

# 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 aggregating Open Access content
2. The CORE system

# Outline

- 1. The need for aggregating 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]

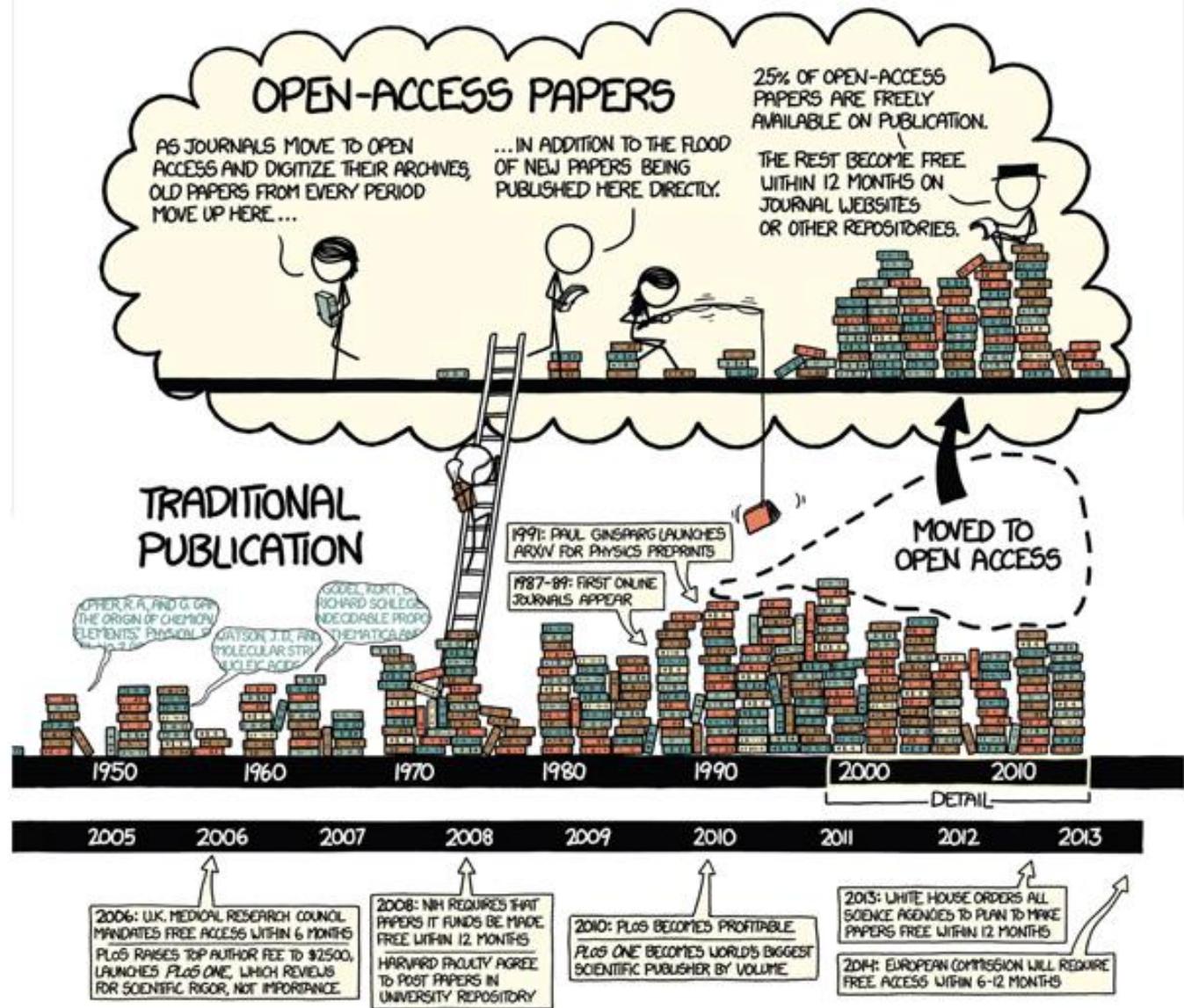
Open Access = Access + **Reuse**

# How to achieve OA?

Two routes:

