LIBER QUARTERLY
The Journal of the Association of European Research Libraries
Vol. 25, no. 4 (2016) 172–188 | ISSN: 1435-5205 | e-ISSN: 2213-056X

Developing Infrastructure to Support Closer Collaboration of Aggregators with Open Repositories

Nancy Pontika

The Open University, UK
nancy.pontika@open.ac.uk

Petr Knoth

The Open University, UK
petr.knoth@open.ac.uk

Matteo Cacellieri

The Open University, UK
matteo.cancellieri@open.ac.uk

Samuel Pearce

The Open University, UK
samuel.pearce@open.ac.uk

Abstract

The amount of open access content stored in repositories has increased dramatically, which has created new technical and organisational challenges for bringing this content together. The COnnecting REpositories (CORE) project has been dealing with these challenges by aggregating and enriching content from hundreds of open access repositories, increasing the

Uopen Journals | http://liberquarterly.eu/ | DOI: 10.18352/lq.10138

discoverability and reusability of millions of open access manuscripts. As repository managers and library directors often wish to know the details of the content harvested from their repositories and keep a certain level of control over it, CORE is now facing the challenge of how to enable content providers to manage their content in the aggregation and control the harvesting process. In order to improve the quality and transparency of the aggregation process and create a two-way collaboration between the CORE project and the content providers, we propose the CORE Dashboard.

Key Words: open access; repositories; harvesting

1. Introduction

Over the past five years the amount of open access content has increased dramatically (Gargouri, Larivière, Gingras, Carr, & Harnad, 2012; Laakso & Björk, 2012; Morrison, 2015). According to the Registry of Open Access Repository Mandates and Policies[1] (ROARMAP), currently there are 79 funder, 54 organisational, 520 institutional and 72 departmental open access mandates. These mandates require the open accessibility of the research manuscripts and call for a shift in the scholars' publishing behaviour towards open access content. As a result, there is a high volume of scientific publications being self-archived in institutional and subject repositories. Even though there is an increasing amount of manuscripts that can be accessed on the web for free, there are still technical challenges in automatically bringing together full-text open access content from different systems and reusing it (Knoth, Anastasiou, & Pearce, 2014).

For the past five years the CORE project has been harvesting research manuscripts from open institutional and subject repositories, and open access journals. CORE's mission is not only to increase the visibility of the open access research manuscripts, but also to enable all research stakeholders to discover, access and reuse this open access content by providing three levels of access:

1. Programmable data access
2. Transaction information access
3. Analytical information access (Knoth & Zdrahal, 2012).

The programmable data access focuses on providing access to raw data and this is made possible with the use of the CORE API.[2] So far we have 135 API registered users, who are in position to gain access to CORE's content and the Data Dumps[3] that permit the text-mining of these manuscripts. This level of access is intended primarily for researchers, developers and companies. The second level is implemented with a set of services; the CORE portal,[4] where users can search and retrieve manuscripts from CORE; the Mobile application, which enhances the user flexibility of searching the CORE content; and the Plug-in,[5] a tool that, when integrated with repositories, provides research paper recommendations hosted in the CORE collection. The CORE portal receives a high traffic every year; in 2015 we had 70,465 new visitors, while 8,513 returned to our page. In addition, throughout this time, 14,704,530 full-text documents were downloaded from CORE. The transactional access applies mainly to researchers, students and life-long learners. For the third level of access, the analytical information access, CORE has newly implemented the Repositories Dashboard,[6] which is presented in this article. The target group of this application is primarily the repositories that act as CORE's data providers and their repository managers.

Currently there are other products that offer services similar to CORE, like Google Scholar[7] or CiteSeerX,[8] but there are some major differences between them. First, none of these were designed to aggregate repository systems. These products crawl and index research papers located anywhere on the web, providing access to them either directly through their own system, like CiteSeerX, or by linking to the original source, like Google Scholar. Once the content is aggregated by these systems, the originator has no control over this content and the aggregation system is not accountable to the original repository. CORE aims to strike a balance between the need for aggregating, promoting and exploiting the repository content and the need of the repository owners to have control over their content. Another popular project that relates to repositories' harvesting is the European-funded project Open Access Infrastructure for Research in Europe (OpenAIRE).[9] While OpenAIRE works with the full-text content of the articles, it does not store the full-text content, while CORE caches the full-text file. In this perspective, CORE is mostly similar to PubMed,[10] a free of cost search engine on medical literature, since it collects and disseminates papers from many content providers, both publishers and repositories, but serves the needs of the providers of the open access content instead.

2. The Harvesting Process

In order to collect the world's resources, CORE implements a harvesting technique with which it aggregates the open access content via the Open Archives Initiative Metadata Harvesting Protocol (OAI-PMH).[11] The OAI-PMH is one of the most widely used standards (Horwood, Sullivan, Young, & Garner, 2004) in content collection and the vast majority of repositories are supporting it.[12] At CORE, the harvesting process is divided into eight different but interdependent tasks.

2.1. Metadata Download, Extraction and Cleaning

As this first step, a repository's metadata are being downloaded into the CORE database. Since CORE uses the OAI-PMH protocol, it is essential for the harvesting process that the metadata are formatted in the Dublin Core schema, a collection of conditions that are used to describe objects in an online environment.[13] Afterwards, the metadata are being extracted in our database for local storage and they are cleaned; for example the order of the authors is corrected and normalised when necessary, or the digital object identifiers (DOIs) are extracted in case they appear in the wrong field.

2.2. Full-text Harvesting

Apart from downloading a record's metadata, CORE also downloads the article full-text and stores it in the CORE database. Users are in position to retrieve the cached content either via the CORE or any other search engine.

2.3. Text Extraction

After the full-text harvesting task, CORE extracts the full-text of an output into a text file, which is indexed to facilitate full-text searching.

2.4. Language Detection

Based on the fact that repositories hold large collections of manuscripts that are written in many languages, CORE has a dedicated task that recognizes

the language that an output uses. Thanks to the language detection task, CORE's users are in position to filter manuscripts in specific languages.

2.5. Citation Extraction

This task extracts the citations from an output's references. During this task the titles of all references are extracted and then CORE searches for the referenced output in the CORE collection. If the item is available in our collection then the two items are linked together; if not, CORE submits the titles of the referenced manuscripts to a DOI resolution service, CrossRef,[14] which sends back to CORE the output's DOI, if available.

2.6. Related Content Identification

This step relates with the discoverability and matching of semantically related manuscripts using information retrieval techniques.

2.7. Detection of Duplicates

In this step, the CORE system detects the duplicate records and groups these duplicates together in the database.

2.8. Indexing

Once the whole harvesting process is completed and a large volume of data is stored in the CORE database, the data is indexed. This task enables the searching of the CORE content and is also necessary for the functionality of the CORE API as well as to enable the creation of the Data Dumps.

3. The Need for a Repositories Dashboard

Existing research studies (Allard, Mack, & Feltner-Reicher, 2005; Walters, 2007; Wickham, 2010) that describe the roles of repository managers indicate that they "*manage the repository service by identifying goals and future strategies for improvement in the repository service based on new developments, usage*

statistics and feedback from users" (Wickham, 2010, p. 5). Furthermore, Allard et al. (2005) discovered that repository managers do not always have specialized technical skills, which indicates that perhaps they cannot take a direct advantage of the available information that CORE offers through the use of the API and the data dumps.

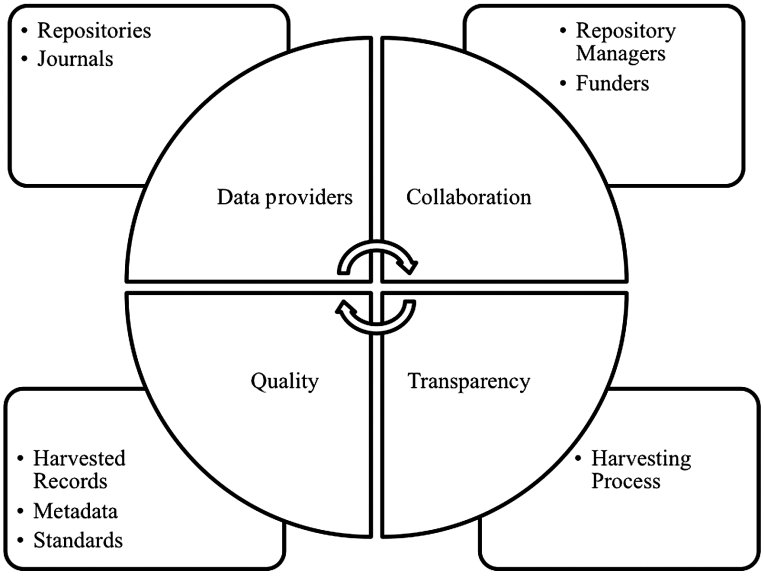
At the time when these aforementioned studies were conducted, five to ten years ago, the numbers of open access mandates were not as high as currently. According to ROARMAP, by the end of 2007 there were 22 funder and 137 institutional open access mandates, in 2010 there were 34 funder and 258 institutional, while the third quarter of 2015 ROARMAP has recorded 67 funder and 430 institutional open access mandates. In addition, SPARC Europe, an organization focusing on scholarly communications, in 2013 conducted an analysis of the global funder open access policies and discovered that from the 48 mandatory policies, 33 were green open access policies, which means that compliance is met via self-archiving in a repository (SPARC Europe, 2013). Therefore, the repositories' landscape has been significantly shifted by these open access mandates. In 2010 SHERPA Services surveyed the United Kingdom Council of Research Repositories members (Wickham, 2010) and discovered that it is the repository manager's responsibility to "*develop workflows to manage the capture, description and preservation etc. of research outputs*". In this new environment, repository managers have to further develop more skills and deal with timely deposits and publishers' embargo periods, count compliance percentages and assist authors with licensing their manuscripts (Pontika & Rozenberga, 2015).

CORE has been dealing with the aggregation challenges over the past four years by harvesting and enriching content from open access repositories, allowing the discoverability and reusability of millions of open access manuscripts via its own search engine and the API. While CORE has been able to provide the aggregated content from a single harmonised endpoint, it is now facing a challenge of how to enable the content providers to manage their content through the aggregation and control the harvesting process. Throughout the past four years of its existence, CORE has harvested 687 repositories from all over the world. All this time, CORE has received dozens of opt-in requests and only a few repositories have opted-out from the service. The primary reason for the opt-out requests was fear of institutions losing control of their content through the aggregation process. On the other hand, those repository managers and library directors that have opted-in, often email us requesting

access to details regarding their aggregated content and wishing to gain control over it.

In order to improve the quality and transparency of the aggregation process of the open access content and create a two-way collaboration between the CORE project and the providers of this content, CORE has created the Repositories Dashboard. The purpose of the Dashboard is to provide an online interface for CORE's data providers, the vast majority of which are repositories' managers (Figure 1). This online interface enables data providers to acquire more control of their content that appears in CORE by them gaining access to information that they did not have in the past. This allows the repository managers to efficiently manage the aggregation process, by, for example, requesting metadata updates or managing takedown requests directly in the CORE aggregation. The tool also provides information with regards to the frequency the content is being aggregated, including all detected technical issues, suggestions for improving the efficiency both of the harvesting process and the quality of metadata, and compliance with existing metadata guidelines. Furthermore, the CORE dashboard provides a range of statistics about the aggregated content.

Fig. 1: Repositories Dashboard Purpose.



4. Repositories Dashboard Overview

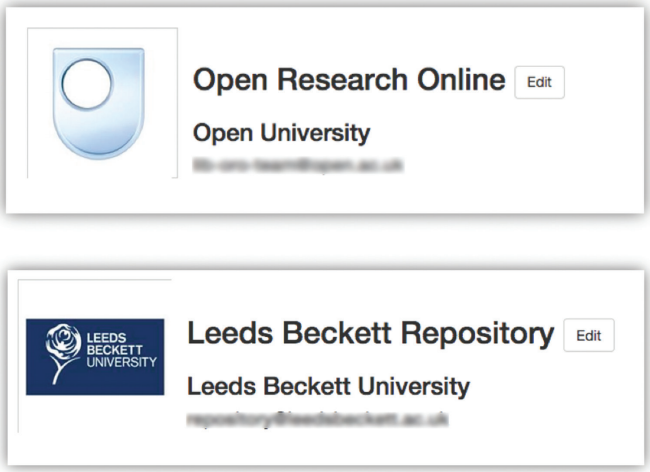
4.1. Institution Main Page

The vast majority of the repositories hosted in CORE are institutional, hosted and maintained by academic or research institutions. In the dashboard each institution and their affiliated repositories—there are cases where one institution may host more than one repository—have a dedicated page that includes the name and logo of the institution, the repository name and corresponding email. This page is intentionally left blank and it is the responsibility of the repository manager to fill in all this information (Figure 2).

4.2. Invite users to the Dashboard

There are two requirements for repositories to gain access to the dashboard. First they need to be CORE's data providers and second the repository manager needs to allocate and manage the dashboard invitations to the members of their own institution. In order to register one repository manager in the Dashboard, CORE applies the following process: initially, CORE uses either the personal email address of the repository manager or the generic

Fig. 2: Institution Main Page in the Dashboard.

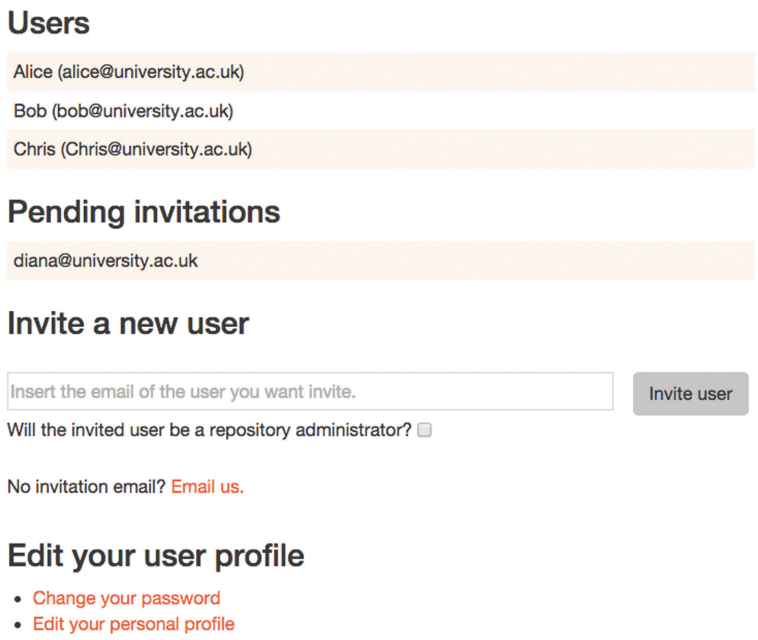


repository email address and sends an invitation to the new user, granting this account with administrative privileges. Afterwards, the person who handles this account has the right to create as many accounts for her/his own institution members as s/he wishes. These accounts have two levels of access: advanced and standard. The standard account allows users to view only the content hosted in CORE and the information related to their repository (Figure 3). Users with advanced accounts are able to perform actions, for example take down material, request the re-harvesting of the repository, or download Comma Separated Values (CSV) files, which contain the same fields as the information in the Content tab explored in section 4.3.

4.3. Content Tab

The content tab of one repository lists all the manuscripts that are harvested from this content provider. This page contains the title of the harvested

Fig. 3: Invite Users to the Dashboard.



