CORE: Aggregation Use Cases for Open Access

Petr Knoth
Knowledge Media institute
The Open University
Milton Keynes, United Kingdom
petr.knoth@open.ac.uk

Zdenek Zdrahal
Knowledge Media institute
The Open University
Milton Keynes, United Kingdom
zdenek.zdrahal@open.ac.uk

ABSTRACT
The push for free online availability of research outputs promoted by the Open Access (OA) movement is undoubtedly transforming the publishing industry. However, the mere availability of research outputs is insufficient. To exploit the full potential of OA, it must be possible to search, discover, mine, analyse, etc. this content. To achieve this, it is essential to improve the existing OA technical infrastructure to effectively support these functionalities. Many of the vital benefits of OA are expected to come with the ability to reuse OA content in unanticipated ways. Access to the OA content must therefore be flexible, yet practical, content-based and not just metadata based. In this demonstration, we present the CORE system, which aggregates millions of OA resources from hundreds of OA repositories and journals. We discuss the use cases aggregations should support and demonstrate how the CORE system addresses them, including searching, discovering, mining and analyzing content. We also show how aggregated OA content can be reused to build new applications on top of CORE's functionality.

Categories and Subject Descriptors
H.3.7 Digital Libraries, H.3.6 Library Automation

Keywords
digital libraries, cyberinfrastructure architectures, open access, content harvesting

1. INTRODUCTION
In the context of digital libraries, we understand *aggregations* as systems or cyberinfrastructures supporting the acquisition, storage, management, integration, enrichment and other types of information processing of digital resources, distributed beyond the scope of a single institution. In our previous paper [1], we have analysed the role of aggregations in supporting the Open Access movement. We have discussed the functionalities aggregations should provide and articulated the need for an OA technical infrastructure providing seamless access at three levels: *raw data*[1] *access, transaction access* and *analytical access* (see Table I). We also showed that there is currently no aggregation of all OA materials that would provide harmonised, unrestricted, transparent and convenient access to OA metadata and content, on top of which it would be possible to build services forming the necessary infrastructure. The CORE system aims to fill this gap.

In this paper, we provide a set of use cases for each of these access levels, building on the experience gathered during talks with stakeholders while building the CORE service and during aggregation requirements gathering carried out for the UK RepositoryNet+ project. We demonstrate how CORE can support them. During the demonstration, we will give short step-by-step tutorials on the implementation of these use cases in CORE. However, the ideas and concepts will be generally applicable.

2. THE CORE AGGREGATOR
The main goal of the CORE system[2] is to deliver a system aggregating research papers from multiple sources and providing a wide range of value-added services on top of this data. The purpose of CORE is to use this platform to establish the technical infrastructure for Open Access content taking into account the different needs of users participating in the OA ecosystem. CORE has been built to support a wide range of users [1] including those who need programmable access to metadata and content to build new applications (e.g. for the purposes of text-mining or bibliometrics), general users who need to search and explore content, repository administrators who need to manage collections and increase their visibility or Open Access analysts and advocates.

2.1 Raw data access
Raw data access is needed to enable passing the aggregated information to programs for further processing. It is a flexible type of access allowing others to extend existing infrastructures. It can be considered as the foundational access type, since all other functionalities can be built on this access level. Quite surprisingly, the existing aggregation systems in most cases do not provide this functionality or provide it in a very incomprehensive way [1], which might be in some cases intentional (protecting their own market, not wanting 3rd parties to easily build new services, etc.).

Aggregations should provide a free API allowing access to the aggregated metadata and content. Others should be able to use this API to build new services and software applications, for example making use of text-mining, semantic web or other technologies. The aggregation APIs should provide value-added services, i.e. services that are not or cannot be offered by the individual data providers, but can be useful to users.

CORE offers a free API, which provides functions for searching, downloading and finding similar (and duplicate) resources across the metadata and content aggregated from many sources. Additionally, CORE allows downloading content directly in plain text, accessing citation and reference information and offers also experimental services for content classification, URL extraction and data source benchmarking and analysis. The list of functions offered by the CORE API is growing to allow more convenient access to this large dataset. At the moment, we are also working towards releasing downloadable datasets complemented by a software library to provide flexible access for those who might have specific application or hardware requirements.

[1] In this paper, the concept of raw data refers to structured or unstructured publication manuscript data provided by repositories, or resulting from processing at the level of aggregation, to be used for further machine processing. Our concept of raw data is different from the more frequently discussed concept of research data which typically refers to data in the form of facts, observations, images, computer program results, recordings, measurements or experiences, on which an argument, test or hypothesis, or another research output, is based.