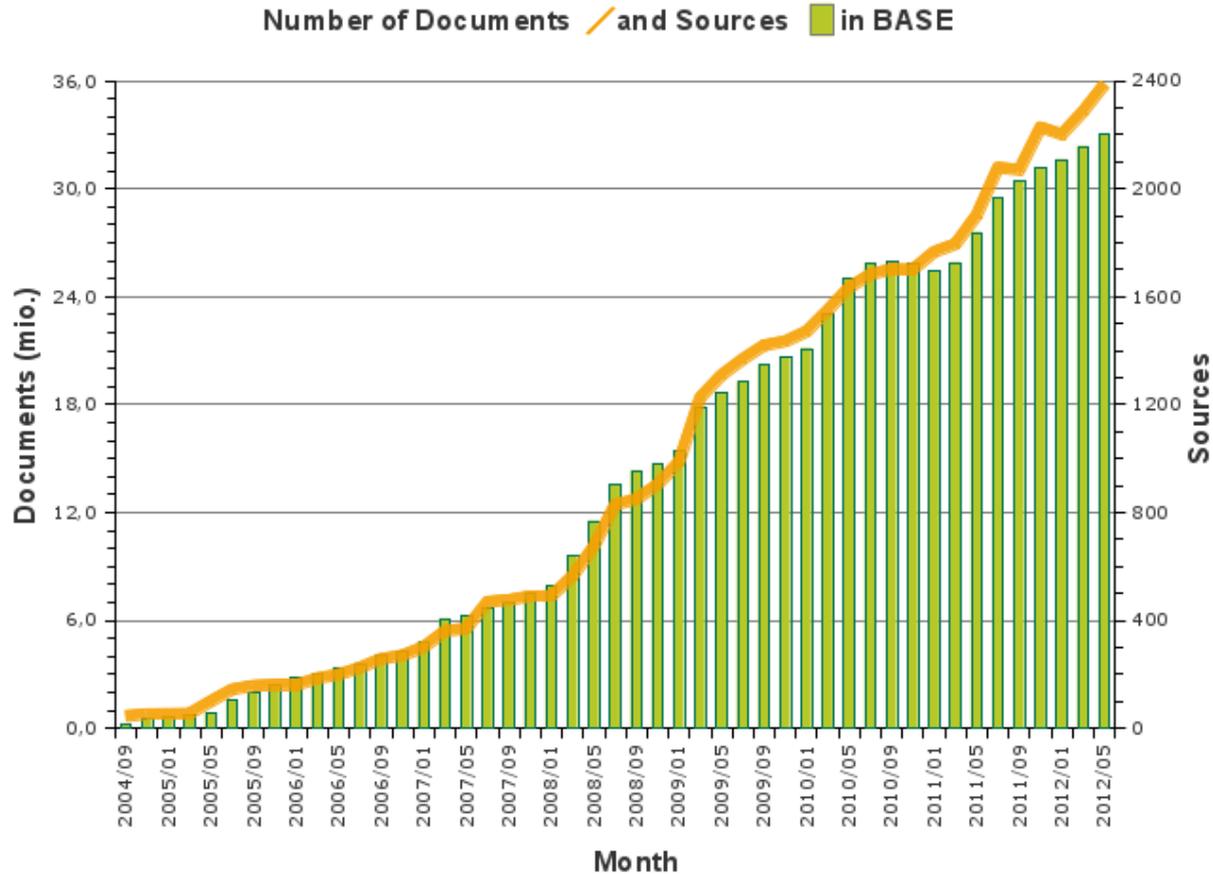
- Self-archiving: Institutional/Open Repositories
- Open Access Journals

# OA growth



BY RANDALL MURKIN • REPORTING BY JOCELYN KASER AND DAVID MUKHOFF

# Growth of items in Open Access repositories

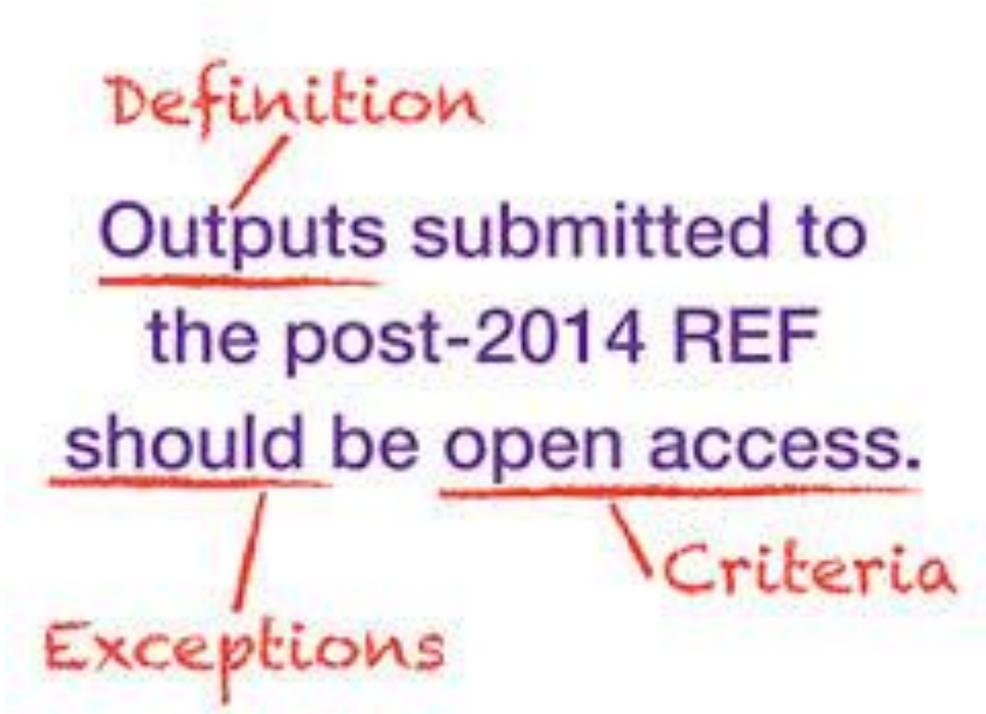


BASE - Bielefeld Academic Search Engine (<http://www.base-search.net/>)

## 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.”*

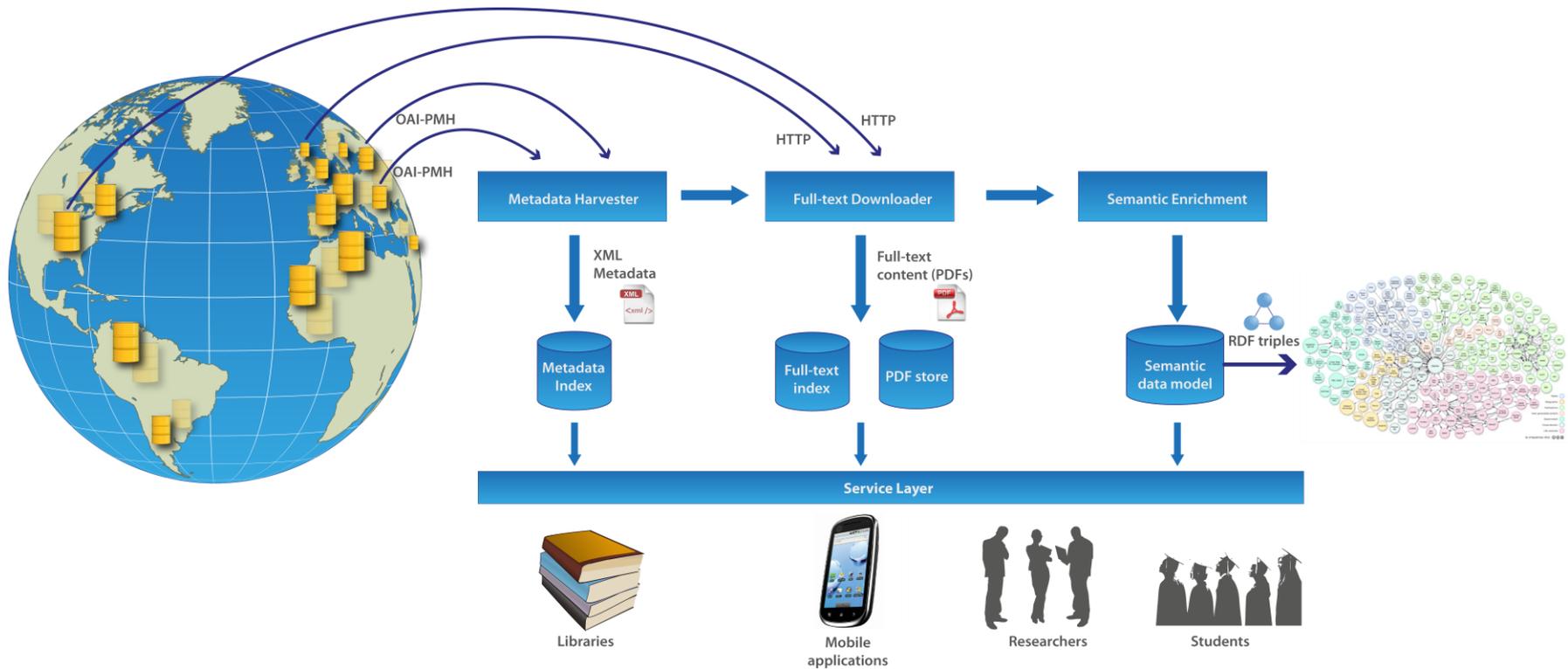
# Outline

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

## The mission of CORE

*Aggregate all open access content distributed 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*

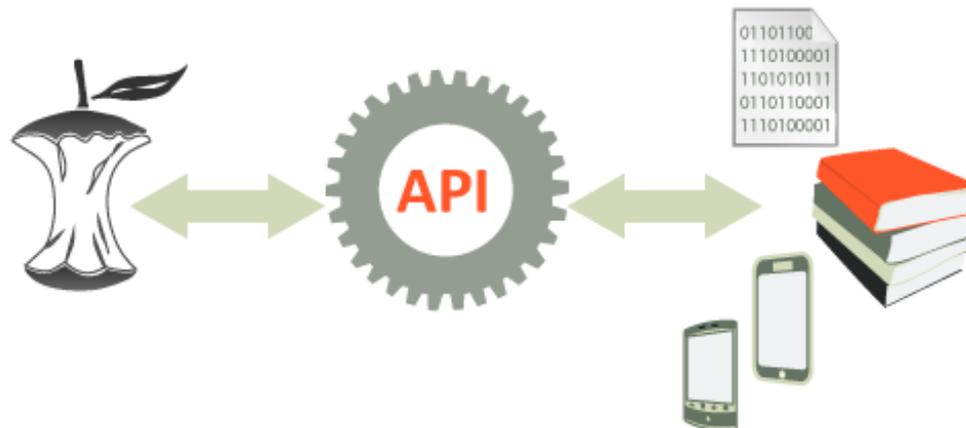
- *Analytical information access.* Funders, government, bussiness intelligence ...

*Apps: Repository Analytics, CORE Policy Compliance Analytics*

# CORE API

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

- Search, download metadata and content
- 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



The screenshot shows the CORE website interface. At the top, there is a navigation menu with links for Search, Repository Analytics, CORE API, About CORE, Contact us, and Login. Below the menu is the CORE logo and a search bar with a 'Search' button. The main content area features four application tiles:

- Explore Open Access scholarly papers from across the World.** (Illustrated with a world map and arrows)
- CORE Mobile: Try our mobile application.** (Illustrated with a smartphone and tablet)
- Repository Analytics: CORE analytical tools to support you repository.** (Illustrated with a line graph and an upward arrow)
- CORE API: Offering programmable access to millions of resources.** (Illustrated with a gear and the text 'API')

Below the tiles, there are three columns of text:

- About CORE**: CORE (CConnecting REpositories) aims to facilitate free access to scholarly publications distributed across many systems. As of today, CORE gives you access to millions of scholarly articles aggregated from many Open Access repositories. We believe in free access to information. The mission of CORE is to:
  - Support the right of citizens and general public to access the results of research towards which they contributed by paying taxes.
  - Facilitate access to Open Access content for all by targeting general public, software developers, researchers, etc., by improving search and navigation using state-of-the-art technologies in the field of natural language processing and the Semantic Web.
- Applications**: CORE offers four applications:
  - CORE Portal** - Allows to search and navigate scientific publications aggregated from a wide range of Open Access Repositories (OARs)
  - CORE Mobile** - An Android application that enables you to search and download open access articles.
  - CORE Plugin** - A Plugin to Open Access repositories that enables them to search for related scientific publications.
  - CORE API** - Enables external systems and services to interact with the CORE repository.
- Supporting your repository**: CORE aims to contribute to Open Access by providing services that support Open Access providers. We are offering our repository analytics tools and also try to increase the visibility of content exposed by Open Access providers. CORE can help you to:
  - Increase the visibility of your repository content.
  - Track the harvesting and the usage of your repository content.
  - Enrich the metadata provided by your repository.
  - Enable mobile access to your content.
  - Improve the search and navigation of related papers directly in your repository using the CORE Plugin.

At the bottom of the Applications column, there is a link: [Join CORE repositories](#)

# CORE Applications

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

The screenshot displays the CORE mobile application interface. At the top, there is a navigation bar with the CORE logo and tabs for 'Search', 'Library', and 'History'. To the right of these tabs are icons for 'SAVE', 'FIND SIMILAR', 'SHARE', 'DELETE', and 'PREFERENCES'. Below the navigation bar, a list of search results is shown. The selected result, 'The Semantic Web Revisited', is displayed in a larger view on the right. This view includes the article title, publisher information (e-Prints Soton), authors (Nigel Shadbolt, Tim Berners-Lee and Wendy Hall), and the date (2006-07). A PDF icon and the text 'Document available on server' are also present. Below the article information, an abstract is provided, discussing the evolution of the Semantic Web from 2001 to the present.

**Search Results List:**

- Wendy Hall and Kieron Hara  
e-Prints Soton 2009-05
- Semantic Web**  
Wendy Hall and Kieron Hara  
Electronics & Computer Science EPrints Service - University of Southa
- Notes on Semantic Web Services (Greek langua**  
Thanassis Tiropanis  
Electronics & Computer Science EPrints Service - University of Southa
- Notes on Semantic Web Services (Greek language)**  
Thanassis Tiropanis  
e-Prints Soton 2006-07
- Mining the Semantic Web**  
Mr Chakravarthy  
Advanced Knowledge Technologies EPrints Archive 2005-01-01
- Mining the Semantic Web**  
Ajay Chakravarthy  
Pattern Analysis Statistical Modelling & Computational Learning EPrint
- The Semantic Web Revisited**  
Nigel Shadbolt, Tim Berners-Lee and Wendy Hall  
e-Prints Soton 2006-07

