Better together: building services for public good on top of content from the global network of open repositories

UKSG webinar – 19 November 2019

Petr Knoth, Nancy Pontika

Big Scientific Data and Text Analytics Group
Knowledge Media Institute, The Open University
Papers OA sooner and sooner

Study quantifies the growing traction of open access

JCDL 2019 Best Paper Award

The delay between publication and OA availability decreasing globally

CORE’s work on REF2021 audit

• CORE data will be used in the REF 2021 Open Access Policy Audit
  Related sections: 40, 46 and 49.

• Deposit date of articles to be available in the CORE Repository Dashboard.

• Under development, release date depends on RE and Jisc.
To capture data, CORE recommends to institutions:

1. Make sure your institutional repository/ies are harvested by CORE
2. Adopt RIOXX as a data format
3. Add DOIs to records you are submitting to REF
4. Release deposit dates for harvesting purposes
5. Ensure “dateAccepted” field is used where known
6. Ensure all records submitted for REF have full text linked directly from dc:identifier
7. Ensure repository’s OAI-PMH endpoint is operational
Faster open access makes repository infrastructure more important
We don’t need just open access
we need fast open access
This study was only possible to conduct because of repositories and aggregators working together
By "open access" to this literature, we mean its free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself …

_Budapest Open Access Initiative, 2002_
“A single scientific repository is of limited value, real benefits come from the ability to exchange data within a network … … interoperability allows us to exploit today's computational power so that we can aggregate, data mine, create new tools and services, and generate new knowledge from repository content.”

– Confederation of Open Access Repositories (COAR)
“To achieve open access to scholarly journal literature, we recommend two complementary strategies.

• **Self-Archiving**: First, scholars need the **tools and assistance** to deposit their refereed journal articles in open electronic archives, a practice commonly called, self-archiving. When these archives conform to standards created by the Open Archives Initiative, then search engines and other tools can treat the separate archives as one. Users then need not know which archives exist or where they are located in order to find and make use of their contents.

• ...”

*Budapest Open Access Initiative, 2002*
CORE’s mission

Aggregate all open access research articles worldwide ...

... enrich this content and provide **seamless access** to it through a set of **data services** ...
Introducing CORE
Harvesting data is challenging

Mostly OAI-PMH

OA Repositories

OA Journals
Types of harvested outputs

CORE harvests all records and leaves quality control to data providers
Types of content providers

CORE harvests from everywhere:

- Repositories
  - Institutional
    - Includes repository and CRIS platforms
  - Disciplinary
- Journals
  - Pure Open Access
  - Gold Open Access
Harvesting is challenging

Many OAI-PMH implementations challenges …

- No content harvesting support
- Restrictions on full text downloading
- Failing resumption tokens
- Incremental updates
- Reliability
- Scalability
- Sequential nature of OAI-PMH
- Locating full text URLs in metadata
- Metadata interoperability
Scalability of repository platforms

<table>
<thead>
<tr>
<th>IR Software</th>
<th># Repos</th>
<th>AVG(Rec/s)</th>
<th>MED(Rec/s)</th>
<th>$\sigma^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSpace</td>
<td>659</td>
<td>147.29</td>
<td>71.64</td>
<td>1.07e+06</td>
</tr>
<tr>
<td>EPrints</td>
<td>402</td>
<td>35.48</td>
<td>29.14</td>
<td>1.98e+03</td>
</tr>
<tr>
<td>Digital Commons</td>
<td>149</td>
<td>28.57</td>
<td>11.47</td>
<td>3.70e+04</td>
</tr>
<tr>
<td>OPUS</td>
<td>74</td>
<td>39.56</td>
<td>23.84</td>
<td>2.42e+03</td>
</tr>
<tr>
<td>OJS</td>
<td>70</td>
<td>11.13</td>
<td>10.13</td>
<td>5.77e+01</td>
</tr>
<tr>
<td>dLibra</td>
<td>55</td>
<td>13.39</td>
<td>0.35</td>
<td>2.71e+03</td>
</tr>
<tr>
<td>Fedora</td>
<td>16</td>
<td>71.59</td>
<td>32.13</td>
<td>2.18e+04</td>
</tr>
<tr>
<td>Invenio</td>
<td>14</td>
<td>136.17</td>
<td>70.42</td>
<td>6.97e+04</td>
</tr>
</tbody>
</table>

Table 1. Comparing different IR software download speed
Need for more interoperability across systems
Harvesting data is challenging
Harvesting data is challenging
CORE Processing pipeline

- Metadata download, extraction and harmonisation
- Full text download
- Text extractions, sections extraction
- Metadata validation and enrichment (DOI, ORCID, etc.)
- Thumbnails generation
- References and citation contexts extraction
- API enrichment (e.g. finding DOIs, linking to other systems)
- Document type classification
- Deduplication
- Indexing
- Exposing (data dumps, API, FastSync)
How often is the CORE content updated

• Data providers harvested as frequent as hardware allows

• Harvesting time is specified by the CORE scheduler
  • Last time the repository was harvested
  • Repository size
  • Repository location
  • Repository harvesting performance
  • Previous information about harvesting errors

• Schedule functionality reviewed on a regular basis
World’s largest dataset of Open Access full texts

- 14,389,274 Hosted full texts
- 24,936,921 Access to free to read full texts
- 135,539,113 Metadata records
- 9,645 Data providers
• January 2019 – CORE reached **over 10M monthly active users** for the first time
• 571% increase from January 2018
CORE Usage

- September 2019 – CORE within top 5k websites globally by user engagement.
- A combination of daily visitors and page views on a website over a 3 month period.
- core.ac.uk by usage in the top 0.0009% of global websites.
Discovery landscape

• Search services
  • Title, author, keywords, etc.
  • And integrations into them

• Recommender services
  • Similar resources

• OA Discovery/Delivery services
  • Overcoming publishers’ paywalls by discovering open access versions of papers
CORE Search

- Full text search for OA content
- Faceted searching
- What you find is what you get
- Real change of data providers wanting to be included
CORE Recommender

• Recommending relevant content to users from across all free content
• Recommender plugin for repositories
• https://core.ac.uk/services/recommender/
**CORE Recommender**

- Recommending relevant content to users from across all free content
- Recommender plugin for repositories
- [https://core.ac.uk/services/recommender/](https://core.ac.uk/services/recommender/)
Introducing CORE Discovery

• High coverage of freely available content
• Free service for researchers by researchers. No company controlling the pipes.
• Best grip on open repository content.
• Repository integration
• Discovering documents without a DOI.

Beat the paywall
Read for FREE

https://core.ac.uk/services/discovery/
CORE Discovery demonstration
CORE Discovery Repository integration

- Majority of articles in repositories metadata only.
- CORE Discovery repository plugin:
  - turns dead ends of user journeys into journeys fulfilling users’ information needs
  - makes repository content more discoverable.
Example integrations with existing library systems

Microsoft Academic

Scientific and ethical issues related to deep brain stimulation for disorders of mood, behavior, and thought

2009 ARCHIVES OF GENERAL PSYCHIATRY Volume: 66, Issue: 9, pp 931-937 DOI: 10.1001/archgenpsychiatry.2009.113
P. Rubin, B. Appleby, J. Brandt, M. DeLong
1 Johns Hopkins University, 2 Emory University, 3 University of California, San Francisco, 4 UZ Ghent University, 5 Brown University, 6 University of Rochester, 7 Minnesota Center for Health Care Ethics, 8 Cleveland Clinic

Context A 2-day consensus conference was held to examine scientific and ethical issues in the application of deep brain stimulation for treating mood and behavioral disorders, such as major depression, obsessive-compulsive disorder, and Tourette syndrome. Objectives The primary objectives of the conference were to (1) establish consensus among participants about the design of future clinical trials of deep brain stimulation for disorders of mood, behavior, and thought and (2) develop standards for the protection of human subjects participating in such studies. Results Conference participants identified 16 key points for guiding research in this growing field. Conclusions The adoption of the described guidelines would help to protect the safety and rights of research subjects who participate in clinical trials of deep brain stimulation for disorders of mood, behavior, and thought and have further potential to benefit other stakeholders in the research process, including clinical researchers and device manufacturers. That said, the adoption of the guidelines will require broad and substantial commitment from many of these same stakeholders.
Ongoing integrations ...

• Ontochem, Wheesbee and others
  • CORE articles searchable from within specialised search engines

• PMC LinkOut
  • Full texts of CORE articles to be discoverable from within PMC.

• ProQuest Summon and Primo
  • Singed contract and awaiting from technical work to be commenced

• Lean Library
  • Recent integration with Lean Library’s content.
Increasing discoverability of content

• Experiment: March 2019, 21 UK institutions
• Logged downloads statistics via CORE and from repository
• On average CORE increases discoverability of content by 15%
• For University of Lincoln 32.5% increase of downloads due to CORE
• Better results for repositories supporting interoperability standards
CORE’s raw data services
Raw data services – CORE API

• Enables the development of new applications
• Real-time machine access to the world's largest collection of open access papers
• Harmonised access to data from across the network of CORE providers
• Direct *machine access to full texts* of research papers
Raw data services – CORE Dataset

• Download millions of research papers for text and data analysis
• Prototype, analyse and mine your data in your infrastructure
Raw data services – CORE FastSync

• Keeps your data in sync with research content from around the world
• Fast and incremental updates as soon as they become available. No usage restrictions
• Based on ResourceSync
Use powered by CORE

- It is beyond human capacities to read all scientific literature
- Example use cases in which CORE is applied:
  - Improving discovery
  - Plagiarism detection
  - Question answering in science
  - Literature based discovery
  - Fact checking and detection of misinformation
  - Analysing research trends
  - Finding experts in a particular domain
  - Research evaluation and scientometrics
  - Exploratory and visual search
  - ...
Types of collaborations

Companies

HEIs

Funders

IRIS.AI

Microsoft

ontochem

IT SOLUTIONS

UNIVERSITY OF CAMBRIDGE

Open Access Button

Research England

COAR

The University Of Sheffield
CORE wins an award

• Outstanding Impact of Research on Society and Prosperity Award 2019

• Greatly motivated to serve the community even more!
CORE Opportunities

• Monitor compliance with Plan S
  • Help institutions to comply as well as monitor their compliance

• Growing demand for raw data access services
  • Help companies to develop innovative services with OA research papers

• Development of products to serve the needs of HEIs
  • Help institutions to increase the discoverability of their research outputs via CORE services (Recommender, Discovery, Search, integrations with other systems, etc.)
  • Make repositories more engaging
Take home

• Data providers (repositories, preprint servers, journals, etc.) and aggregators need to work together to allow text and data analysis, processing and reuse of large volumes of research papers.
• CORE provides the tools for programmatically processing open access data fast, reliably and from across the global network of repositories.
• If you are a repository manager or a librarian:
  • CORE Discovery and Recommender
• If you are a developer or analyst:
  • Build your own stuff using CORE’s data services on top of the global full text open access corpus
CORE looks for Ambassadors

Would you like to become a CORE Ambassador?

How HEIs can help with CORE?

Are you a repository manager or a librarian?

- Make sure that you are a CORE data provider
- Use CORE Search
- Integrate CORE Search in Library Guides
- Install CORE Discovery and CORE Recommender
- Promote Open Access with the use of Open Licenses

Are you a developer or analyst:

- Build your own stuff using CORE’s raw data services on top of the global full text open access corpus
Thank you!

https://core.ac.uk