CORE: Improving access and enabling re-use of open access content using aggregations

Petr Knoth
CORE (Connecting REpositories)
Knowledge Media institute
The Open University

@petrknoth

Outline

1. The need for aggregationg Open Access content
2. The CORE system

Outline

1. **The need for aggregationg Open Access content**
2. The CORE system

What is Open Access exactly?

By "open access" to [peer-reviewed research 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.

[BOAI, 2002]

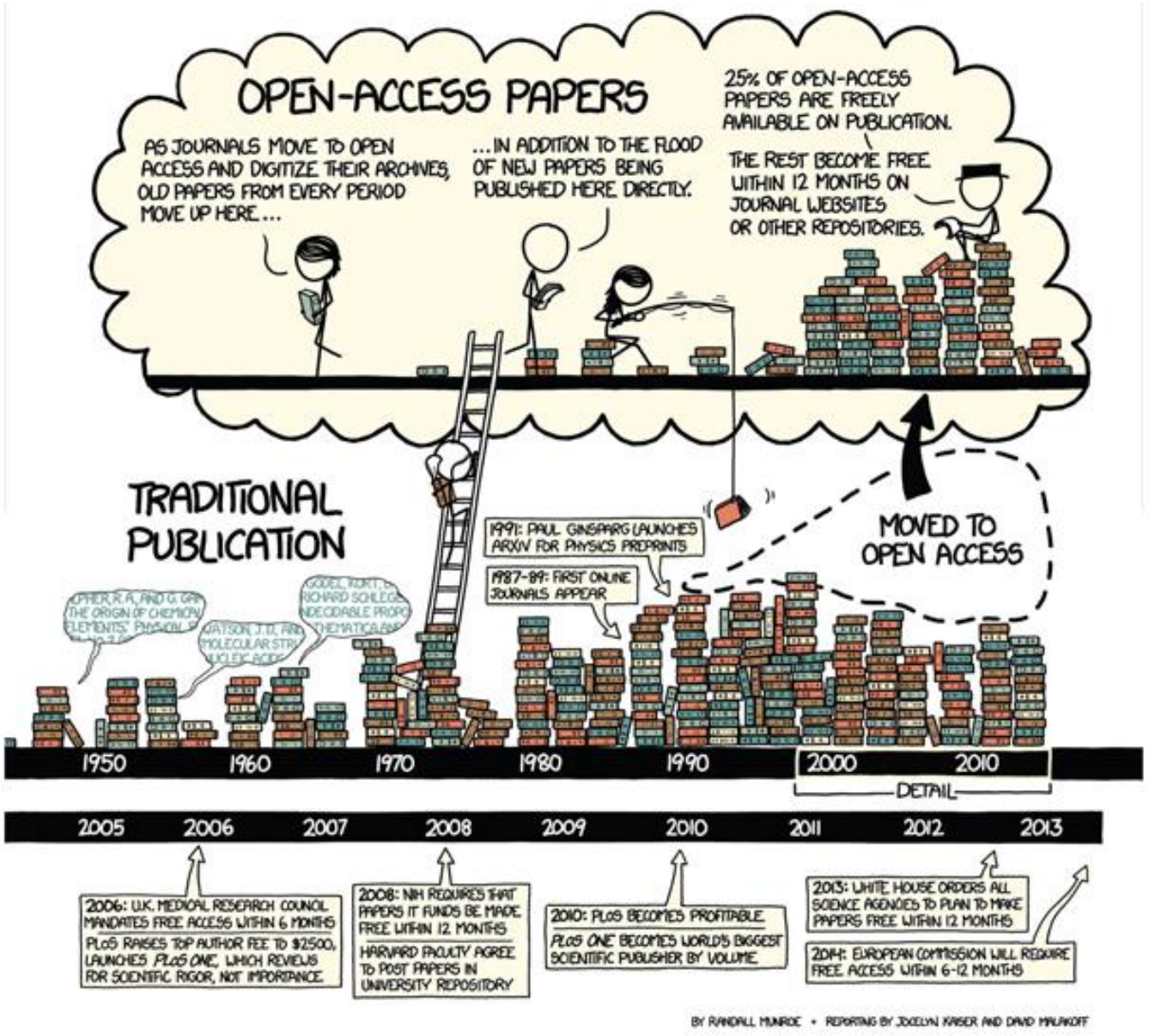
Open Access = Access + **Reuse**

How to achieve OA?

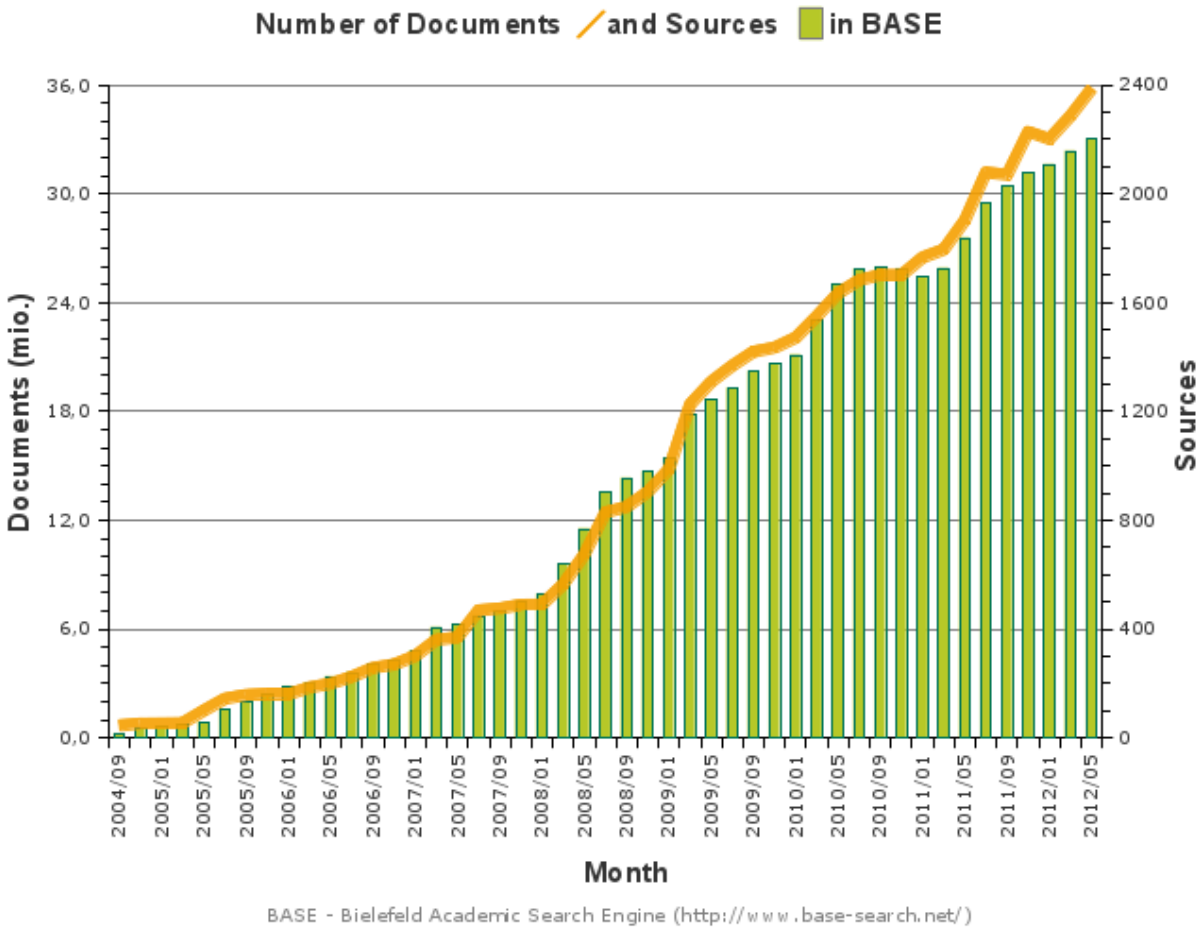
Two routes:

- Self-archiving: Institional/Open Repositories
- Open Access Journals

OA growth


OPEN-ACCESS PAPERS
AS JOURNALS MOVE TO OPEN ACCESS AND DIGITIZE THEIR ARCHIVES, OLD PAPERS FROM EVERY PERIOD MOVE UP HERE...
...IN ADDITION TO THE FLOOD OF NEW PAPERS BEING PUBLISHED HERE DIRECTLY.
25% OF OPEN-ACCESS PAPERS ARE FREELY AVAILABLE ON PUBLICATION.
THE REST BECOME FREE WITHIN 12 MONTHS ON JOURNAL WEBSITES OR OTHER REPOSITORIES.
TRADITIONAL PUBLICATION
MOVED TO OPEN ACCESS
1991: PAUL GINSPARG LAUNCHES ARXIV FOR PHYSICS PREPRINTS
1987-89: FIRST ONLINE JOURNALS APPEAR
1950
1960
1970
1980
1990
2000
2010
DETAIL
2005
2006
2007
2008
2009
2010
2011
2012
2013
2006: U.K. MEDICAL RESEARCH COUNCIL MANDATES FREE ACCESS WITHIN 6 MONTHS
PLOS RAISES TOP AUTHOR FEE TO $2500, LAUNCHES PLOS ONE, WHICH REVIEWS FOR SCIENTIFIC RIGOR, NOT IMPORTANCE
2008: NIH REQUIRES THAT PAPERS IT FUNDS BE MADE FREE WITHIN 12 MONTHS
HARVARD FACULTY AGREE TO POST PAPERS IN UNIVERSITY REPOSITORY
2010: PLOS BECOMES PROFITABLE
PLOS ONE BECOMES WORLD'S BIGGEST SCIENTIFIC PUBLISHER BY VOLUME
2013: WHITE HOUSE ORDERS ALL SCIENCE AGENCIES TO PLAN TO MAKE PAPERS FREE WITHIN 12 MONTHS
2014: EUROPEAN COMMISSION WILL REQUIRE FREE ACCESS WITHIN 6-12 MONTHS
BY RANDALL MUNROE • REPORTING BY JOCELYN KAISER AND DAVID MALAKOFF

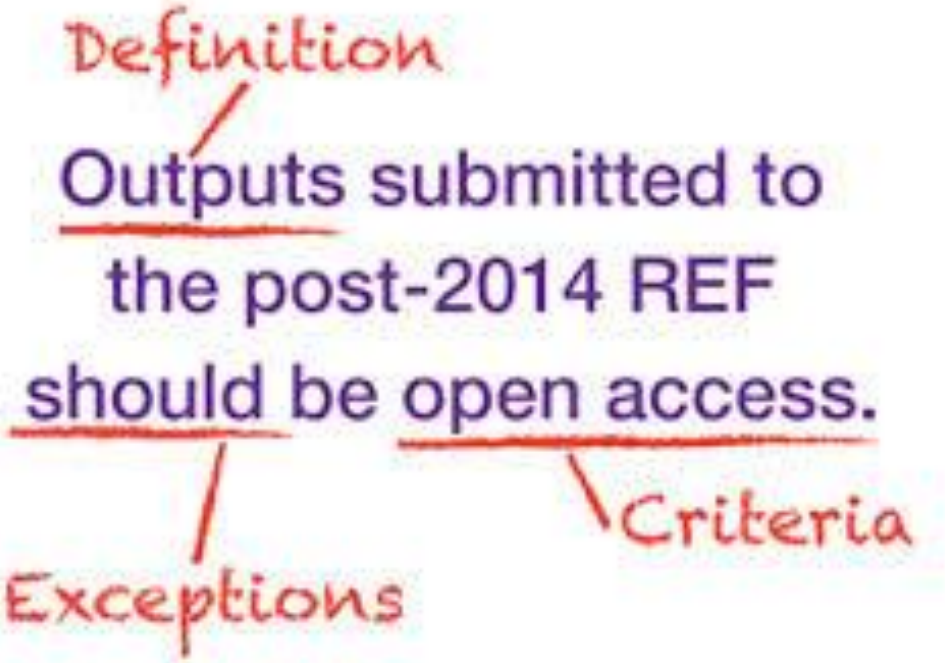
Growth of items in Open Access repositories



Records stored across all OARs

164,259,752 records across 2,531 repositories as estimated by OpenDOAR

The aim of the open access (OA) post-2014 REF policy



COAR: About harvesting and aggregations …

"Each individual repository is of limited value for research: the real power of Open Access lies in the possibility of connecting and tying together repositories, which is why we need interoperability. In order to create a seamless layer of content through connected repositories from around the world, Open Access relies on interoperability, the ability for systems to communicate with each other and pass information back and forth in a usable format. 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."

