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

June 21, 2019 – OAI-11, Geneva

Petr Knoth

Big Scientific Data and Text Analytics Group

Knowledge Media Institute, The Open University

Papers OA sooner and sooner

MENU SIGN IN/R

PHYSICS TODAY

HOME BROWSE INFO RESOURCES JOBS

DOI:10.1063/PT.6.2.20190418a

18 Apr 2019 in Politics & Policy

Study quantifies the growing traction of open access

Spurred by the policies of funding agencies, academics are getting much better at posting their peer-reviewed research in freely available online repositories.

Dalmeet Singh Chawla

1 COMMENTS 203 SHARES

Credit: h_pampel, CC BY-SA 2.0

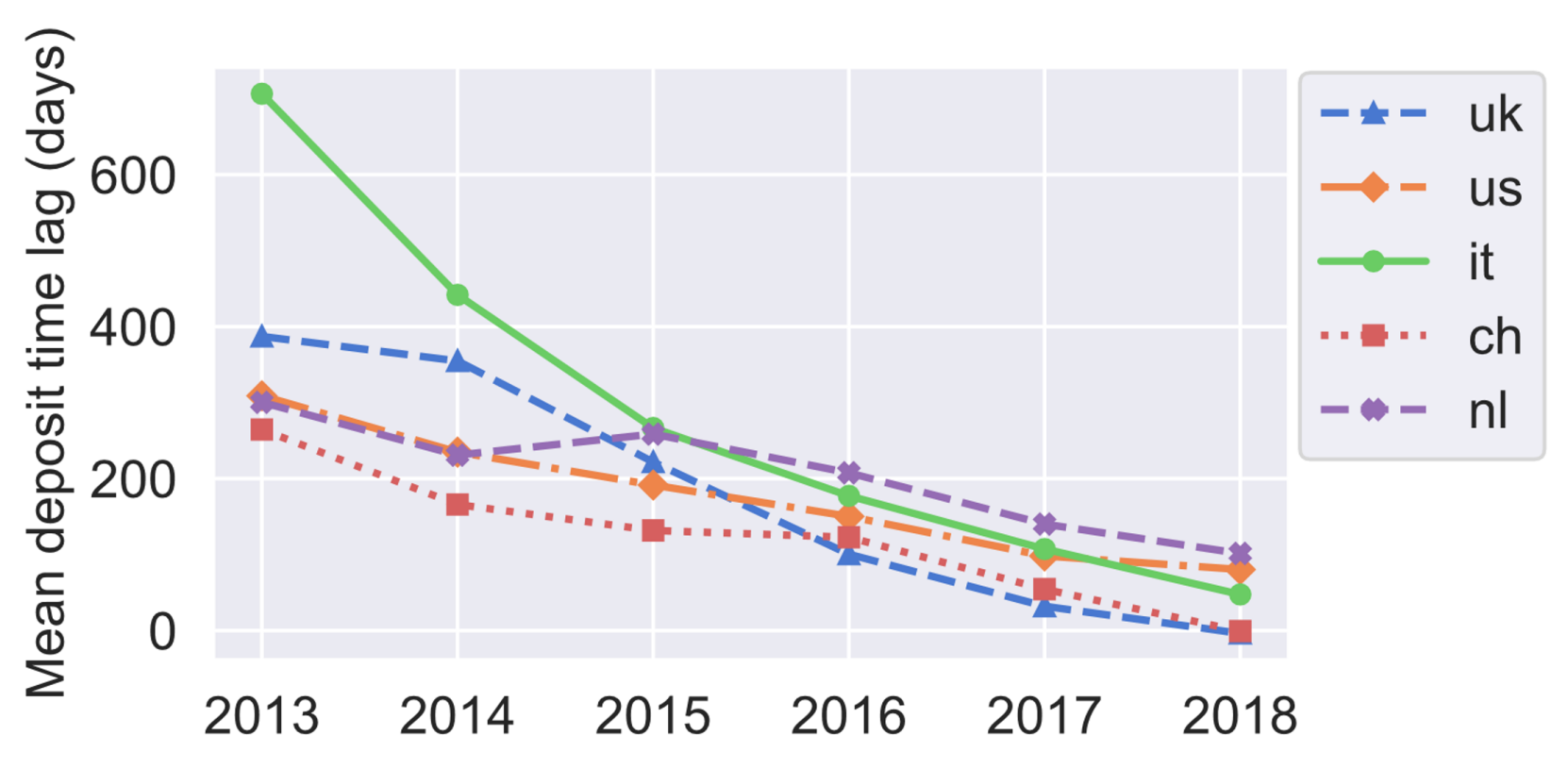
Five years ago, the UK funding body Research England, then known as the Higher Education Funding Council for England, announced an ambitious policy designed to speed up the transition to open-access publishing. To become eligible for a slice of billions of pounds of government money distributed to UK universities, academics would have to post their research on free-to-access websites such as preprint servers and institutional repositories

JCDL 2019 Best Paper Award

https://physicstoday.scitation.org/do/10.1063/PT.6.2.20190418a/full/

Papers OA sooner and sooner

The delay between publication and OA availability decreasing globally



Herrmannova, Drahomira; Pontika, Nancy and Knoth, Petr (2019). Do Authors Deposit on Time? Tracking Open Access Policy Compliance. In: *2019 ACM/IEEE Joint Conference on Digital Libraries*, 2-6 Jun 2019, Urbana-Champaign, IL

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

Global network of repositories

"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)*

OA Aggregations and BOAI 2002

“To achieve open access to scholarly journal literature, we recommend two complementary strategies.

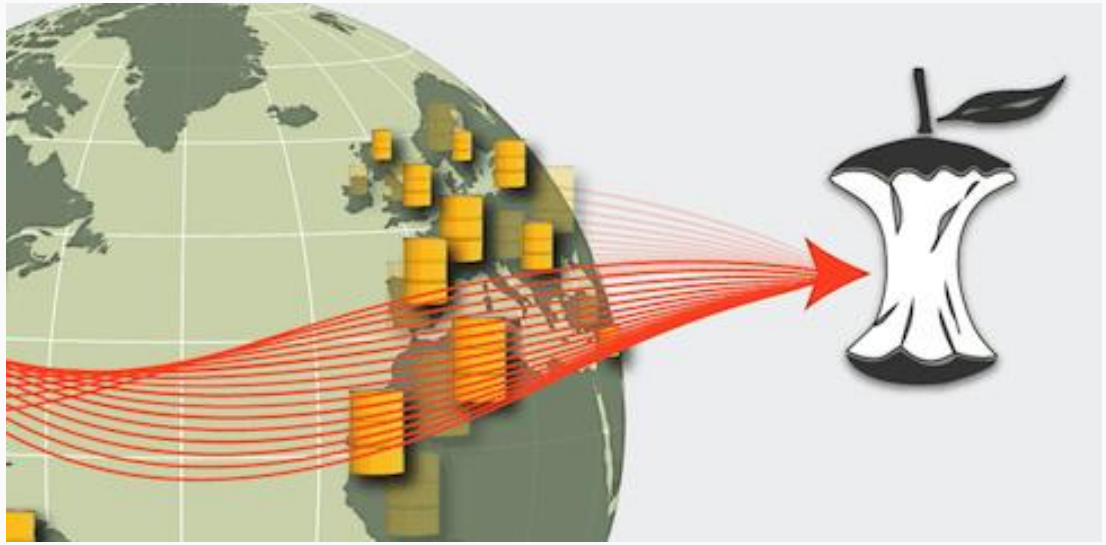
- Self-Archiving: First, scholars need the tools and assistance to deposit their refereed journal articles in open electronic archives, a practice commonly called, self-archiving. When these archives conform to standards created by the Open Archives Initiative, then search engines and other tools can 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

