Text and data mining

Patricia Feeney
Product Support
pfeeney@crossref.org
About Crossref

• 8,000+ publisher members
• over 89 million scholarly content items (now including preprints)
• Persistent citation linking (through DOIs)
• Funder identifiers
• Report and display corrections & retractions
• Open REST API & Search
• Data about content activity e.g. social, dataset links
<table>
<thead>
<tr>
<th>Academic Analytics, LLC</th>
<th>Hindawi Publishing Corporation</th>
<th>PTFS Europe</th>
</tr>
</thead>
<tbody>
<tr>
<td>Airiti, Inc.</td>
<td>Instytut Bioinfobank Sp. z o.o.</td>
<td>Reprints Desk</td>
</tr>
<tr>
<td>Altmetric, LLC</td>
<td>Labtiva, Inc.</td>
<td>Research Square</td>
</tr>
<tr>
<td>American Chemical Society (ACS)</td>
<td>mapegy GmbH</td>
<td>ScholarMate, Ltd.</td>
</tr>
<tr>
<td>Cambia</td>
<td>MESUR</td>
<td>Science-Metrix</td>
</tr>
<tr>
<td>CHORUS</td>
<td>Microsoft Research</td>
<td>Sciencescape, Inc.</td>
</tr>
<tr>
<td>Colwiz Limited</td>
<td>NASA ADS</td>
<td>Sunmedia</td>
</tr>
<tr>
<td>Copyright Clearance Center, Inc.</td>
<td>OCLC</td>
<td>Suweco, CZ</td>
</tr>
<tr>
<td>DeepDyve, Inc.</td>
<td>Ontario Council of Universities</td>
<td>Symplectic, Ltd.</td>
</tr>
<tr>
<td>EasyBib.com (ImagineEasy Solutions)</td>
<td>Libraries (OCUL)</td>
<td>Talis Education Limited</td>
</tr>
<tr>
<td>Elsevier, Inc.</td>
<td>PhilPapers Foundation</td>
<td>Technical Information</td>
</tr>
<tr>
<td>EMBL-EBI</td>
<td>Plum Analytics</td>
<td>Center of Denmark</td>
</tr>
<tr>
<td>Ex Libris Ltd.</td>
<td>ProQuest Information &amp; Learning</td>
<td>The Korean Federation of Science and Technology</td>
</tr>
<tr>
<td>Global Digital Central</td>
<td></td>
<td>Societies</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Third Iron</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UberResearch</td>
</tr>
</tbody>
</table>
Metadata enable connections
Our registered content

Over 89,000,000 million content items:
- Journals
- Books
- Conference Proceedings
- Datasets
- Dissertations
- Reports
- Standards
- Preprints (new!)
Citation metadata

including...

author names, ORCIDs, affiliations, article titles, ISSN, ISBN, pages, issue #, volume #s, dates, identifiers
more metadata

reference lists, funding data, ORCID, license data, clinical trial numbers, errata, retractions, updates and more through our Crossmark service, JATS-formatted abstracts, relationships between items...
Identifying licenses and how to retrieve them
Steps for TDM

1. Define problem
2. Identify potential corpus to mine
3. Discovery (full text links)
4. Identification of subset which can be accessed (license information)
5. Download identified corpus
6. Text and data mine corpus
Text and data mining metadata

- URLs directly to the full text content on publisher sites
- License information (URL)

⭐ consistent with NISO ALI recommendations ⭐
JSON (from http://api.crossref.org)

```json
link: [
  - {
    content-type: "application/pdf",
    content-version: "vor",
    intended-application: "text-mining"
  },
  - {
    content-type: "application/xml",
    content-version: "vor",
    intended-application: "text-mining"
  }
],
```
License URL (XML deposit)
license: [
  - {
      URL: "http://www.crossref.org/license",
      start: {
        date-parts: [
          [2013, 2, 3]
        ],
        date-time: "2013-02-03T00:00:00Z",
        timestamp: 1359849600000
      },
      delay-in-days: 0,
      content-version: "vor"
    },
  - {
      URL: "http://creativecommons.org/licenses/by/3.0/deed.en_US",
      start: {
        date-parts: [
          [2014, 2, 3]
        ],
        date-time: "2014-02-03T00:00:00Z",
        timestamp: 1391385600000
      },
      delay-in-days: 365,
      content-version: "vor"
    }
]
TDM step-by-step

1. Fetch Metadata
2. Check for license info / check license against whitelist
3. Point your TDM tools to full text URL and download
Fetch Metadata

• HTTP per-DOI request:
  
curl -L -iH "Accept: application/vnd.crossref.unixsd+xml" http://dx.doi.org/10.5555/515151
curl -L -iH "Accept: application/vnd.crossref.unixsd+xml" http://dx.doi.org/10.5555/515151

• REST API
REST API

How many members? 8,415
  http://api.crossref.org/members?rows=0

How many records does Crossref have? 89,377,410
  http://api.crossref.org/works?rows=0

How many journal article records? 66,739,806
  http://api.crossref.org/types/journal-article/works?rows=0

Book chapters? 10,425,160
  http://api.crossref.org/types/book-chapter/works?rows=0
How many works have license information? 25,799,562
http://api.crossref.org/works?filter=has-license:true&rows=0

How many license types are there? 352
http://api.crossref.org/licenses?rows=0

How many works have a CC-BY license? 254,986
http://api.crossref.org/works?rows=0&filter=license.url:http://creativecommons.org/licenses/by/3.0/

How many works have full text links? 25,527,916
http://api.crossref.org/works?filter=has-full-text:true&rows=0

Get license info for a single journal
http://api.crossref.org/works?facet=license:*&filter=issn:2090-8091
Records containing the term “blood” 480,587
http://api.crossref.org/works?query=%22blood%22&rows=0

How many records with the word "blood" in the metadata have license information and full text links? 170,388
http://api.crossref.org/works?filter=has-license:true,has-full-text:true&query=blood&rows=0

Get all of those records
http://api.crossref.org/works?filter=has-license:true,has-full-text:true&query=blood

Get matching records for a given publisher
look up member ID: http://api.crossref.org/members?query=hindawi
look up using ID:
http://api.crossref.org/members/98/works?filter=has-license:true,has-full-text:true&query=blood&rows=50
#2 Check for license info

- researcher should have list of licenses they accept
- extract items that use that license (using REST API or locally)

#3 Point TDM software to full text links

- perform a standard GET request on the url to download the full text from the Publisher’s site.
Resources

• Full REST API tour: https://github.com/CrossRef/rest-api-doc/blob/master/rest_api_tour.md

• REST API docs: http://api.crossref.org

• TDM docs: https://support.crossref.org/hc/en-us/sections/203047543

• Crossref Support: http://support.crossref.org, @CrossrefSupport
Thank you!

Patricia Feeney
pfeeney@crossref.org