[2] http://core.kmi.open.ac.uk

Access level	What it provides	Users group
Raw data access	Access to the raw metadata and content as downloadable files or through an API. The content and metadata might be cleaned, harmonised, pre-processed and enriched.	Developers, digital libraries, eResearchers, companies developing software, ...
Transaction information access	Access to information primarily with the goal to find and explore content of interest typically realised through the use of a web portal and its search and exploratory tools.	Researchers, students, life-long learners, general public, ...
Analytical information access	Access to statistical information at the collection or sub-collection level often realised through the use of tables or charts.	Funders/government, business intelligence, repository/library managers, ...

Table I: The three access levels defined [1].

2.2 Transaction access

This type of access is perhaps the area mostly associated with the role of aggregations. It refers to the ability of aggregations to help users searching, exploring and discovering the aggregated information typically at the granularity of individual documents or small sets of documents. Yet, in the era of Google Scholar, Microsoft Academic Search and other academic search engines, the role aggregations should play in this domain is often misunderstood and not sufficiently appreciated.

First of all, OA aggregations should provide (fulltext-based and not just metadata-based) cross-search facilities exclusively for OA content. Current commercial search engines do not distinguish between subscription based and Open Access content, which often leads to users' frustration (when they only discover content they cannot download). Second, aggregations should work hand in hand with search engines to ensure that OA content is indexed thoroughly by exposing content and metadata according to the search engine recommendations (this is often not the case and OA content is paradoxically difficult to find). Third, aggregations should offer alternative cross-search interfaces for those users who require some specialist search functionality not important to the general users. For example, it has been reported that many users require faceted browsing (e.g. when looking into content provenance) or various exploratory and visual interfaces. The strength of aggregations should also be in the ability to utilize their raw data access layer, allowing the implementation of novel search and exploration interfaces. The right to deploy these interfaces at a web scale should no longer be reserved to a few powerful companies.

The CORE system offers a faceted cross-search interface tailored to Open Access content. The offered facets are currently the repositories (content providers), language of the content, full-text availability and publication date. The system exposes metadata tags according to Google Scholar recommendations, helping to increase the visibility of the aggregated content. CORE also offers content recommendation and visualization tools. There is even a native interface for mobile devices (Apple and Android) and a recommendation plugin that can be embedded to 3rd party systems. Furthermore, the CORE API allows building new search interfaces on top of the aggregation, such as an exploratory search interface discussed in [2].

2.3 Analytical access

The ability to generate transparent and justifiable content statistics is without doubts one of the main reasons why we need OA aggregations. Today, we are unable to accurately answer even the simplest questions, such as how many full-texts are available as Open Access. Aggregations have the potential to change this. Apart from providing various types of metadata and content statistics, aggregations should provide information about growth of content in different disciplines, research trends, subject clusters, funder statistics, source and collection citation statistics, etc. across repositories. These statistics are already needed and should be monitored on an ongoing basis. Furthermore, the characteristic sign of these statistics should be transparency (it is clear which data sources are used to generate these statistics) and extensibility (it is possible to generate new statistics from the data available on the raw data access level). Apart from statistical information, aggregations should also serve the community as a tool that encourages standardization. Aggregations can detect data inconstancies across sources and inform about them thereby helping to speed up adoption of standards and improving interoperability.

The CORE system offers an experimental tool called Repository Analytics. Repository Analytics report content statistics from the different content sources. The tool also indicates the ability or inability to perform content harvesting and shows the log of the harvests. The statistics are available through a dashboard as well as an API. There is lot of potential for providing more sophisticated analytical services in CORE, which is a direction we are following.

3. CONCLUSIONS

In this demonstration paper, we introduced some of the main use cases in which (Open Access) aggregation systems should assist different types of users and described how this functionality is currently implemented in the CORE system. More details will be provided during the demonstration.

4. ACKNOWLEDGMENTS

Our work has been supported by JISC. The use cases have been discussed with and provided to UK RepositoriesNet+ as part of the requirements-gathering for UK aggregation services.

5. REFERENCES

[1] 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. DOI= http://dx.doi.org/10.1045/november2012-knoth

[2] Herrmannova, D. and Knoth, P. (2012) Visual Search for Supporting Content Exploration in Large Document Collections, D-Lib Magazine, 18, 7/8, Corporation for National Research Initiatives. DOI= http://dx.doi.org/10.1045/july2012-herrmannova