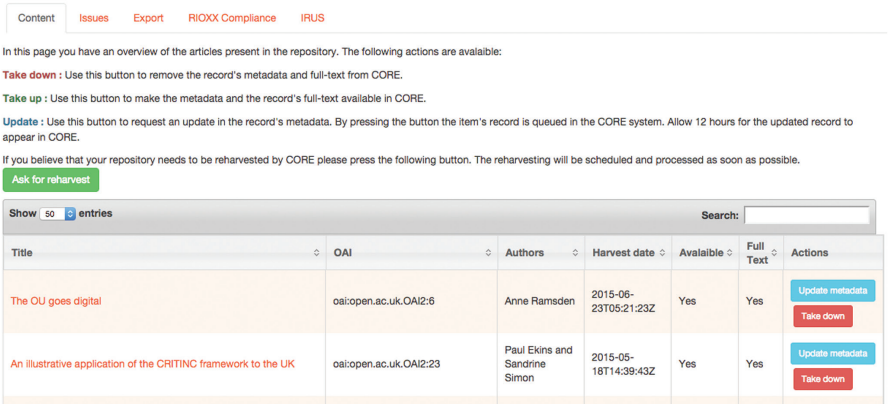
document, the output's unique identifier (OAI ID), the author name, the date the output was harvested and whether there is an openly accessible version through CORE (Figure 4). On this page repository managers can perform four tasks: take down and take up content, update metadata records and request a full re-harvesting of their repository.

The take down and take up buttons are considered to be critical for repository managers, who often receive take down requests of thesis and dissertations from authors or publishers and need to act promptly on them. CORE's intention was not only to make the process of taking down a document as simple and as fast as possible, but we also wanted to hand over the control of the harvested content to its data providers. Our future goal is to integrate this functionality with the repository software. For example, every manuscript that will be taken-down from an EPrints repository, will then be automatically removed from the CORE collection as well or vice versa.

4.4. Issues Related to Harvesting

During the harvesting process, explained above, various issues may occur. These issues can be critical to the ability of CORE completing the harvesting task and can lead to the whole corpus of a repository being not accessible in CORE, or to poor harvesting, where only some items are retrieved. As it has already been mentioned, repository managers may not have the technical

Fig. 4: Content Tab in the Dashboard.



skills to deal with these issues and in most cases they receive support from technical staff in their institution. In an effort to improve the communication between the repository managers and Information Technology staff, which would also result into the improvement of the harvesting process and the quality of the harvested content, CORE has created the "Issues" tab, where all possible issues are explained in further detail in a way that should be understood by both technical and non-technical staff.

First, the issues are divided into three types, a) error, b) warning and c) info, and a related description is provided for each one of them:

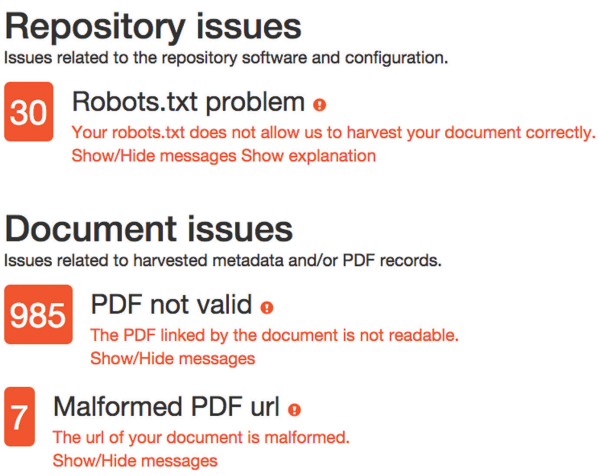
a) Error: When harvesting your repository/document we encountered an error that we couldn't resolve. These errors need to be fixed in order to harvest your repository/document.
b) Warning: We encountered an error but we are still able to harvest the repository document. We strongly recommend that these issues are resolved as they may lead to incompatibility problems in the future.
c) Info: This may not be a problem but it may be a clue for misconfiguration or future incompatibilities.

Apart from these generic instructions, the Dashboard software provides also issues that relate to a specific repository, as they were recorded during the harvesting process by CORE's systems. These issues are divided into two sections, repository and document issues (Figure 5). The first category relates to issues accessing the repository, mainly because the CORE crawlers are blocked from accessing items in a repository[15] or the OAI endpoint has changed. The second category relates to issues that the CORE harvester has encountered with regards to the aggregated full-text. For example in this page repositories' managers can find out information as to whether there are links where the resource locator (URL) in the <dc:indentifier> tag was not properly formulated, or if there are documents that require the use of a username and password to permit access to their content.[16]

4.5. IRUS-UK Statistics

In the past, CORE often received emails from repository managers requesting download and usage statistics from CORE. Generally speaking, the importance of a repository manager being aware of their repository's statistics can be

Fig. 5. Repositories Issues During the Harvesting Process.

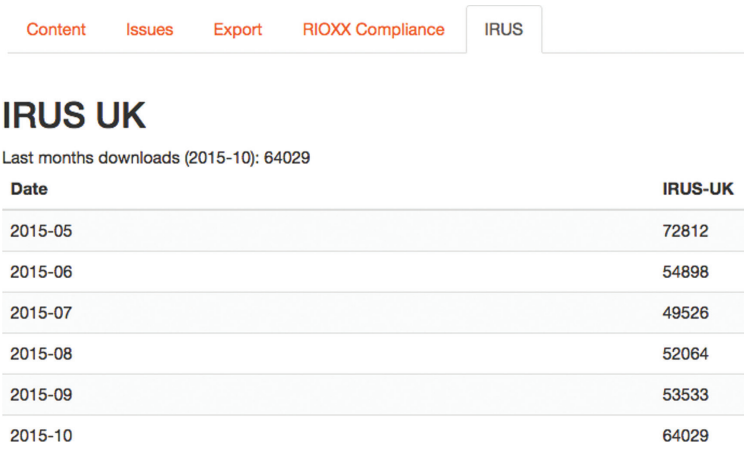


summarised into four main reasons. The statistics indicate the level of exposure of the research that is being conducted in an institution; it can serve as information regarding the return on investment both for the conducted research and the maintenance of the repository (Schöpfel & Boukacem-Zeghmouri, 2011); in some subject fields they can verify an increase in the citation rate of the papers (Gentil-Beccot, Mele, & Brooks, 2010); and it is a signal for prospective citations (Watson, 2009). Apart from downloading the metadata files of the repositories collections, CORE, during the harvesting process, downloads also the full-text of the manuscripts and caches this PDF version in its own database. As an effort to provide to repository managers information regarding to the manuscripts' downloads from CORE, we have integrated in the Dashboard the Institutional Repository Usage Statistics (IRUS-UK, see Figure 6). IRUS-UK[17] is a Jisc-funded project that serves as a national repository usage statistics aggregation service. IRUS-UK aims to provide article download statistics for content from UK repositories. Repositories who participate in the IRUS-UK project, which are currently close to 90, have access to these statistics from the CORE Dashboard as well.

4.6. RIOXX Metadata

The RIOXX Metadata application profile[18] aims to assist repository managers in tracking compliance with the Research Councils UK Policy on Open

Fig. 6: IRUS Statistics in the Dashboard.

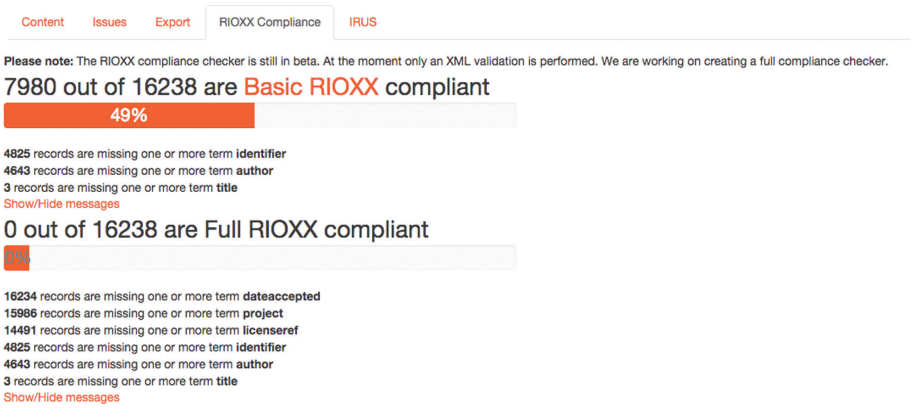


Access and Guidance.[19] Via the UK Metadata Guidelines for Open Access Repositories,[20] RIOXX provides mainly directions on the discoverability of the research manuscripts across different systems with the use of a set of metadata elements, resulting in the automated detection of the RCUK compliant manuscripts in a repository. During the harvesting process, CORE is in position to detect those UK repositories that support the RIOXX metadata and run a compliance check. The purpose of this task is to provide repository managers with the ability to validate the metadata inserted in their repositories. The dashboard supports two validation types, the "Basic" and the "Full", similar to the RIOXX application (Figure 7). The difference between these two types is that the first one has less strict constraints with less fields, while the latter has more fields and requires the input of more data with rigid metadata rules; the latter could be proved more important from a funder perspective.

4.7. Benefits of the Repositories Dashboard

CORE foresees some distinctive benefits with the implementation of the CORE Dashboard. First, it is expected that it will bring an increased and simplified collaboration between the aggregator, that is the CORE service, and the content providers, which are the repositories and their administrators. The Dashboard will also be the tool to improve the control of the content providers over the harvested content, something that we hope will reduce the

Fig. 7: RIOXX Compliance in the Dashboard.



scepticism and fear of sharing open access content with third party systems, which provide this content openly as well. We also hope that the technical issues notification system, not only will improve the harvesting process, but it will also provide a mutual understanding and closer collaboration between the repositories' managers and the technical staff, who support the repository. Finally, CORE's foremost goal is to broaden the discoverability of the open access content and its reuse when permitted.

5. Conclusion

The idea of facilitating the collaboration between CORE and repositories using the CORE Dashboard can be generalised to the collaboration of any aggregator with content providers, such as national libraries and archives. The overall aim of this approach is to strike a balance between the ability of aggregators to more effectively disseminate content, while allowing content providers to keep full control over it at all times.

6. Acknowledgement

This paper was presented at the 44th LIBER International Conference in London, on 24–26 June, 2015. The CORE team would like to thank the

Dashboard volunteer testers, Nick Sheppard and Chris Biggs for their comments and feedback.

References

Allard, S.L., Mack, T.R., & Feltner-Reichert, M. (2005). The librarian's role in institutional repositories: a content analysis of the literature. *Reference Services Review*, *33*(3), 325–336. doi:10.1108/00907320510611357.

Gargouri, Y., Larivière, V., Gingras, Y., Carr, L., & Harnad, S. (2012). Green and gold open access percentages and growth by discipline. In *17th International Conference on Science and Technology Indicators (STI)*, Montreal, CA, 05–08 September 2012. Retrieved February 16, 2016, from http://eprints.soton.ac.uk/340294/.

Gentil-Beccot, A., Mele, S., & Brooks, T.C. (2010). Citing and reading behaviours in high-energy physics. How a community stopped worrying about journals and learned to love repositories. *Scientometrics*, *84*(2), 345–355. Retrieved February 16, 2016, from http://arxiv.org/abs/0906.5418.

Horwood, L., Sullivan, S., Young, E., & Garner, J. (2004). OAI compliant institutional repositories and the role of library staff. *Library Management*, *25*(4/5), 170–176. doi:10.1108/01435120410533756.

Knoth, P., & Zdrahal, Z. (2012). CORE: Three access levels to underpin open access. *D-Lib Magazine*, *18*(11/12). Retrieved February 16, 2016, from http://www.dlib.org/dlib/november12/knoth/11knoth.html.

Knoth, P., Anastasiou, L., & Pearce, S. (2014). My repository is being aggregated: a blessing or a curse? In *9th International Conference on Open Repositories*, Helsinki, Finland, 9–13 June 2014. Retrieved February 16, 2016, from http://blog.core.ac.uk/files/OpenRepositories2014_v2.pdf.

Laakso, M., & Björk, B.-C. (2012). Anatomy of open access publishing: a study of longitudinal development and internal structure. *BMC Medicine*, *10*(124), 9. Retrieved February 16, 2016, from http://www.biomedcentral.com/1741-7015/10/124.

Morrison, H. (2015). The dramatic growth of open access June 30, 2015. *The Imaginary Journal of Poetic Economics*. [Weblog] Retrieved February 16, 2016, from: http://poeticeconomics.blogspot.co.uk/2015/06/dramatic-growth-of-open-access-june-30.html.

Pontika, N., & Rozenberga, D. (2015). Developing strategies to ensure compliance with funders' open access policies. *Insights*, *28*(1), 32–36. doi:10.1629/uksg.168. Retrieved February 16, 2016, from http://insights.uksg.org/articles/10.1629/uksg.168/.

Schöpfel, J., & Boukacem-Zeghmouri, C. (2011). Assessing the return on investments in GL institutional repositories. In *Gray Literature in Library and Information Studies,* De Gruyter Saur (pp. 1–20). Retrieved February 16, 2016, from https://halshs.archives-ouvertes.fr/sic_00601568/document.

SPARC Europe. (2013). *Analysis of funder open access Policies around the world*. Retrieved February 16, 2016, from http://sparceurope.org/analysis-of-funder-open-access-policies-around-the-world/.

Walters, T.O. (2007). Reinventing the library: How repositories are causing librarians to rethink their professional roles. *portal: Libraries and the Academy*, *7*(2), 213–225. doi:10.1353/pla.2007.0023. Retrieved February 16, 2016, from https://courses.washington.edu/mlis550/au10/pdf/Module_4_Walters_Reinventing_the_Library.pdf.

Watson, A.B. (2009). Comparing citations and downloads for individual articles at the Journal of Vision. *Journal of Vision*, *9*(4):i, 1–4. doi:10.1167/9.4.i. Retrieved February 16, 2016, from http://jov.arvojournals.org/article.aspx?articleid=2193506.

Wickham, J. (2010). Repository management: an emerging profession in the information sector. In Online Information 2010. London, UK, 30 November–2 December. Retrieved February 16, 2016, from http://eprints.nottingham.ac.uk/1511/.

Notes

[1] http://roarmap.eprints.org/.

[2] http://core.ac.uk/intro/api.

[3] http://core.ac.uk/intro/data_dumps.

[4] http://core.ac.uk/.

[5] http://core.ac.uk/intro/plugin.

[6] http://core.ac.uk/intro/dashboard.

[7] https://scholar.google.co.uk/.

[8] http://citeseerx.ist.psu.edu.

[9] https://www.openaire.eu/.

[10] https://www.ncbi.nlm.nih.gov/pubmed.

[11] https://www.openarchives.org/OAI/openarchivesprotocol.html.

[12] https://www.openarchives.org/pmh/tools/tools.php.

[13] http://dublincore.org/.

[14] http://www.crossref.org/.

[15] Read more at point 3 http://blog.core.ac.uk/2015/10/19/7-tips-for-successful-harvesting/.

[16] Read more at point 5 http://blog.core.ac.uk/2015/10/19/7-tips-for-successful-harvesting/.

[17] http://www.irus.mimas.ac.uk/.

[18] http://rioxx.net/.

[19] http://www.rcuk.ac.uk/research/openaccess/policy/.

[20] http://rioxx.net/guidelines/RIOXX_Metadata_Guidelines_v_3.0.pdf.