Machine accessibility of Open Access scientific publications from publisher systems via ResourceSync

Petr Knoth, Lucas Anastasiou, Giorgio Basile, Samuel Pearce and Nancy Pontika. CORE, United Kingdom

The number of scholarly research papers being published is gradually growing; it is estimated that approximately 1.5 million of research papers are produced each year and about 4% of them are offered via Open Access journals[1]. The high volume of scientific papers introduces new opportunities for content discoverability and facilitates a growth in various scientific disciplines via text and data mining (TDM)[2]. One of the greatest barriers to TDM is caused by the difficulty of programmatically accessing open access content from a wide range of publishers[3].

CORE has created a publisher connector, which harvests the open access scientific papers from publishers and exposes the content in a standardised API. Our contribution can be summarised as:

a) creating a seamless layer for accessing content from across publishers,
b) offering a generic integrated access point to millions of scientific papers, which will be constantly updated and
c) providing the first large-scale and highly scalable ResourceSync server deployment.

This is the first service to provide a harmonised access layer over non-standardised publisher APIs for retrieving scholarly content as well as the first implementation of ResourceSync scaling to millions of documents with the potential for fast real-time updates. With the connector we have made available over 1.5 million Gold and Hybrid Gold Open Access articles for text mining that used to be locked on publisher platforms.

This work has been conducted within the OpenMinTeD[4] project by the team developing the CORE aggregator[5,6]. In addition to the connector, we also offer an expertise directory[7], where we provide the following information per publisher:

- Publisher API
- Harvesting approach
- Publisher’s available information
- Features table
- Recommendations

References

CORE Publisher Connector

Proprietary APIs
Connector layer

Discovery services:
- Crossref
- PMC

Every record contains metadata and full text

All resources are accessible via ResourceSync

Every resource is automatically synchronised across all clients

Master copy through ResourceSync sitemaps

Immediate propagation of deletion

Open Access articles seamlessly accessible by everyone

7% of the total content available from the above publishers is Open Access

and more publishers on the way...

The largest datasets for text mining
Gold Open Access
- arXiv: 1,261,533
- PubMed Central (OA subset): 1,582,188
- CORE Publisher Connector: 1,660,625

For the largest collection of Green & Gold Open Access content, look at https://core.ac.uk/services#dataset

Access the connector: http://publisher-connector.core.ac.uk/resourcesync

The open access scientific papers from publishers are harvested and exposed as a standardised API.