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

Petr Knott, 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][2][3]. 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)[4][5]. One of the greatest barriers to TDM is caused by the difficulty of programmatically accessing open access content from a wide range of publishers[6][7].

1,005,768
Springer

436,340
frontiers

58,426
PLOS

160,091

Discovery services:
Crossref
PMC

Proprietary APIs
Connector layer

1,660,625
Open Access articles seamlessly accessible by everyone

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

Every record contains metadata and full text.

AI resources are accessible via ResourceSync

Synchronised copy

The largest datasets for text mining
Gold Open Access
- arXiv: 1,231,533
- 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

This work has been conducted within the OpenMinTeD[8] project by the team developing the CORE aggregator[9][10]. In addition to the connector, we also offer an expertise directory[11], where we provide the following information per publisher:

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

References: