SEMSCI - ISWC2018, OCTOBER 9TH, MONTEREY

Creating Open Citation Data with BCite

Daquino Marilena, Ilaria Tiddi,
Silvio Peroni, David Shotton

Open Citation Data

OpenCitations Corpus

Once upon a time a Cell Biologist with a love for open data . . .

Crossref

made open only 1% of deposited article references

The Initiative for Open Citations (I4OC)

Born in 2017 to persuade publishers to make references deposited in Crossref open for reuse

Crossref

By late 2017, about 500 million article references to 19 million articles are openly made available through its APIs

OpenCitations

Creates and serves open RDF citation data, and provides services and tools for accessing the OpenCitations Corpus

What Crossref references are not open?

Incomplete openness in Crossref

Not all the publishers deposit article references with Crossref due to financial or technical restrictions.

Not all publishers that do deposit reference make them open.

Expensive curatorial activities in small publishers

Lack of software for checking references and formatting them in RDF

Why are some published references incorrect?

Incomplete openness in Crossref

Expensive curatorial activities in small publishers

Editorial teams must check each reference in the reference list of the article to be published for data correctness and respect for the journal reference style.

Lack of software for checking references and formatting them in RDF

Why are references not published in RDF?

Incomplete coverage in citation indexes

Expensive curatorial activities in small publishers

Lack of software for checking references and formatting them in RDF

Lack of easy-to-use tools for ingesting, curating and serving (RDF) reference data.

What do we need?

- easy-to-use interfaces for citation metadata discovery and reference text correction
- curation of RDF reference data conforming to a shared bibliographic data model
- facilities for the deposit of verified citations into open citation corpora

