FutureTDM
Explore . Analyse . Improve

ORGANISATIONS

IMPROVING UPTAKE OF TEXT AND DATA MINING IN THE EU

Facts

Project No:	665940
Program:	H2020 | CSA | GARRI
Duration:	09/2015 - 08/2017

CORE, aggregating the world's open access research papers

CORE is a global large-scale Open Access aggregation platform that offers access to a large volume of free and open access content (https://core.ac.uk). Its aim is to aggregate all open access research outputs from repositories and journals worldwide and make them available to the public. In this way, CORE facilitates free unrestricted access to research for all. Such a technical infrastructure is vital to demonstrate the advantages of open access policy over traditional publishing models. However, CORE has also encountered a number of challenges in the process.

Background

The last decades have seen a massive increase in the amount of Open Access publications in journals and institutional repositories. Having such large volumes of state-of-the-art knowledge freely available online provides benefits in many fields. It helps for example to reduce time and money spent on getting access to these publications for research.

"By 'open access' to this literature, we mean its free availability on the public internet, [..] without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself."

Budapest Open Access Initiative

CORE harvests openly accessible content available according to this open access definition: the platform currently offers approximately 70 million of bibliographic metadata records and over 6 million of full-text research outputs.

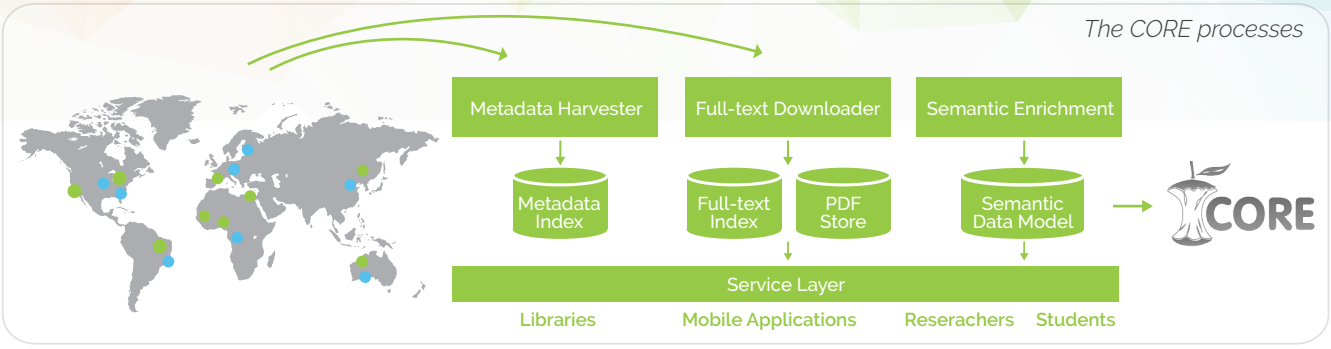
How does CORE work?

The CORE system harvests metadata records and the associated full-text content from Open Access repositories and journals listed in CORE. As a next step, the metadata is harmonised and enriched using both the harvested metadata as well as the full-text content. After running a standard text preprocessing pipeline including tokenisation, filtering, stemming and indexing of the metadata and text, a number of text mining tasks is then performed:

- Discovery of semantically related content
- Metadata extraction
- Extraction of citations and citation resolution.

In the information exposure phase, the system provides a range of services for accessing and exposing the aggregated data.

This project has received funding from the European Union's Horizon 2020 (H2020) Research and Innovation Programme.



Relevant for text and data mining of the scientific literature is that CORE provides access in various ways (such as through a web-based portal, through a plugin and through an API), which also helps to enable the development of new artificial intelligence-based applications for scientists.

What challenges does CORE hope to solve?

In developing this platform, CORE hopes to help solve a number challenges related to text and data mining.

Technical and Infrastructure

At the moment, there is no technical infrastructure for Open Access (OA) research papers that provides search functionality and support at different access levels for different user groups. One of the most important user groups for example is that of researchers and developers who need access to raw data so that they can analyse, mine and develop new applications. The CORE system attempts to fill this gap, providing support at different access levels.

Legal and content

Copyright law and other barriers are limiting the use of semantic enrichment technologies, namely text-mining. If semantic enrichment technologies are applied as part of an OA technical infrastructure in a way that provides significant benefits to users, users will prefer OA resources and this will create pressure on commercial publishers.

Education and skill

To fully exploit the OA reuse potential, it is important to better inform the OA community about both the benefits and commitments resulting from OA publishing. In particular, publishers should be aware of the fact that the content they publish might be processed and enriched by machines and the results further distributed. Similarly, the academic community should be better informed about the benefits of increased exposure and reuse potential of their research outputs due to these technologies.

Conclusions and recommendations

In order to achieve the full potential of having knowledge available, it is necessary to develop systems that:

make it easy for users to discover and access this knowledge at the level of individual resources,

explore and analyse this knowledge at the level of

- collections of resources and
- provide infrastructure and access to raw data in order to lower the barriers to the research and development of systems and services on top of this knowledge.
- CORE addresses these needs by providing a system that helps institutional repositories, individuals, researchers, developers, funding bodies and governments.

Furthermore what OA needs is a technical infrastructure demonstrating the advantages of OA policy over traditional publishing models.