
  "source_token": "36c35e23-8757-4a9d-aacf-345e9b7eb50d",
  "occurred_at": "2017-05-17T16:18:29Z",
  "id": "000053b8-4eaf-460f-80fd-84d8bd0ff640",
  "evidence_record": "https://evidence.eventdata.crossref.org/evidence/20170517-wikipedia-8397e91a-c10b-4058-96a7-e0f83e7e988a",
  "terms": "https://doi.org/10.13003/CED-terms-of-use",
  "action": "add",
  "subj": {
    "url": "https://en.wikipedia.org/wiki/Gliotoxin",
    "title": "Gliotoxin",
  },
  "source_id": "wikipedia",
  "obj": {
    "pid": "https://doi.org/10.1007/s00253-011-3689-1",
    "url": "https://doi.org/10.1007/s00253-011-3689-1"
  },
  "timestamp": "2017-05-17T16:19:23Z",
  "relation_type_id": "references"
}
```

http://query.eventdata.crossref.org/events/000053b8-4eaf-460f-80fd-84d8bd0ff640
Non-scholarly links across the web, as many places as possible.

Interoperable. We are only one source. Broad community.

Open API.
Now in Beta!
Check the user guide

https://www.eventdata.crossref.org/guide/
https://www.crossref.org/categories/wikipedia

@joewass