**The Semantic Web Revisited**

Published by  
e-Prints Soton

Authors  
Nigel Shadbolt, Tim Berners-Lee and Wendy Hall

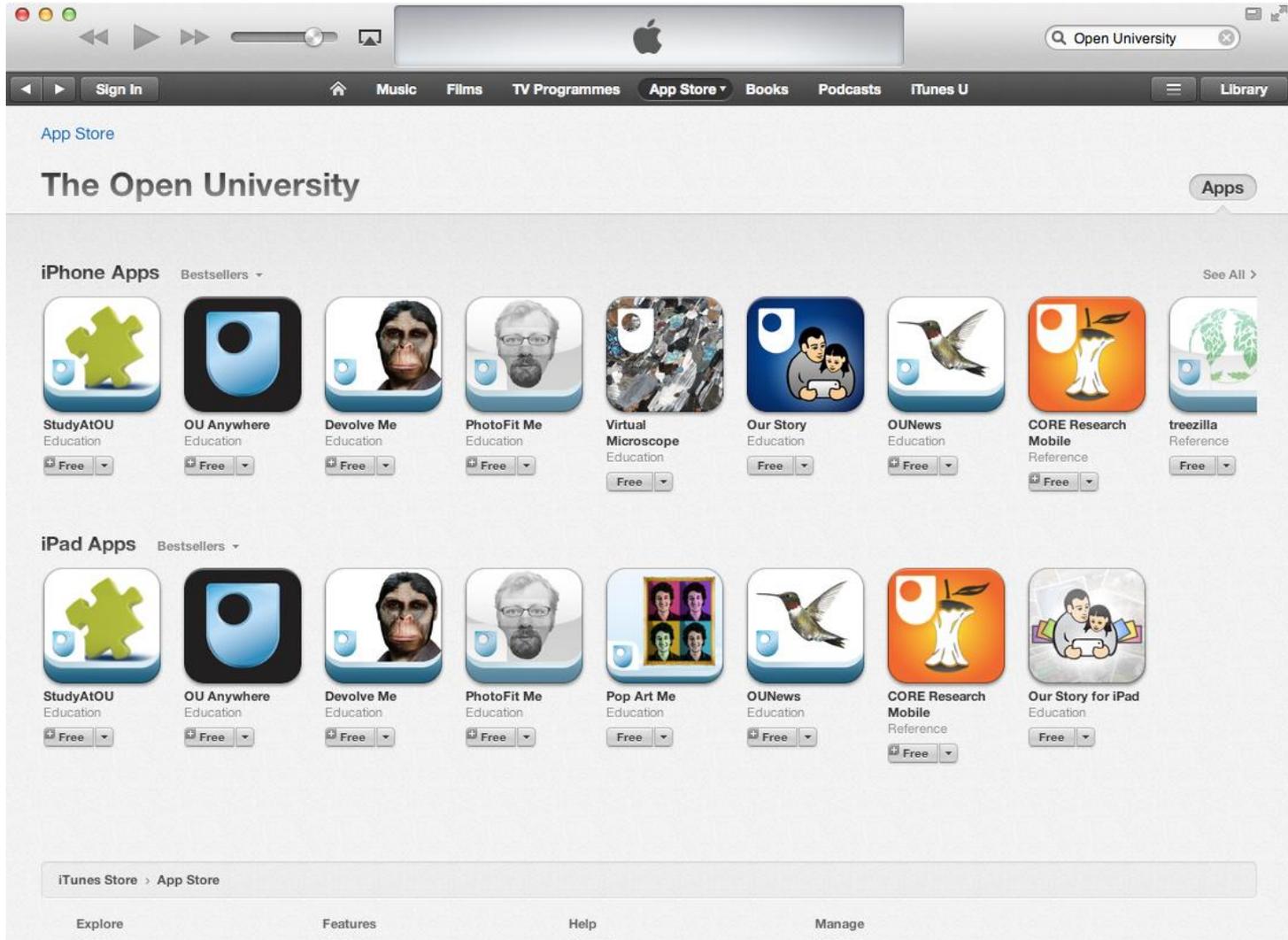
Date  
2006-07

Status  
Document available on server

**Abstract**

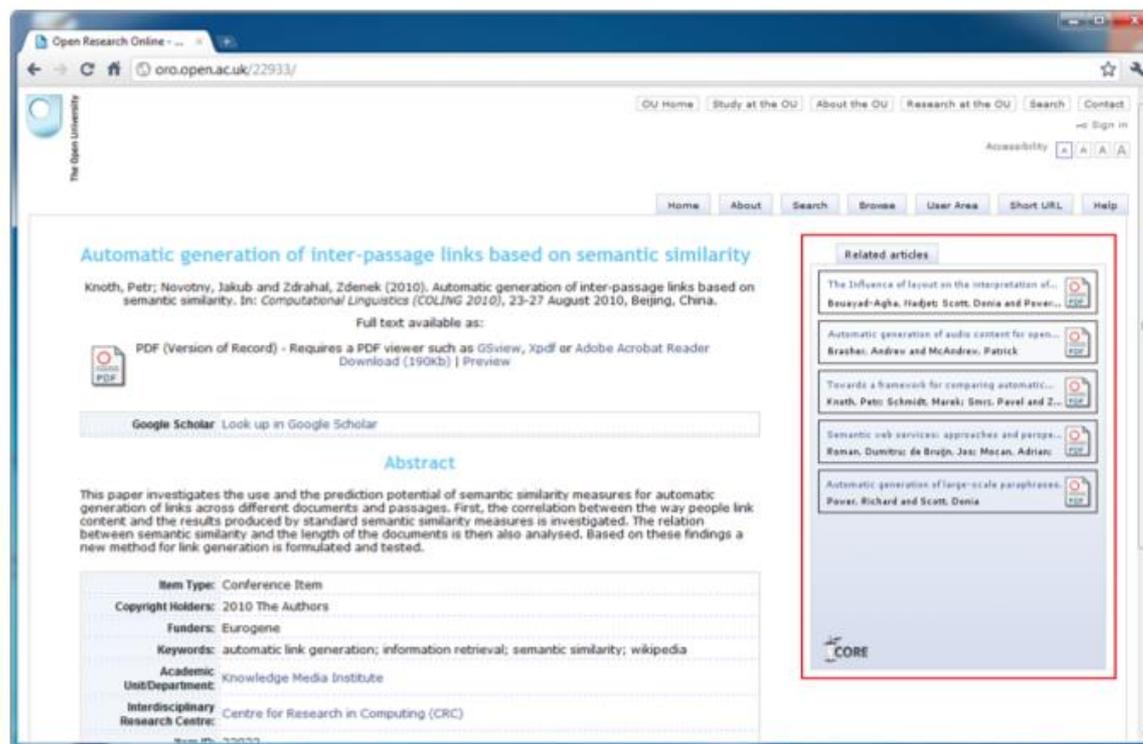
The original Scientific American article on the Semantic Web appeared in 2001. It described the evolution of a Web that consisted largely of documents for humans to read to one that included data and information for computers to manipulate. The Semantic Web is a Web of actionable information--information derived from data through a semantic theory for interpreting the symbols. This simple idea, however, remains largely unrealized. Shoppbots and auction bots abound on the Web, but these are essentially handcrafted for particular tasks; they have little ability to interact with heterogeneous data and information types. Because we haven't yet delivered large-scale, agent-based mediation, some commentators argue that the Semantic Web has failed to deliver. We argue that agents can only flourish when standards are well established and that the Web standards for expressing shared meaning have progressed steadily over the past five years. Furthermore, we see the use of ontologies in the e-science community presaging ultimate success for the Semantic Web--just as the use of HTTP within the GDM portal design community led to the realization of...

