Contribution ID: **45** Type: **not specified**

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

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 (OA) journals . 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) . One of the greatest barriers to TDM is caused by the difficulty of programmatically accessing open access content from a wide range of publishers .

In this poster, we will outline these technical difficulties and will present how we succeeded in harvesting metadata records and full text content of millions of OA articles from publisher APIs. We will also show how we have managed to provide an interoperable layer over these data using ResourceSync.

To achieve this we have 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) creation of a seamless layer for accessing content from across publishers, b) offering of a generic integrated access point to these data via ResourceSync and c) provision of a high performance access interface, which will be constantly updated. This is first service to provide a harmonised access layer over non-standardised publisher APIs for retrieving gold and hybrid gold scholarly content as well as the first implementation of ResourceSync scaling to millions of documents with the potential for fast real-time updates.

Currently, we are investigating the four largest, in the amount of published papers, scientific publishers - Elsevier, Springer, Wiley and Sons, Taylor and Francis - while the rest of the publishers are accessible via CrossRef. Apart from the connector, we also offer an expertise directory, where we provide the following information per publisher:
● Publisher API: what does it offer, providing descriptions and a critical analysis.
● Harvesting approach: we describe the approach followed to harvest open access content
● Publisher's available information: investigate the publishers'information available on their websites and cross-check it with our harvesting procedures to discover possible inconsistencies.
● Features table: we offer a table for easy access to essential information, such as download limitations rates, how to discover DOIs and access full text.
● Recommendations: we propose changes in the publisher's'systems, which would enable a better TDM performance with regards to OA scientific content.

The work has been conducted in the CORE project, a partner in the OpenMinTeD project. CORE is a global harvesting service offering access to millions of open access research papers, enriches the collected data for text mining and provides unique services to the research community. OpenMinTeD is an EU-funded project, with the aim to establish an open and sustainable TDM platform and infrastructure, where researchers can collaboratively create, discover, share and re-use knowledge from a wide range of text based scientific and scholarly related resources.

Authors: Dr KNOTH, Petr; Mr ANASTASIOU, Lucas; Mr BASILE, Giorgio; Mr PEARCE, Samuel; Dr PONTIKA, Nancy (Open University)

Presenters: Dr KNOTH, Petr; Mr ANASTASIOU, Lucas; Mr BASILE, Giorgio; Mr PEARCE, Samuel; Dr PONTIKA, Nancy (Open University)

Session Classification: Posters and Minute Madness Sessions