[COAR manifesto]

SPARC's position paper on IRs

"For the repository to provide access to the broader research community, users outside the university must be able to find and retrieve information from the repository. Therefore, institutional repository systems must be able to support interoperability in order to provide access via multiple search engines and other discovery tools. An institution does not necessarily need to implement searching and indexing functionality to satisfy this demand: it could simply maintain and expose metadata, allowing other services to harvest and search the content. This simplicity lowers the barrier to repository operation for many institutions, as it only requires a file system to hold the content and the ability to create and share metadata with external systems."

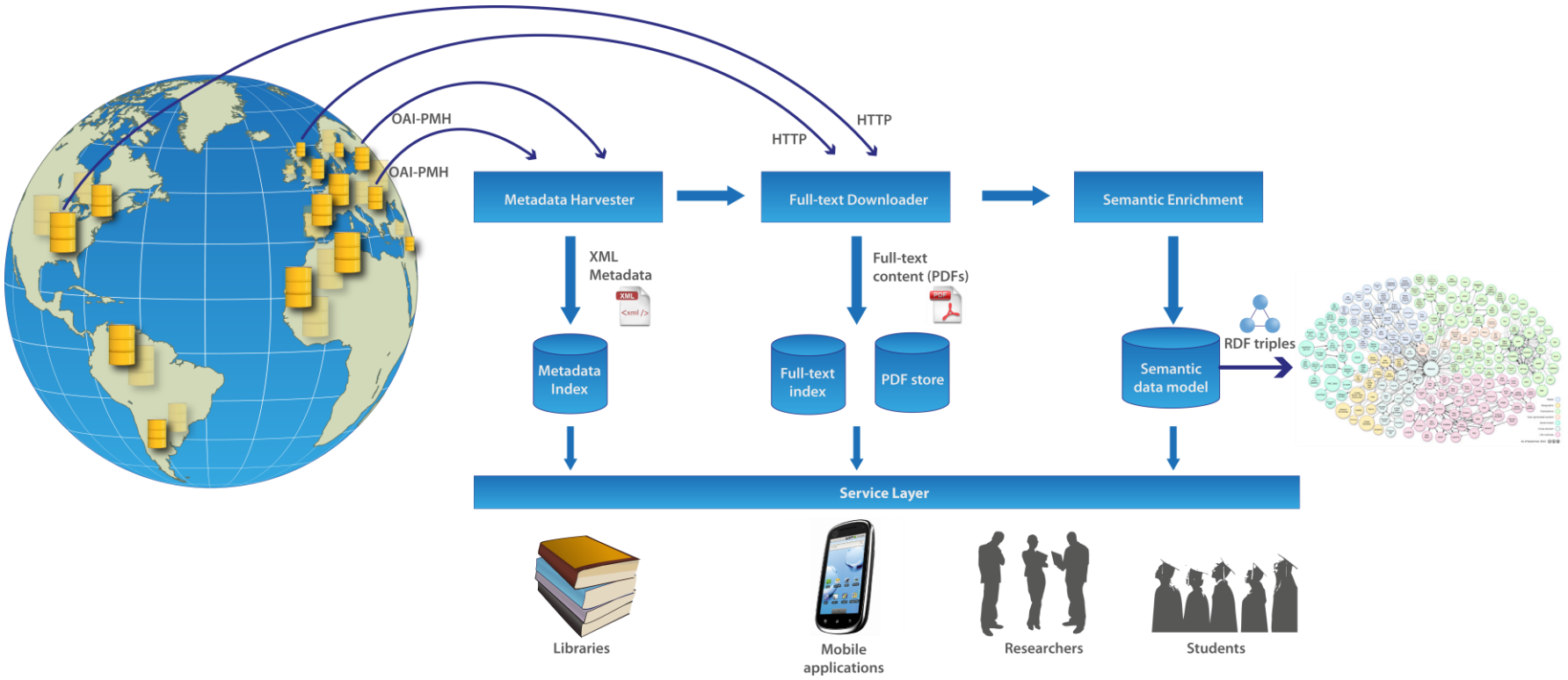
Outline

1. The need for aggregationg Open Access content
2. **The CORE system**

The mission of CORE

Aggregate all open access content ditsributed across different systems worldwide, enrich this content and provide access to it through a set of services …

The CORE aggregator



Processing pipeline

- Metadata download, extraction and cleaning
- Full-text harvesting
- Text-extraction
- Language detection
- Extraction of citation references from text
- Detection of citation reference targets
- Identification of related content
- Detection of duplicate items
- Parsing of author names
- Indexing

CORE statistics

- Content: 18M+ records, 600+ repositories, 1.8M+ full-texts
- The world's largest full-text open access dataset and still growing
- The UK national aggregator (part of Repositories Shared Services project - Jisc)
- Full-text aggregator (not just metadata)
- Over 0.5 million monthly visits
- Placed among Top 10 search engines for research that go beyond Google [JISC, 2013]
- Listed among Top 100 Thesis and Dissertation Resources
- Used by many researchers and organisations, including the European Library and UNESCO, and projects, such as the Open Access Button project

CORE supports a three access levels architecture

- *Raw data access.*
- *Transaction information access.*
- *Analytical information access.*

CORE supports a three access levels architecture

- *Raw data access.* Developers, DLs, DL researchers, companies …
- *Transaction information access.* Researchers, students, life-long learners …
- *Analytical information access.* Funders, government, bussiness intelligence …

CORE supports a three access levels architecture

- *Raw data access.* Developers, DLs, DL researchers, companies …

Apps: CORE API, CORE Data Dumps

- *Transaction information access.* Researchers, students, life-long learners …

Apps: CORE Portal, CORE Mobile, CORE (recommendation) Plugin

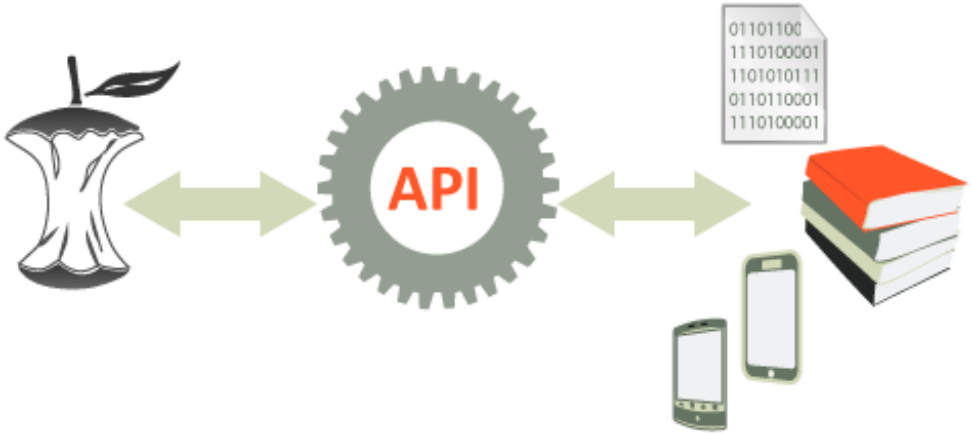
- *Analytical information access.* Funders, government, bussiness intelligence …

Apps: Repository Analytics, CORE Policy Compliance Analytics

CORE API

Enables external systems to interact with OA data (JSON or XML)

- Search, download metadata and cotent
- Content recommendation
- Citation references
- Statistics
- …



Used by: Libraries, Institutional repositories, developers

Data dumps

- Cleaned and enriched with additional information
- Distributed as two large zip files: metadata + full-texts
- Created as part of the Digging into Connected Repositories (DiggiCORE) project

Examples of usage

- Author disambiguation
- Mining URLs from papers to detect trends
- Tagging of chemical compounds for image retrieval
- Citation analysis
- Content recommendation
- Detecting collaboration patterns of scientific communities
- Monitoring of OA growth

- Any form of text or data mining …
- API useful for services and data dumps for offline experiments

Why to use it?

- It is only OA, thus you can legally mine it …
- You can redistribute it: essential for reproducible research
- Very large and growing
- Kept up-to-date
- Ability to rerun experiments with new data

Why to use it?

- Open infrastructure for open science
- Not owned or managed by a for profit company => Ability to run your own services = new opportunities and no give away of your research to commercial companies

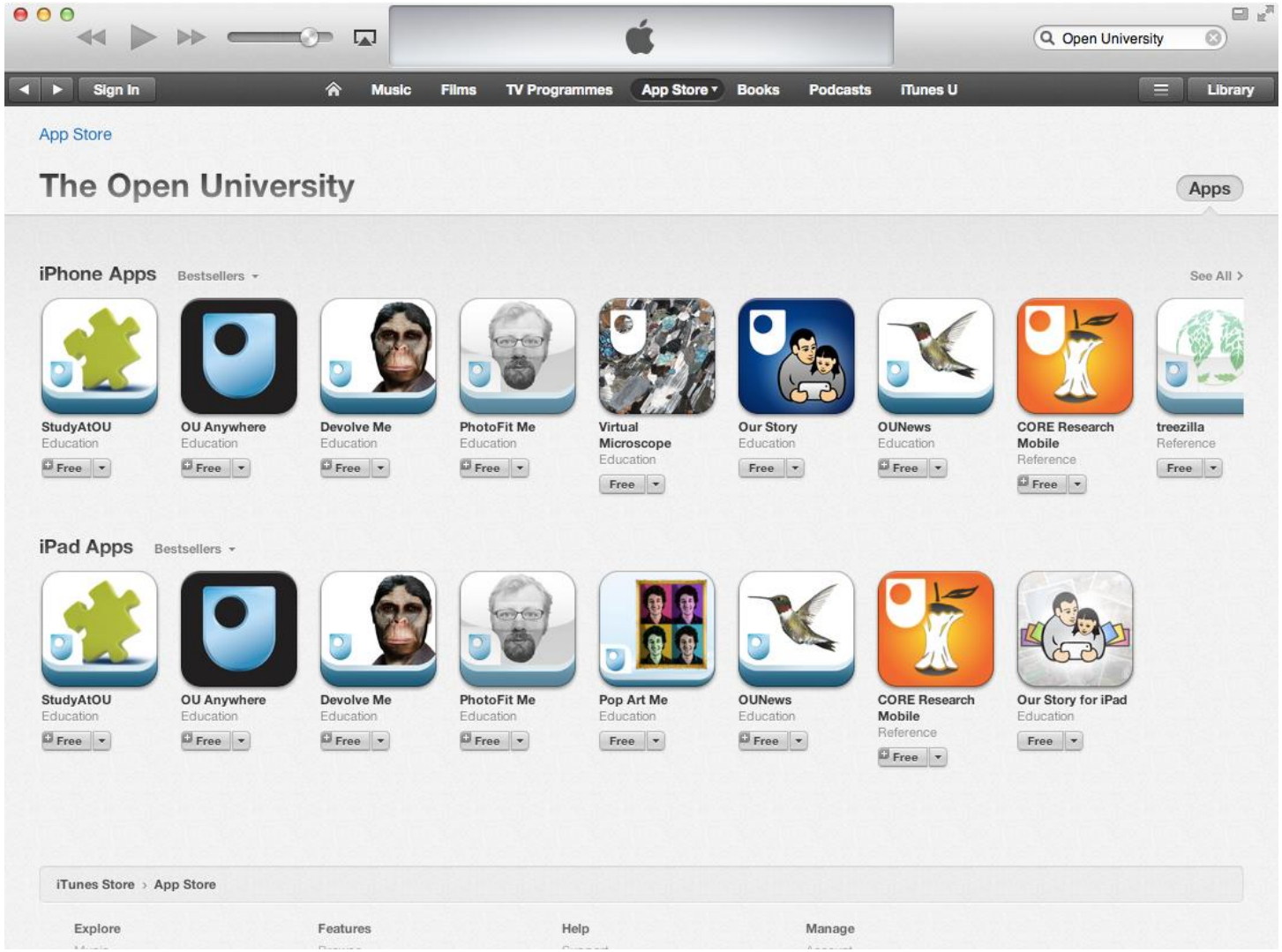
CORE Applications

CORE Portal – Allows searching and navigating scientific publications aggregated from Open Access repositories

CORE Applications

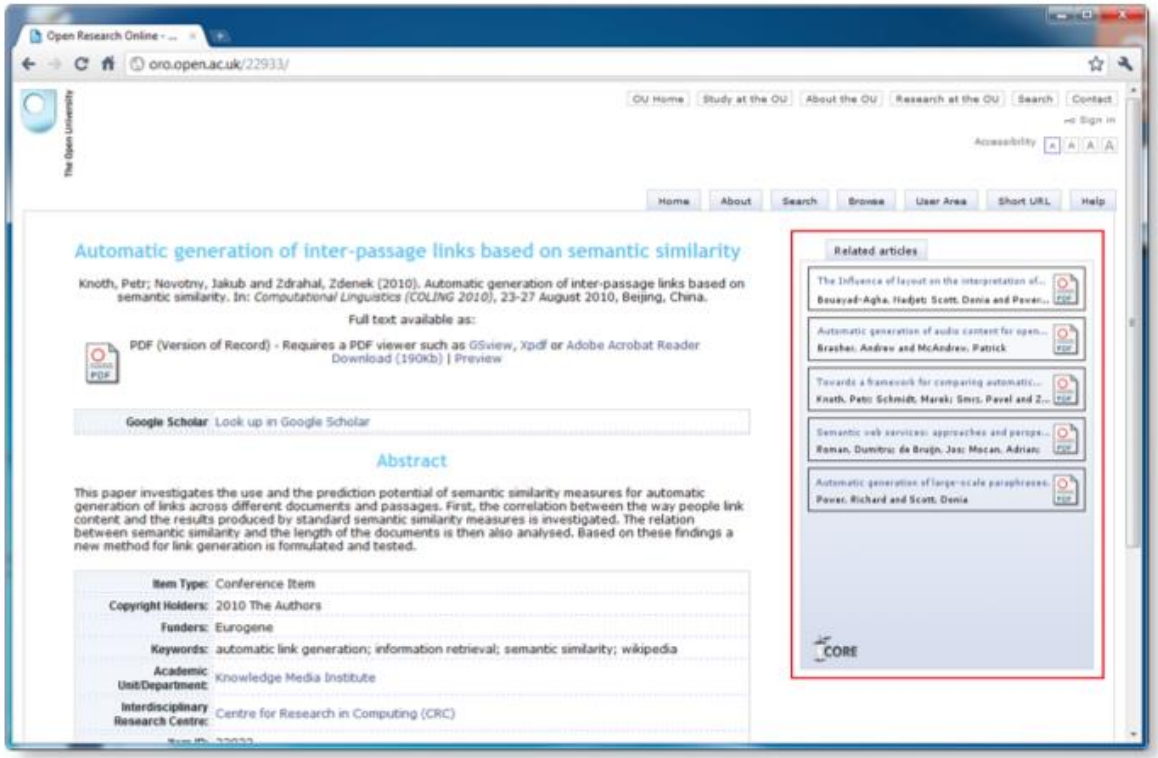
CORE Mobile – Allows searching and navigating scientific publications aggregated from Open Access repositories

CORE Applications



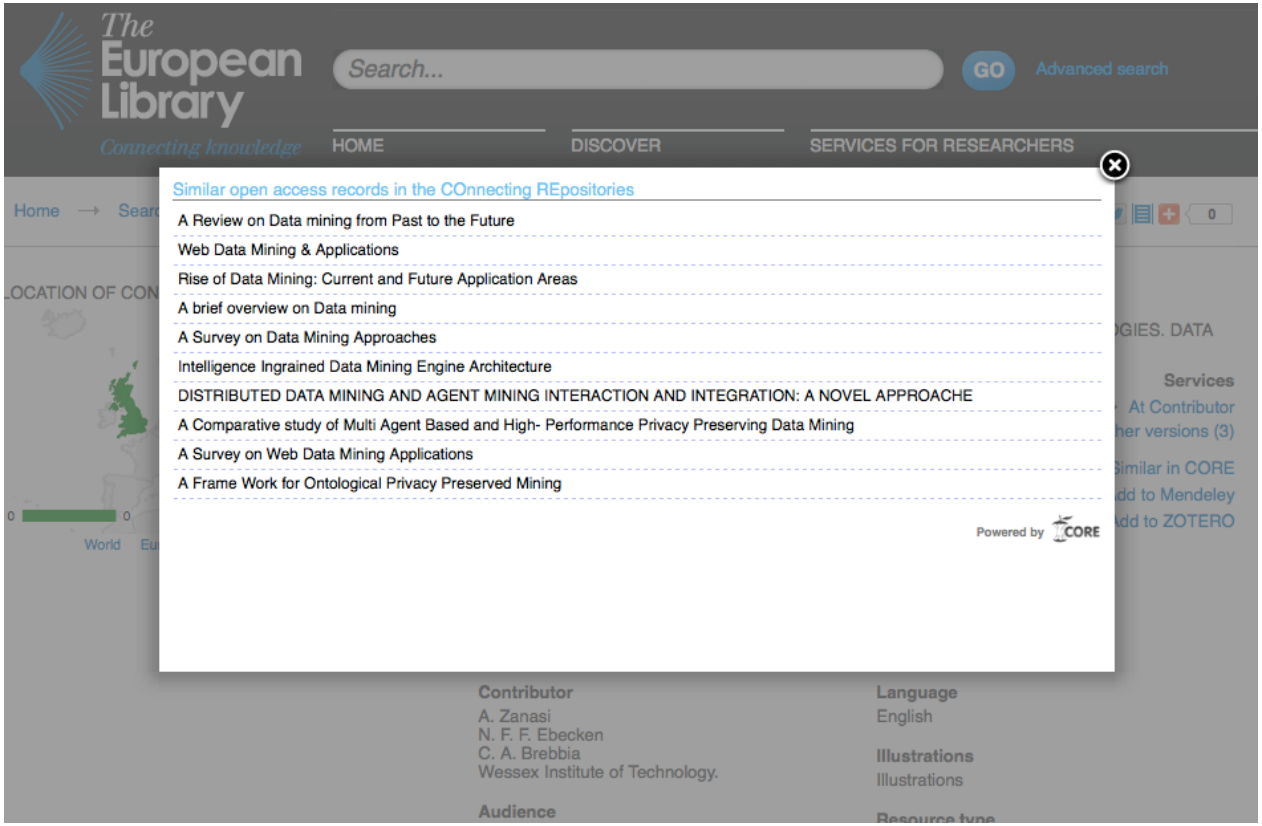
CORE Applications

CORE Plugin – A plugin to system that recommendations for related items.



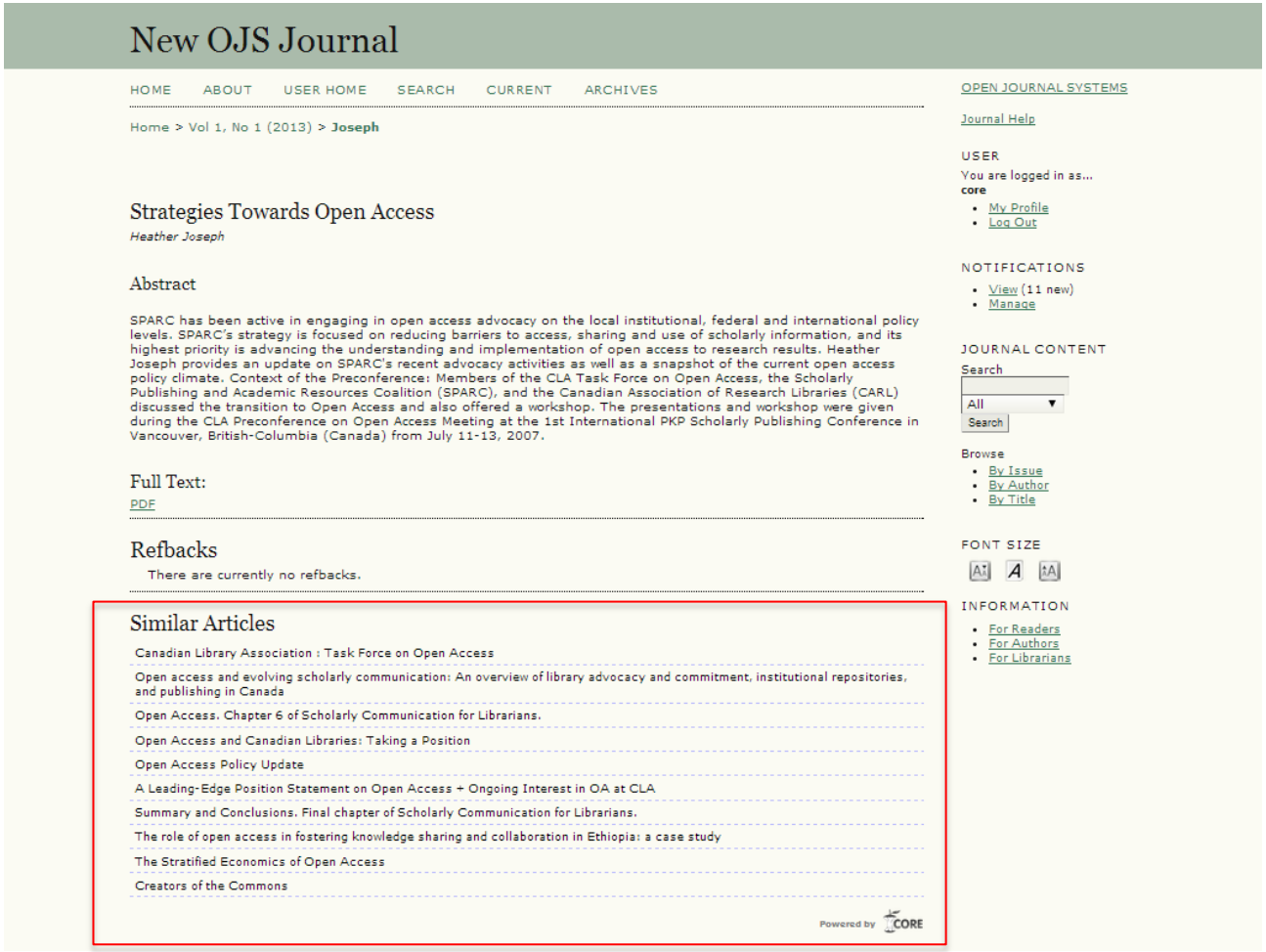
CORE Applications

CORE Plugin – A plugin to system that recommendations for related items.



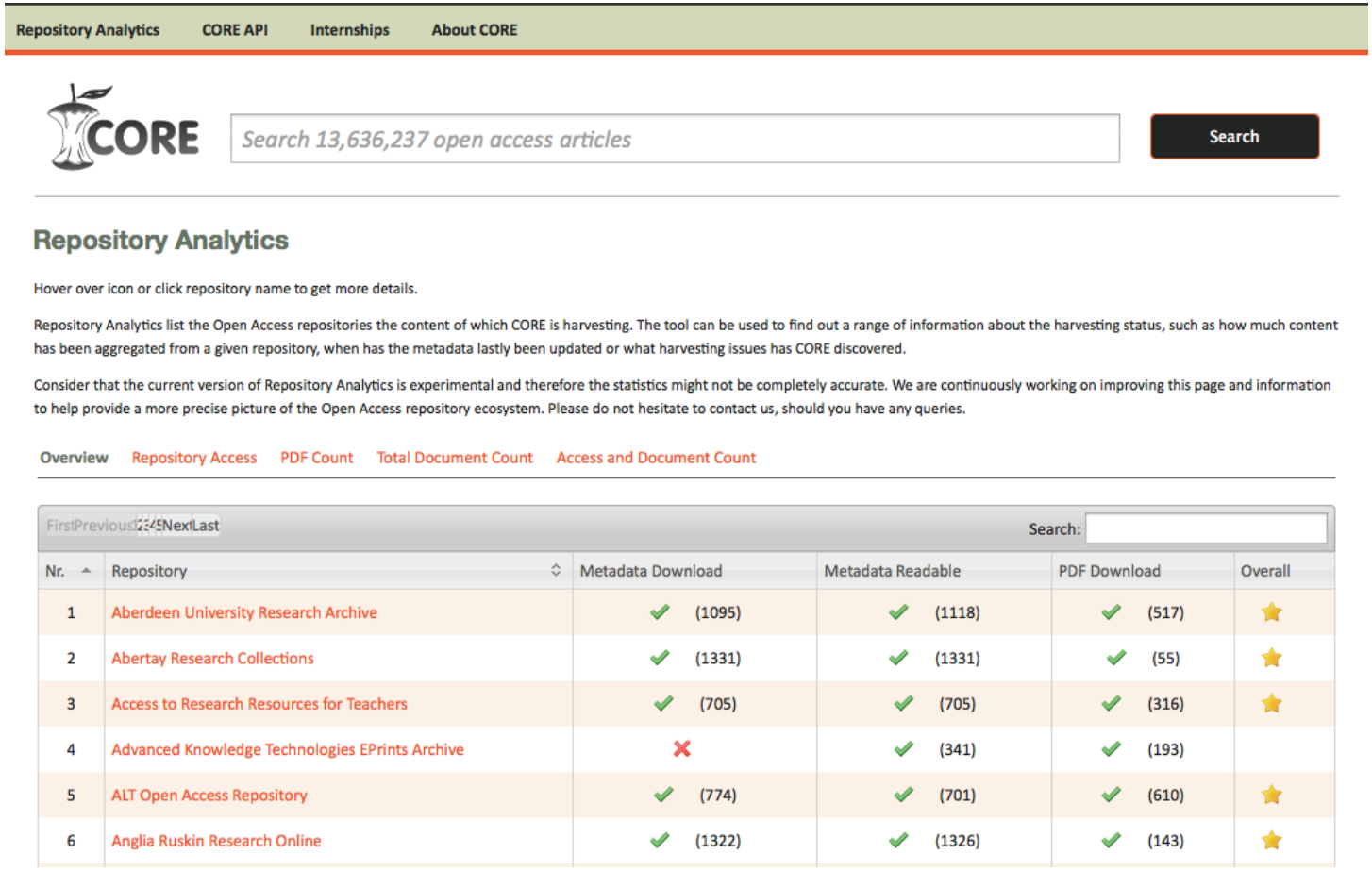
Built on top of CORE API …

CORE Plugin – A cross-repository recommendation system integrated into OJS.



CORE Applications

Repository Analytics – is an analytical tool supporting providers of open access content (in particular repository managers).



Repository Analytics CORE API Internships About CORE

CORE

Search 13,636,237 open access articles

Search

Repository Analytics

Hover over icon or click repository name to get more details.

Repository Analytics list the Open Access repositories the content of which CORE is harvesting. The tool can be used to find out a range of information about the harvesting status, such as how much content has been aggregated from a given repository, when has the metadata lastly been updated or what harvesting issues has CORE discovered.

Consider that the current version of Repository Analytics is experimental and therefore the statistics might not be completely accurate. We are continuously working on improving this page and information to help provide a more precise picture of the Open Access repository ecosystem. Please do not hesitate to contact us, should you have any queries.

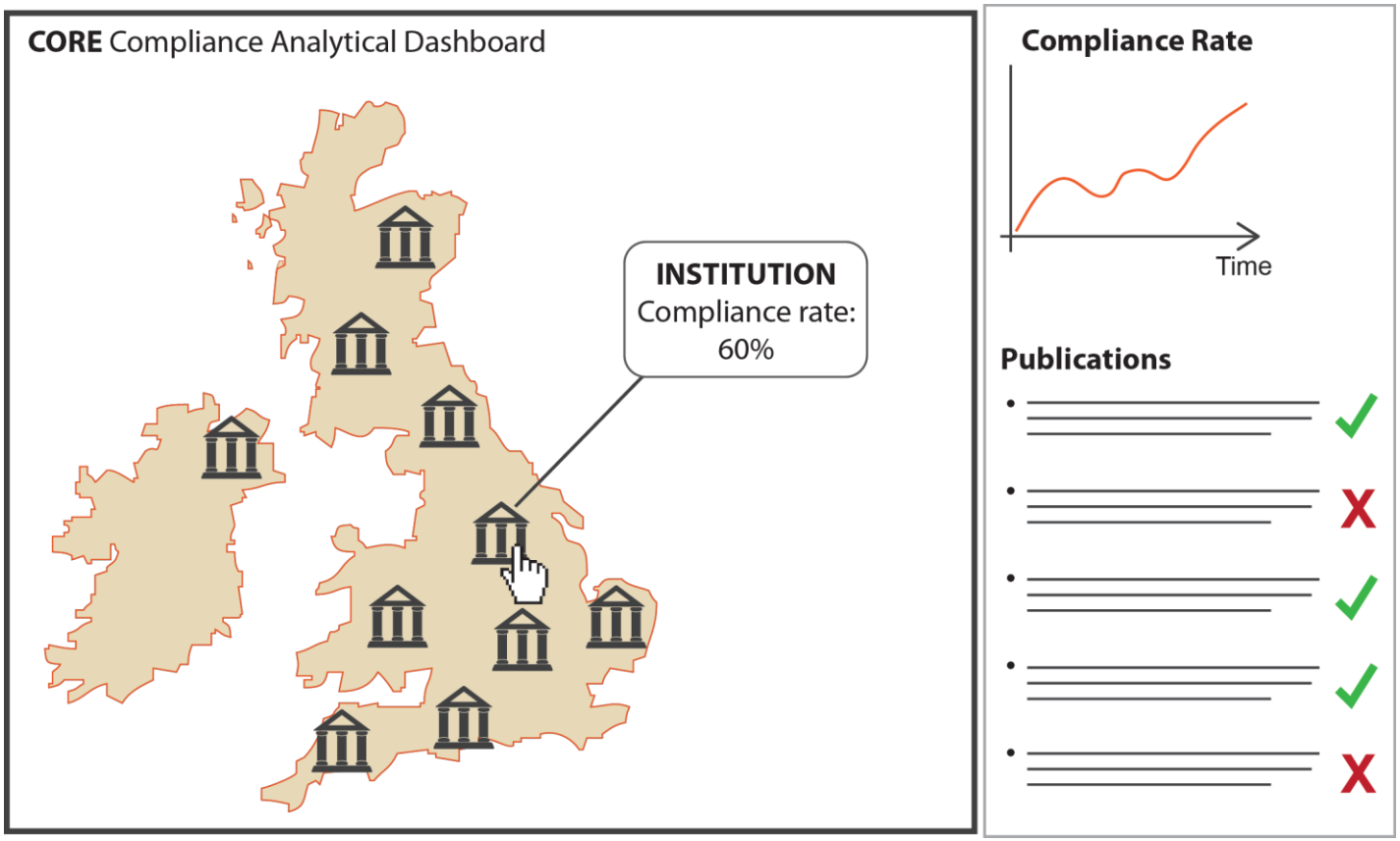
Overview Repository Access PDF Count Total Document Count Access and Document Count

First Previous 1 2 3 4 5 Next Last Search:

Nr.	Repository	Metadata Download	Metadata Readable	PDF Download	Overall
1	Aberdeen University Research Archive	✓ (1095)	✓ (1118)	✓ (517)	★
2	Abertay Research Collections	✓ (1331)	✓ (1331)	✓ (55)	★
3	Access to Research Resources for Teachers	✓ (705)	✓ (705)	✓ (316)	★
4	Advanced Knowledge Technologies EPrints Archive	✗	✓ (341)	✓ (193)	
5	ALT Open Access Repository	✓ (774)	✓ (701)	✓ (610)	★
6	Anglia Ruskin Research Online	✓ (1322)	✓ (1326)	✓ (143)	★

CORE Applications

Policy Compliance Analytics (under development) – Tool to support the implementation and monitoring of the UK HEFCE OA policy.



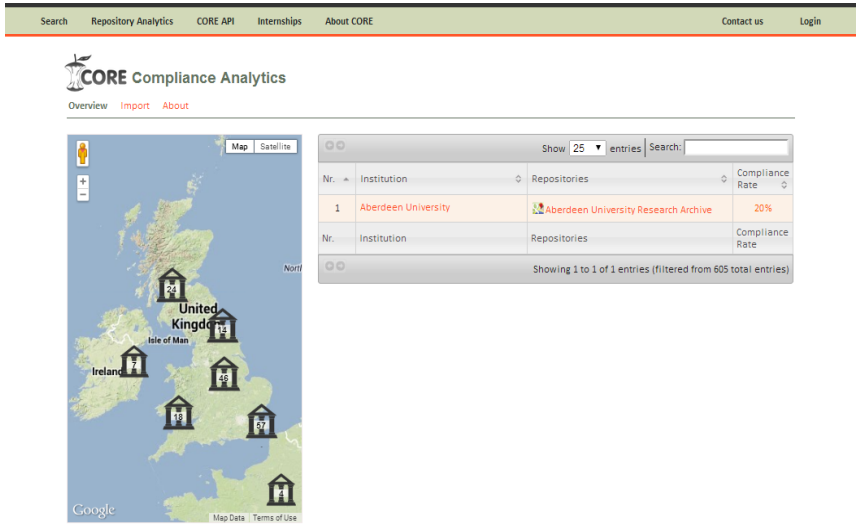
KMi

The definition of OA for post-2014 REF

Consultation on open access in the post-2014 Research Excellence Framework, paragraph 25 says that:

- Accessible through a UK HEI repository (immediately upon acceptance or publication).
- Made available as the final peer-reviewed text (full-text) after a (reasonable) embargo period specified by the publisher.
- Harvestable using automated tools.
- In a machine readable form to allow text-mining
- Unambiguously identifiable in the institutional repository, including items available through a link to another website.

The developed tool



Working for the repositories community

- Aggregations and usage statistics (CORE vs IRUS-UK collaboration)
- Support to repository managers through standards validation tools (e.g. RIOXX or OpenAIRE compliance)
- Support for OA mandates verification processes (e.g. HEFCE)
- Content statistics benchmarking
- Detection of duplicate items across repositories
- Growth monitoring

Conclusions

- Open Access knowledge available online on the rise
- The OA infrastructure (repositories, aggregators) must enable efficient re-use
- CORE provides a single access point to this knowledge and enables its mining
- Opportunities for innovative applications and research

BUT

- The OA infrastructure should be here for the benefit of all and should not be owned by the publishing lobby.
- Aggregations should work hand in hand with the repositories community

Thank you!

CORE: The single access point to open knowledge from repositories worldwide

References 1/2

[BOAI, 2002] Budapest Open Access Initiative. (2002) http://www.opensocietyfoundations.org/openaccess/boai-10-recommendations

[Crow, 2002] Crow, R. (2002). The case for institutional repositories: a SPARC position paper. *ARL Bimonthly Report 223*.

[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, http://dx.doi.org/10.1045/november2012-knoth

[Konkiel, 2012] Konkiel, S. (2012) Are Institutional Repositories Doing Their Job? https://blogs.libraries.iub.edu/scholcomm/2012/09/11/are-institutional-repositories-doing-their-job/

[Laakso & Bjork, 2012] Laakso, M., & Björk, B. C. (2012). Anatomy of open access publishing: a study of longitudinal development and internal structure. *BMC Medicine*, *10*(1), 124.

References 2/2

[Morrison, 2012] Morrison, Louise (2012) 5 reasons why I can't find Open Access publications. http://mmitscotland.wordpress.com/2012/08/06/5-reasons-why-i-cant-find-open-access-publications-2/

[OAI-PMH v2.0, 2008] The Open Archives Initiative Protocol for Metadata Harvesting Version 2.0 (OAI-PMH), Impementation Guidelines (2008). http://www.openarchives.org/OAI/openarchivesprotocol.html

[ResourceSync draft, 2013] ResourceSync protocol draft. 2013 http://www.niso.org/workrooms/resourcesync/

[Salo, 2008] Salo, D. (2008). Innkeeper at the roach motel. *Library Trends*, *57*(2), 98-123.

[Van de Sompel et al, 2004] Van de Sompel, H., Nelson, M. L., Lagoze, C., & Warner, S. (2004). Resource harvesting within the OAI-PMH framework. *D-lib magazine*, *10*(12), 1082-9873.