# CORE Applications



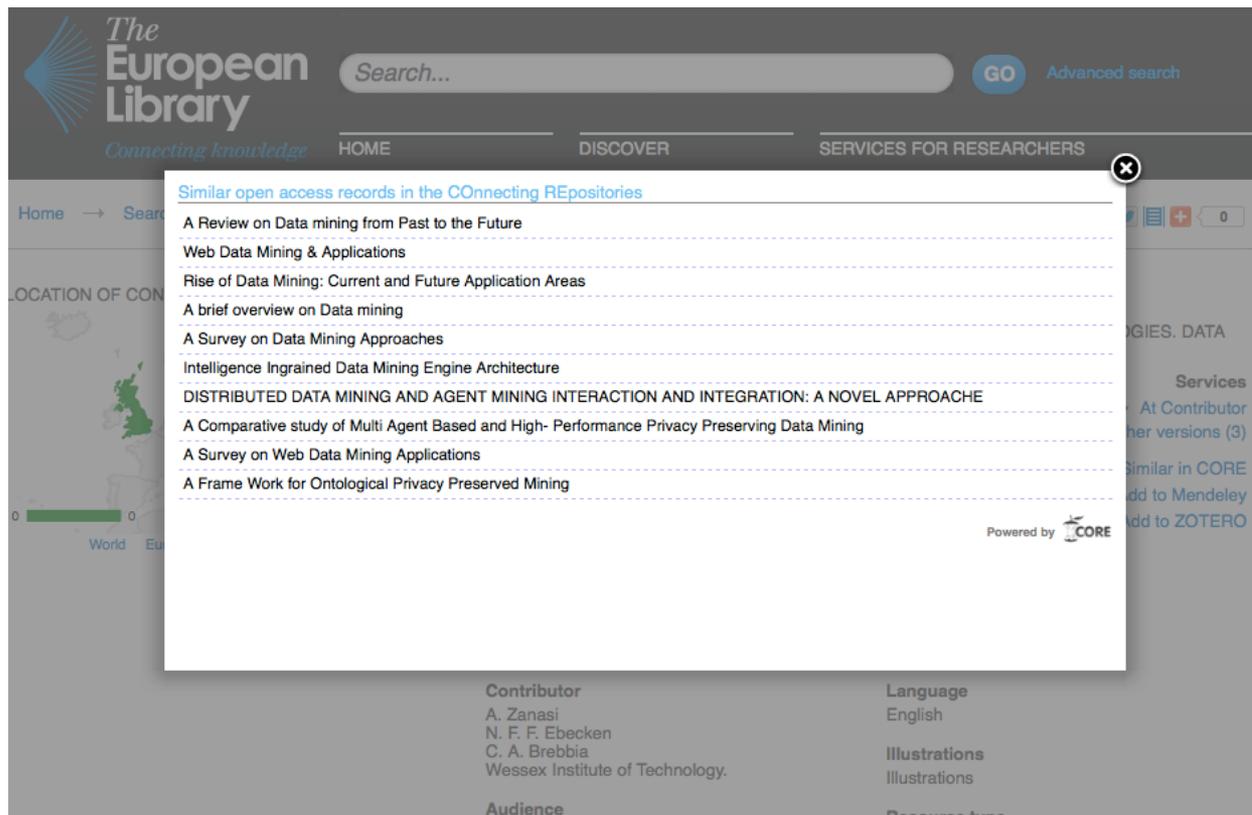
# CORE Applications

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



# CORE Applications

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



The screenshot shows the website interface for The European Library. At the top, there is a search bar with the text "Search..." and a "GO" button. Below the search bar, there are navigation tabs for "HOME", "DISCOVER", and "SERVICES FOR RESEARCHERS". A white overlay window is displayed in the center, titled "Similar open access records in the COncnecting REpositories". This window lists several related records:

- A Review on Data mining from Past to the Future
- Web Data Mining & Applications
- Rise of Data Mining: Current and Future Application Areas
- A brief overview on Data mining
- A Survey on Data Mining Approaches
- Intelligence Ingrained Data Mining Engine Architecture
- DISTRIBUTED DATA MINING AND AGENT MINING INTERACTION AND INTEGRATION: A NOVEL APPROACHE
- A Comparative study of Multi Agent Based and High- Performance Privacy Preserving Data Mining
- A Survey on Web Data Mining Applications
- A Frame Work for Ontological Privacy Preserved Mining

At the bottom right of the overlay, it says "Powered by CORE". Below the overlay, the main page content is partially visible, showing a "Contributor" section with names like A. Zanasi, N. F. F. Ebecken, and C. A. Brebbia, and a "Language" section set to "English".

# Built on top of CORE API ...

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

### New OJS Journal

---

[HOME](#)   [ABOUT](#)   [USER HOME](#)   [SEARCH](#)   [CURRENT](#)   [ARCHIVES](#)

Home > Vol 1, No 1 (2013) > **Joseph**

### Strategies Towards Open Access

*Heather Joseph*

#### Abstract

SPARC has been active in engaging in open access advocacy on the local institutional, federal and international policy levels. SPARC's strategy is focused on reducing barriers to access, sharing and use of scholarly information, and its highest priority is advancing the understanding and implementation of open access to research results. Heather Joseph provides an update on SPARC's recent advocacy activities as well as a snapshot of the current open access policy climate. Context of the Preconference: Members of the CLA Task Force on Open Access, the Scholarly Publishing and Academic Resources Coalition (SPARC), and the Canadian Association of Research Libraries (CARL) discussed the transition to Open Access and also offered a workshop. The presentations and workshop were given during the CLA Preconference on Open Access Meeting at the 1st International PKP Scholarly Publishing Conference in Vancouver, British-Columbia (Canada) from July 11-13, 2007.

