



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

Academic Analytics, LLC	Hindawi Publishing Corporation	PTFS Europe
Airiti, Inc.	Instytut Bioinfobank Sp. z o.o.	Reprints Desk
Altmetric, LLC	Labtiva, Inc.	Research Square
American Chemical Society (ACS)	mapegy GmbH	ScholarMate, Ltd.
Cambia	MESUR	Science-Metrix
CHORUS	Microsoft Research	Sciencescape, Inc.
Colwiz Limited	NASA ADS	Sunmedia
Copyright Clearance Center, Inc.	OCLC	Suweco, CZ
DeepDyve, Inc.	Ontario Council of University Libraries (OCUL)	Symplectic, Ltd.
EasyBib.com (ImagineEasy Solutions)	PhilPapers Foundation	Talis Education Limited
Elsevier, Inc.	Plum Analytics	Technical Information Center of Denmark
EMBL-EBI	ProQuest Information & Learning	The Korean Federation of Science and Technology Societies
Ex Libris Ltd.		Third Iron
Global Digital Central		UberResearch

DOI Registration Agencies



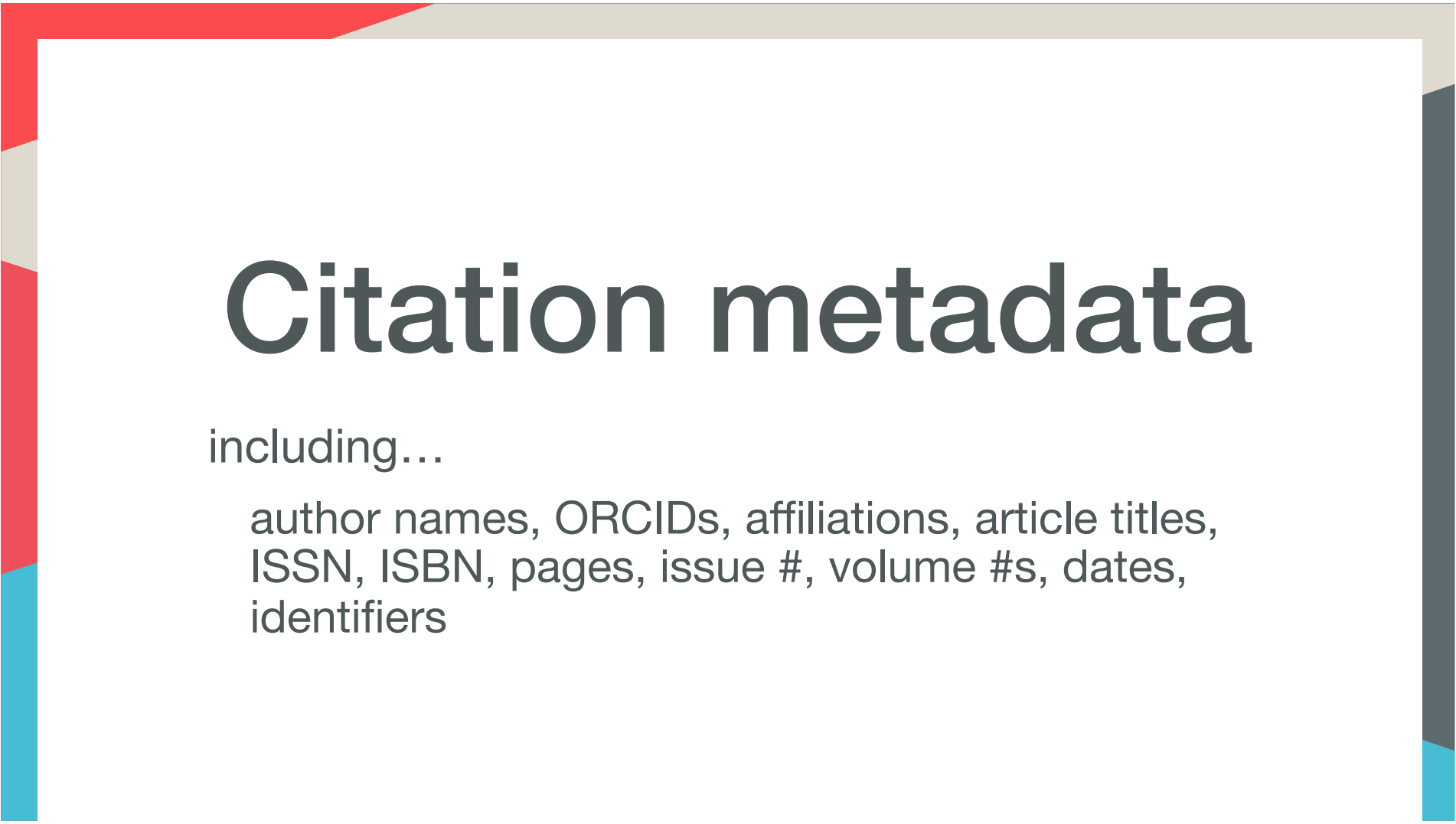
Metadata enable connections



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 IDs, [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](#) ★

```
<doi_data>
  <doi>10.5555/515151</doi>
  <resource>http://annalsofpsychoceramics.labs.crossref.org/abstract/515151/</resource>
  <collection property="text-mining">
    <item>
      <resource mime_type="application/pdf">http://
annalsofpsychoceramics.labs.crossref.org/fulltext/10.5555/515151.pdf
      </resource>
    </item>
    <item>
      <resource mime_type="application/xml">http://
annalsofpsychoceramics.labs.crossref.org/fulltext/10.5555/515151.xml
      </resource>
    </item>
  </collection>
</doi_data>
```

Full text URL (XML deposit)

```
link: [  
  - {  
    URL: "http://annalsofpsychoceramics.labs.crossref.org/fulltext/10.5555/515151.pdf",  
    content-type: "application/pdf",  
    content-version: "vor",  
    intended-application: "text-mining"  
  },  
  - {  
    URL: "http://annalsofpsychoceramics.labs.crossref.org/fulltext/10.5555/515151.xml",  
    content-type: "application/xml",  
    content-version: "vor",  
    intended-application: "text-mining"  
  }  
],
```

JSON (from <http://api.crossref.org>)

```
<program name="AccessIndicators">  
  <license_ref start_date="2013-02-03">  
    http://annalsofpsychoceramics.org/proprietary_license.html  
  </license_ref>  
  <license_ref start_date="2014-02-03">  
    http://creativecommons.org/licenses/by/3.0/deed.en_US  
  </license_ref>  
</program>
```

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"  
  }  
]
```

JSON (from
[http://
api.crossref.org](http://api.crossref.org))

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/10.5555/515151curl -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