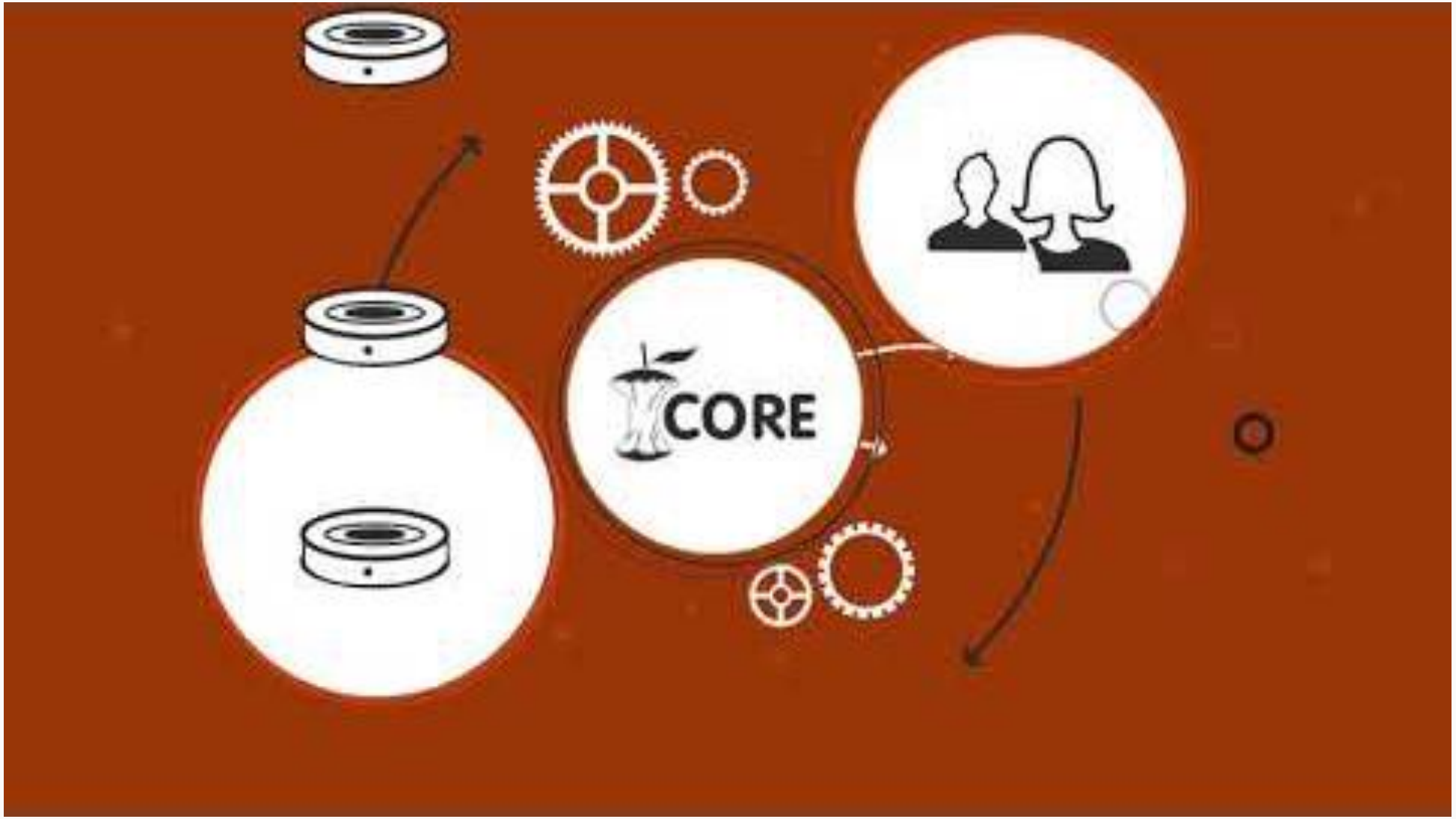
CORE's mission

Aggregate all open access research articles worldwide …

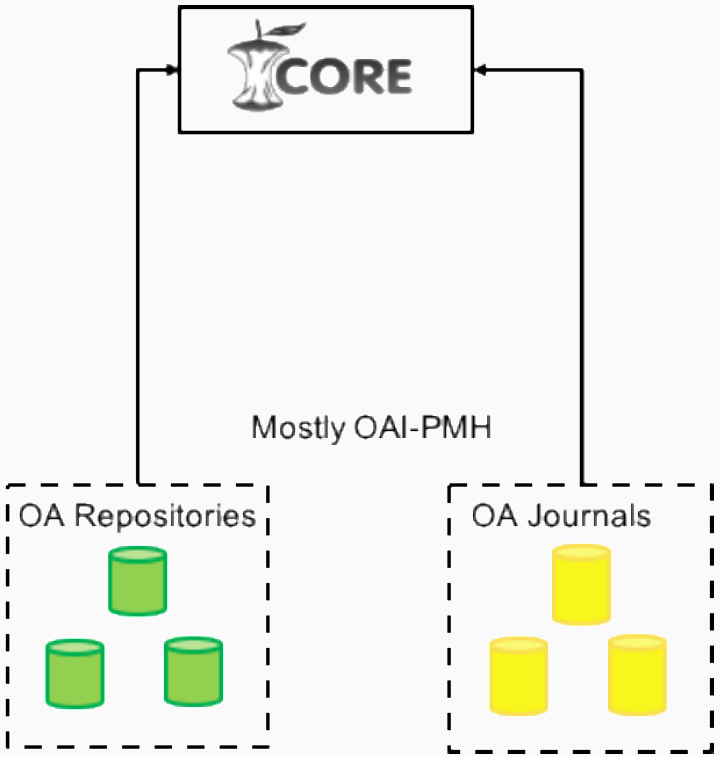
… enrich this content and provide seamless access to it through a set of data services …