A workflow for contributing to open citation corpora

our solution

BCite

A bibliographic reference correction service

BCite Components

facilitates the editor's job while creating curated open citation data

- **BCite web application**

  for data entry, correction and curation

- **BCite triplestore**

  to store the curated RDF citation data

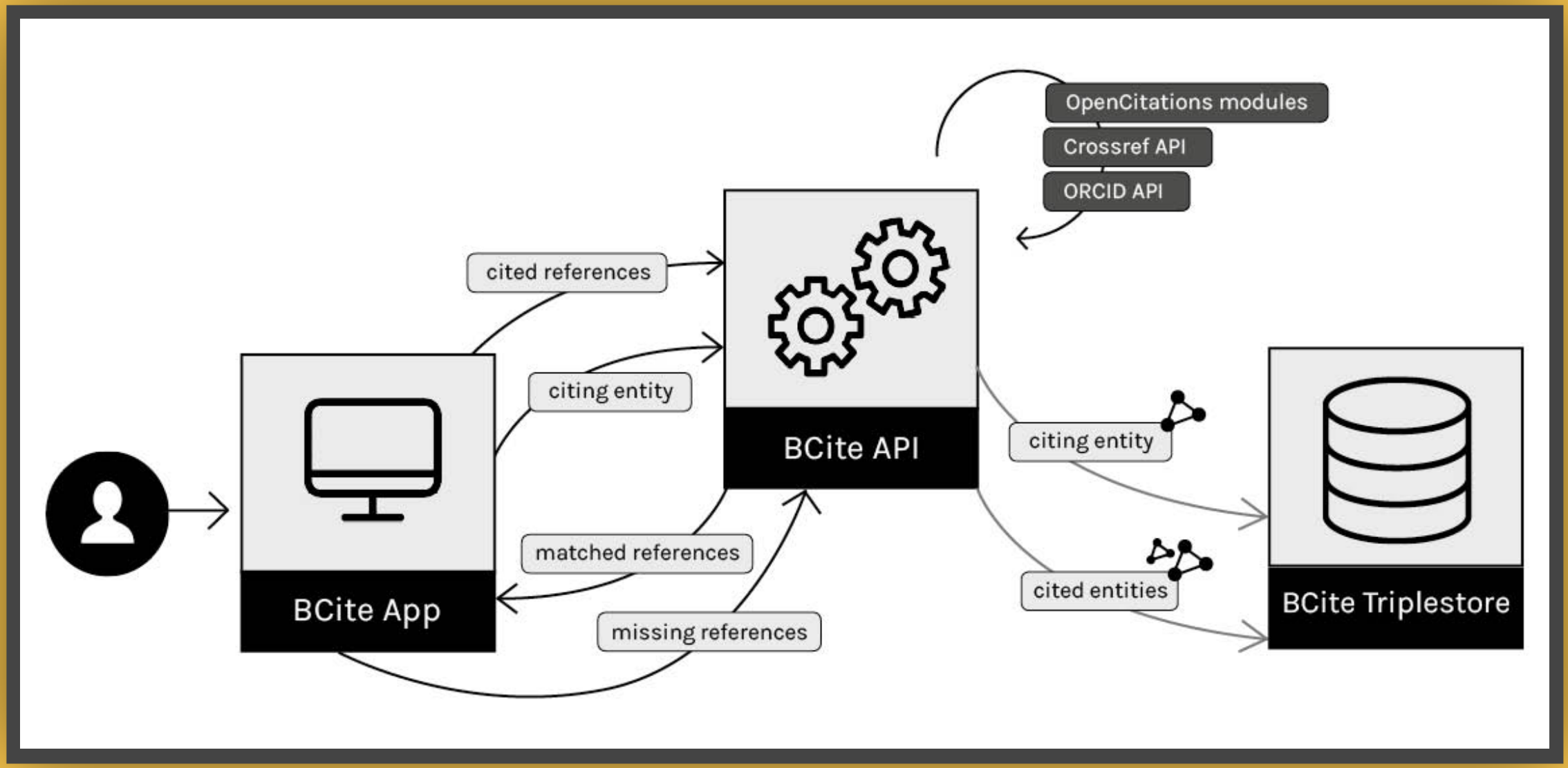
- **BCite API**

  creates RDF citation data ready for ingestion into the OpenCitations Corpus

  + OpenCitations modules
  + Crossref API
  + ORCID API

http://github.com/opencitations/bcite

Components


OpenCitations modules
Crossref API
ORCID API
cited references
citing entity
BCite API
citing entity
BCite App
matched references
missing references
cited entities
BCite Triplestore

How it works

Demonstration

Preliminary evaluation

- **facilitate an editor's curatorial activities**

  measure the number of references matched by the Crossref API, returned according to the specified citation style

- **increase the coverage of an RDF citation index**

  quantify the potential contribution to the Open Citations Corpus

Three scenarios

a “good” paper

most of the references are correct, and are to papers having DOIs and already described in Crossref

–

some of the references checked against Crossref are returned to the editor after correction by BCite.* These references may not already be included in the OCC

an “average” paper

most of the references are correct, but some of the referenced papers are not represented in Crossref, so these references cannot be checked

–

some references that can be checked against Crossref are returned to the editor after correction by BCite. These references may not already be included in the OCC

a “bad” paper

many of the references have errors. the referenced papers are poorly represented in Crossref and so cannot be checked

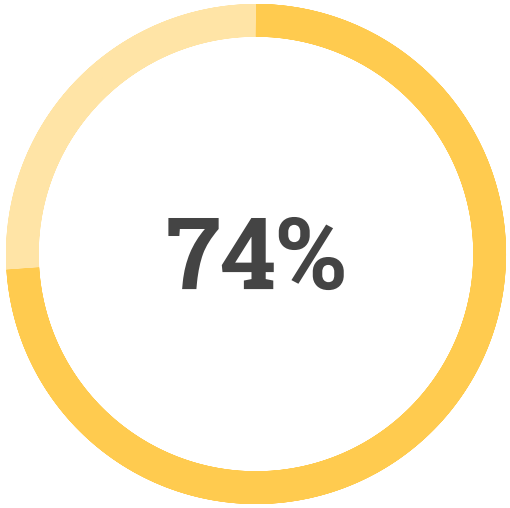
–

few of the input references will be matched, so that BCite will be unable to check and correct them. Most of them will also be new to the OCC

