Digging into Connected Repositories (DiggiCORE) project

Petr Knoth*, Zdenek Zdrahal*, Markus Muhr**, Alastair Dunning**
*Knowledge Media institute, The Open University
**The European Library

The goal of the DiggiCORE (Digging into Connected REpositories) project is to aggregate, at the level of both metadata and content, a vast set of research publications, from institutional repositories, archives (green OA route) and journals (gold OA route) worldwide, and provide novel tools for automatic enrichment of this content with relationships (relatedness, citations). The aggregated data with these relationships are used in turn to generate and publicly expose large and openly available networks of Open Access publications. These networks together with the actual full-text content can be then analysed using natural language processing and social network analysis methods to identify patterns in the behaviour of research communities, to recognise trends in research disciplines, to learn new insights about the citation behaviours of researchers, to discover new features that distinguish papers with high impact, etc.

To enable the analysis, the DiggiCORE project has developed a software infrastructure, building on top of the CORE system [Knoth & Zdrahal, 2012], which provides access to Open Access research outputs acquired by harvesting, cleaning, integrating and processing information from a very large and fast-growing collection of millions of research publications. The DiggiCORE project builds tools that enable access to the raw textual content intended for machine processing and the extracted and generated networks (citation network, article relatedness, author citation network) to the public via a set of web services and also as a downloadable dataset, thus creating a single access point for Open Access research outputs.

The availability of these datasets and the ability of anybody to mine them can significantly influence different disciplines. For example, it will allow researchers to run experiments that can enable the development of better methods for exploratory search and browsing in digital collections or new ways of evaluating research or the researcher's impact. We believe the availability of these datasets can also facilitate the transition to Open Access (by demonstrating the advantages of a uniform and free access to this huge distributed dataset) and can also help carry out experiments to find new impact metrics to improve scholarly communication.

The project provides the following outputs:

- A software infrastructure delivered to users as a free web-service and as a downloadable dataset that enables the analysis of the behaviour of research communities in the Open Access domain.

- New knowledge and understanding resulting from the data analysis.

DiggiCORE users are not only general researchers who want to read researcher papers, but also those who need programmable access to research papers. These are typically organisations, such as libraries or repositories, as well as researchers & developers designing new methods or analysing data. CORE allows these users to investigate the relationships between the impact of publications, citation patterns and the role of the author within the discipline. Users are able to compare these relationships across disciplines and through time. Finally, the data make it possible to measure the coverage of Open Access and help drive the OA agenda forward.

During the DiggiCORE project, we have aggregated over 16 million publications from around 600 systems and thousands of open access journals. These data are now freely available from the CORE portal. We have managed to significantly increase the number of visits to the CORE system to about 0.5 million per month. The system has been listed among Top 10 search engines that go beyond Google [Jacobs & Bruce, 2013] and also among the Top 100 Thesis Dissertation References on the Web [Top 100 Thesis, 2013].

The CORE system has been integrated into the search facilities of the European Library and a number of third party systems. The value of aggregating OA papers in the CORE system has been recognised in the community and CORE has recently won a JISC commisioned tender for the UK open access aggregator, i.e. an important component of the UK Shared Repositories Infrastructure.

The CORE dataset has also attracted the interest of many researchers, who already requested programmable access to the dataset. To promote the usage, the DiggiCORE team has organised the 1st and the 2nd International Workshops on Mining Scientific Publications at JCDL 2012 and JCDL 2013 with the aim to bring the community of researchers in this area together. The team has also published papers and disseminated the project outcomes at a number of major international events including Open Repositories 2012 and 2013, JCDL 2012 and 2013, OAI-8, NTCIR-10, Public Knowledge Project conference 2013 and has also been invited to the panel at the World Summit on the Information Society (WSIS+10) organised by UNESCO. This helped the project to build strong links with relevant stakeholders, which is essential for ensuring the sustainability of the solution in the future.

References

[Knoth & Zdrahal, 2012] Knoth, P. and Zdrahal, Z. (2012) CORE: Three Access Levels to Underpin Open Access, D-Lib Magazine, 18, 11/12, Corporation for National Research Initiatives

[Jacobs & Bruce, 2013] N. Jacobs and R. Bruce, http://www.jisc.ac.uk/inform/inform37/SearchingBeyondGoogle.html (Summer, 2013)

[Top 100 Thesis, 2013] TOP 100 Thesis & Dissertation Reference on the Web http://onlinephdprogram.org/thesisdissertation/ , (2013)