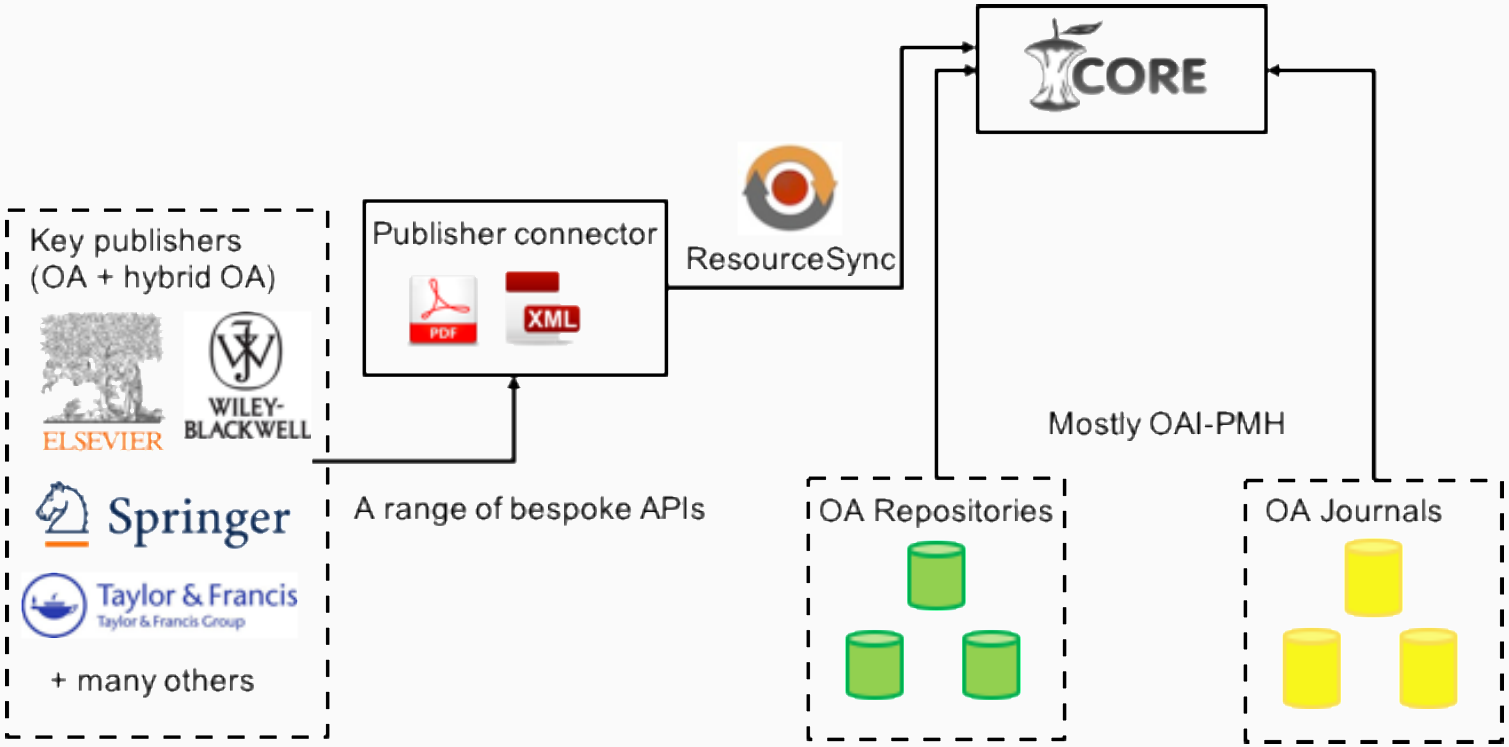
Introducing CORE



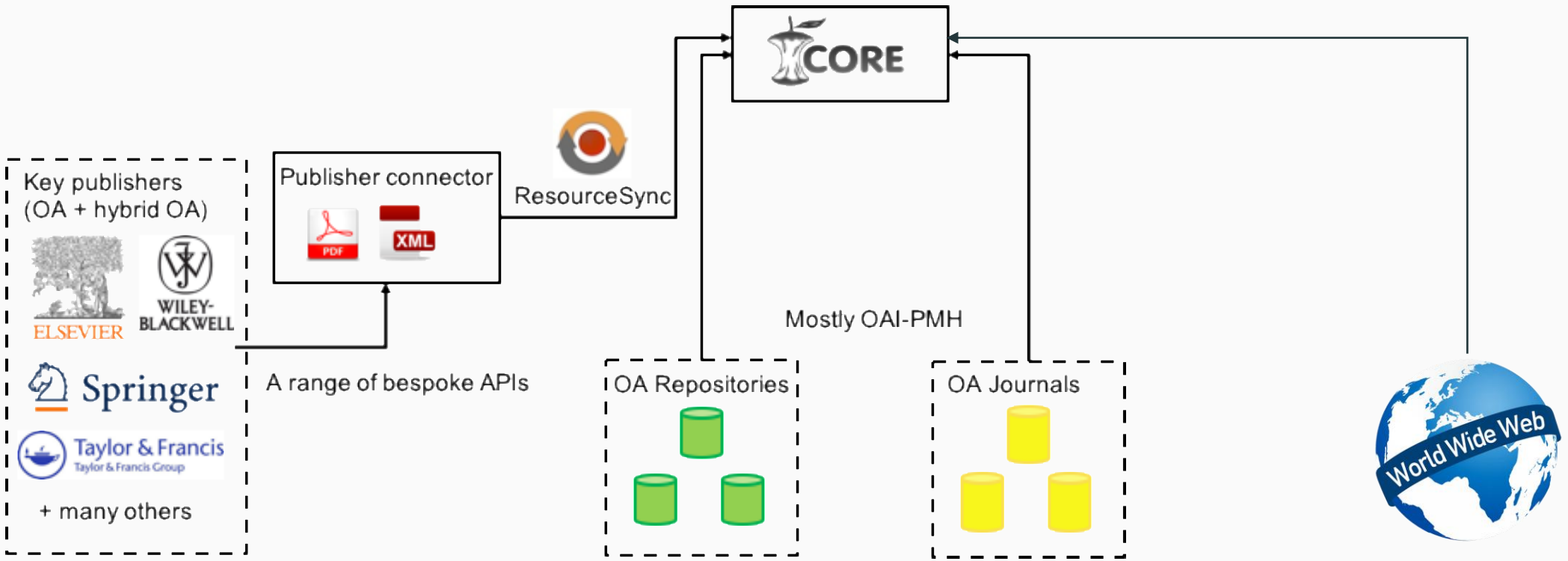
Harvesting data is challenging



Harvesting data is challenging