Full Text:

[PDF](#)

#### Refbacs

There are currently no refbacks.

[OPEN JOURNAL SYSTEMS](#)

[Journal Help](#)

USER

You are logged in as...

**core**

- [My Profile](#)
- [Log Out](#)

NOTIFICATIONS

- [View \(11 new\)](#)
- [Manage](#)

JOURNAL CONTENT

Search

All ▼

Browse

- [By Issue](#)
- [By Author](#)
- [By Title](#)

FONT SIZE

A<sup>+</sup>
A
A<sup>-</sup>

INFORMATION

- [For Readers](#)
- [For Authors](#)
- [For Librarians](#)

#### Similar Articles

- Canadian Library Association : Task Force on Open Access
- Open access and evolving scholarly communication: An overview of library advocacy and commitment, institutional repositories, and publishing in Canada
- Open Access. Chapter 6 of Scholarly Communication for Librarians.
- Open Access and Canadian Libraries: Taking a Position
- Open Access Policy Update
- A Leading-Edge Position Statement on Open Access – Ongoing Interest in OA at CLA
- Summary and Conclusions. Final chapter of Scholarly Communication for Librarians.
- The role of open access in fostering knowledge sharing and collaboration in Ethiopia: a case study
- The Stratified Economics of Open Access
- Creators of the Commons

Powered by

# CORE Applications

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



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.

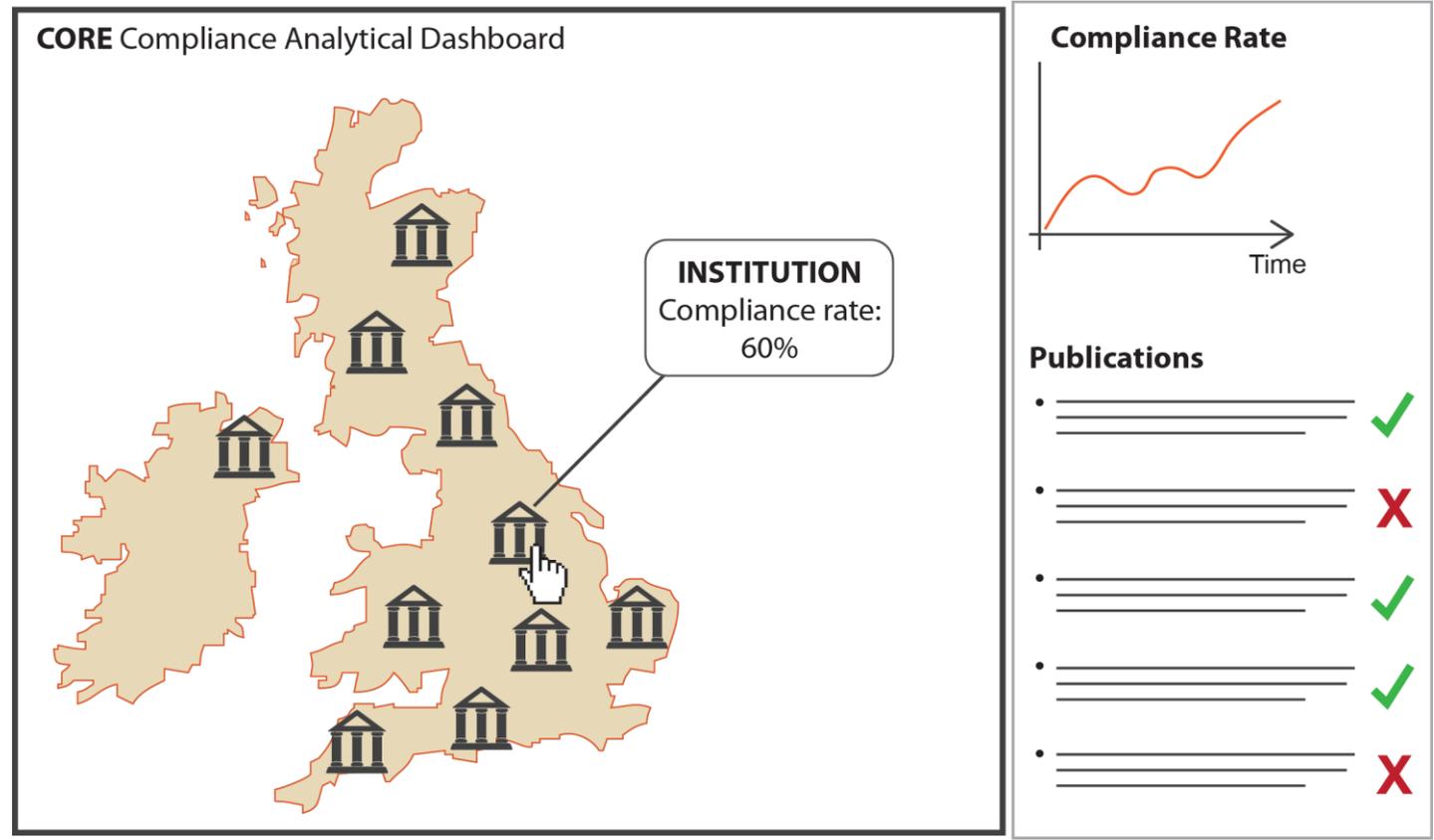
Consider that the current version of Repository Analytics is experimental and therefore the statistics might not be completely accurate. We are continuously working on improving this page and information to help provide a more precise picture of the Open Access repository ecosystem. Please do not hesitate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

[Overview](#)
[Repository Access](#)
[PDF Count](#)
[Total Document Count](#)
[Access and Document Count](#)

| FirstPrevious12345NextLast |   | Search: <input type="text"/> |                   |              |         |
|----------------------------|---|------------------------------|-------------------|--------------|---------|
| Nr. ^                      | Repository  | Metadata Download            | Metadata Readable | PDF Download | Overall |
| 1                          | <a href="#">Aberdeen University Research Archive</a>            | ✓ (1095)                     | ✓ (1118)          | ✓ (517)      | ★       |
| 2                          | <a href="#">Abertay Research Collections</a>                    | ✓ (1331)                     | ✓ (1331)          | ✓ (55)       | ★       |
| 3                          | <a href="#">Access to Research Resources for Teachers</a>       | ✓ (705)                      | ✓ (705)           | ✓ (316)      | ★       |
| 4                          | <a href="#">Advanced Knowledge Technologies EPrints Archive</a> | ✗                            | ✓ (341)           | ✓ (193)      |         |
| 5                          | <a href="#">ALT Open Access Repository</a>                      | ✓ (774)                      | ✓ (701)           | ✓ (610)      | ★       |
| 6                          | <a href="#">Anglia Ruskin Research Online</a>                   | ✓ (1322)                     | ✓ (1326)          | ✓ (143)      | ★       |

# CORE Applications

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



# 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



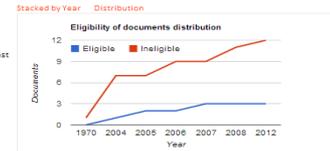
| Nr. | Institution         | Repositories                         | Compliance Rate |
|-----|---------------------|--------------------------------------|-----------------|
| 1   | Aberdeen University | Aberdeen University Research Archive | 20%             |

Showing 1 to 1 of 1 entries (filtered from 605 total entries)

## Aberdeen University



**Repositories**  
 Aberdeen University Research Archive  
 Uri: http://aura.abdn.ac.uk/dspace-09/request  
 Number of metadata records: 1146



Compliance rate  
**20%**

### Advanced filter

| Nr. | Title  | Author   | Year | Eligible |
|-----|--|--|------|----------|
| 1   | Do Social Preferences Increase Productivity? Field experimental evidence from fishermen in Toyama Bay  | Erika Seki and Jeffrey Carpenter   | 2004 | ✗        |
| 2   | Acquisition of Living Things by Specification  | Ernest Metzger   | 2004 | ✗        |
| 3   | Effects of rotation scheme on fishing behaviour with price discrimination and limited durability: Theory and evidence.   | Erika Seki   | 2004 | ✗        |
| 4   | Ethnic enclaves and employment in England and Wales  | Harinder Battu and Macdonald Mwale   | 2004 | ✗        |
| 5   | Job contact networks and the ethnic minorities   | Harinder Battu, Paul Seaman and Yves Zenou   | 2004 | ✗        |
| 6   | Who are the moonlighters and why they moonlight: Evidence for rural communities  | Heather Dickey and Ioannis Theodossiou   | 2004 | ✓        |
| 7   | Corporate culpable homicide: Transcopic v HM Advocate  | James Chalmers   | 2004 | ✗        |
| 8   | Scandinavian Realism   | Geoffrey MacCormack  | 1970 | ✗        |
| 9   | Wagenot, Eigentumsgarantie und Differenzialer Interesse. Der Norweg in Deutschland und Südafrika   | Cornelius Marwe and Michael Hindan   | 2005 | ✓        |
| 10  | Community Hospitals - the place of local service provision in a modernising NHS: an integrative thematic literature review   | David Heaney, Corri Black, Catherine Donnell, Cameron Stark and Edwin Teitlingen   | 2006 | ✗        |
| 11  | Do self-reported intentions predict clinicians behaviour: a systematic review.   | Martin Eccles, Susan Hrisoa, Jillian Francis, Eileen Kaner, DO Dickinson, F Beyer and Marie Johnston   | 2006 | ✗        |
| 12  | Developing the content of two behavioural interventions: using theory-based interventions to promote GP management of upper respiratory tract infection without prescribing antibiotics #1 | Susan Hrisoa, Martin Eccles, Merle Johnston, Jillian Francis, Eileen Kaner, Nick Steen and Jeremy Grimshaw   | 2008 | ✗        |
| 13  | Hydrocortisone therapy for patients with septic shock  | Charles Sprung, Djillali Annana, Didier Kab, Rui Moreno, Marvyn Singer, Klaus Frelvogel, Yoram Weiss, Julie Benbarashy, Armin Kalenka, Halmuth Forst, Pierre Latereke, Konrad Rainhart, Brian Cuthbertson, Didier Pagan, Josef Briegleb and CORTICUS Study Group | 2008 | ✗        |
| 14  | Induced gamma band responses predict recognition delays during object identification   | Jasna Martinovic, Thomas Gruber and Matthias Mueller   | 2007 | ✓        |
| 15  | Source attribution, prevalence and enumeration of Campylobacter spp. from retail liver   | N Strachan, M Macrae, A Thomson, O Rotariu, I Ogdan and K Forbes   | 2012 | ✗        |

## 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), Implementation 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.