*we introduce common mistakes into the sample reference lists to test how many correct references BCite returns

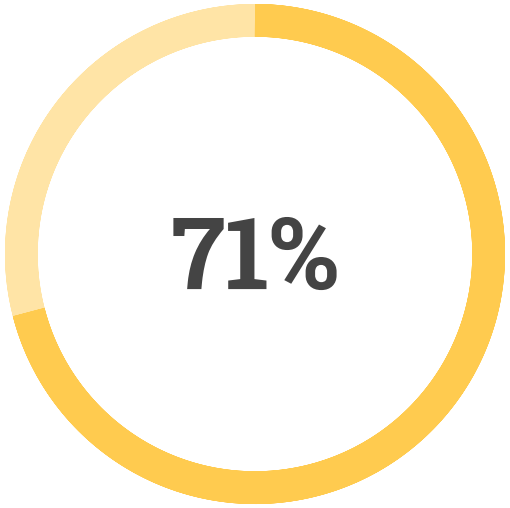
The good paper

Hammarfelt and Haddow 2017 - 42 refs



references correctly returned

26% references have to be manually cleaned and submitted for the inclusion in OCC



references not included in the OCC

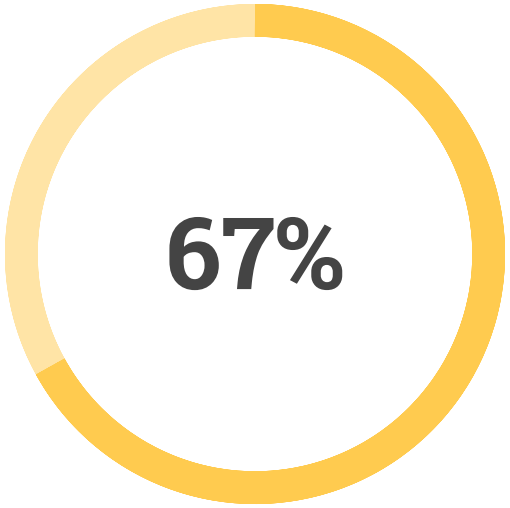
among the references correctly returned by BCite 71% are potential new contributions to be included in the OCC

The average paper

Biagetti 2016 - 30 refs

references correctly returned

50% are correctly matched, while 50% have to be manually cleaned

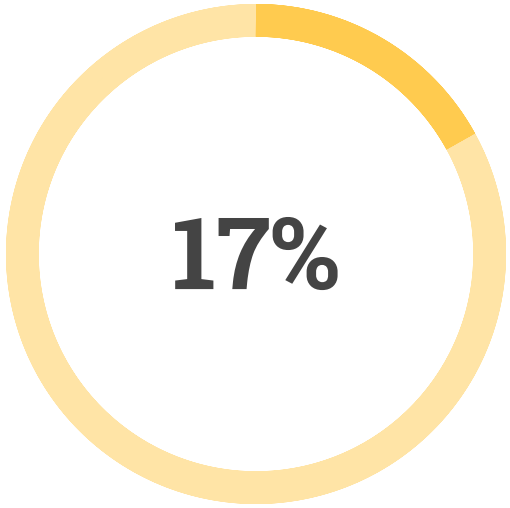


references not included in the OCC

among the references correctly returned by BCite 67% are potential contributions to be included in the OCC.

The bad paper

Citti 2008 - 40 refs



references correctly returned

83% references have to be manually cleaned and submitted for the inclusion in OCC

references not included in the OCC

all the references correctly returned by BCite are potential new contributions to be included in the OCC

Conclusions and future works

- **BCite**

  A simple strategy based on mutual benefit: support editorial teams in citation data cleaning while getting curated RDF citation data

- **Integrate BCite into the OpenCitations workflow**

  for validating and ingesting data created by trusted users (using BCite) into the OCC

- **Export to Crossref**

  Implement mechanisms for depositing open citation data in Crossref

questions?

Thanks