Harvesting data is challenging



Need for more interoperability across systems

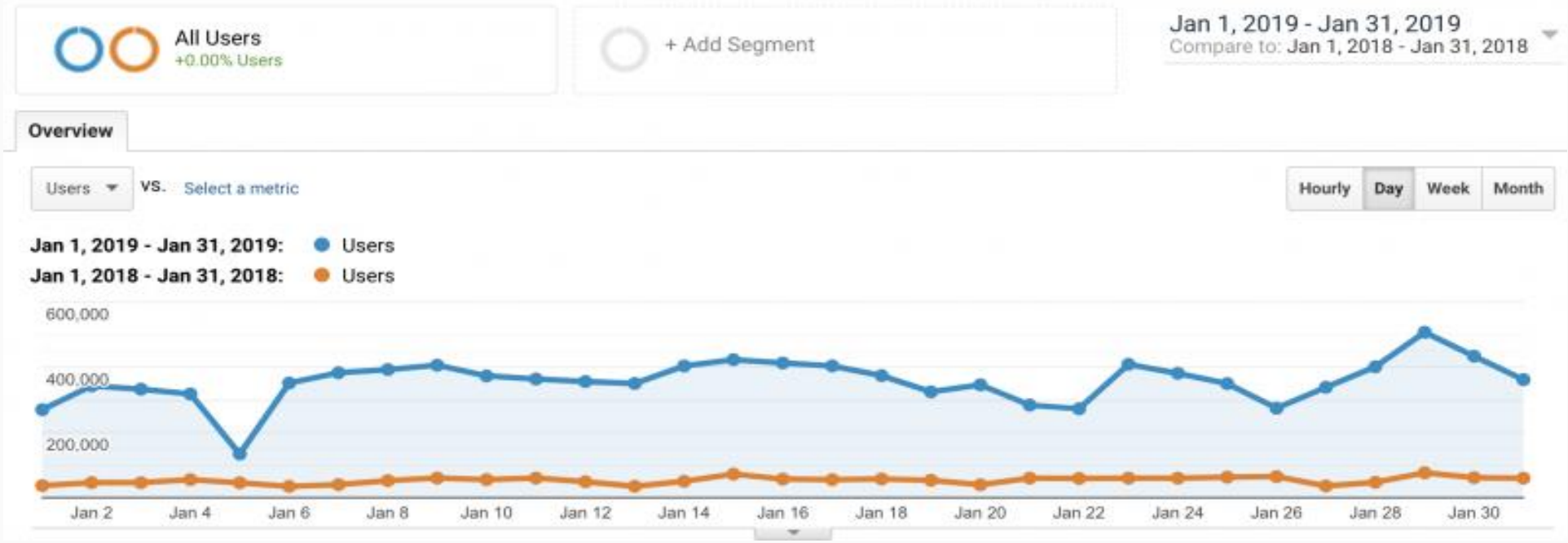
World’s largest dataset of Open Access full texts

- **13,117,488** Hosted full texts
- **135,539,113** Metadata records
- **~78m** Links to full texts
- **15TB** of raw plain text
- **4,123** Data providers

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)

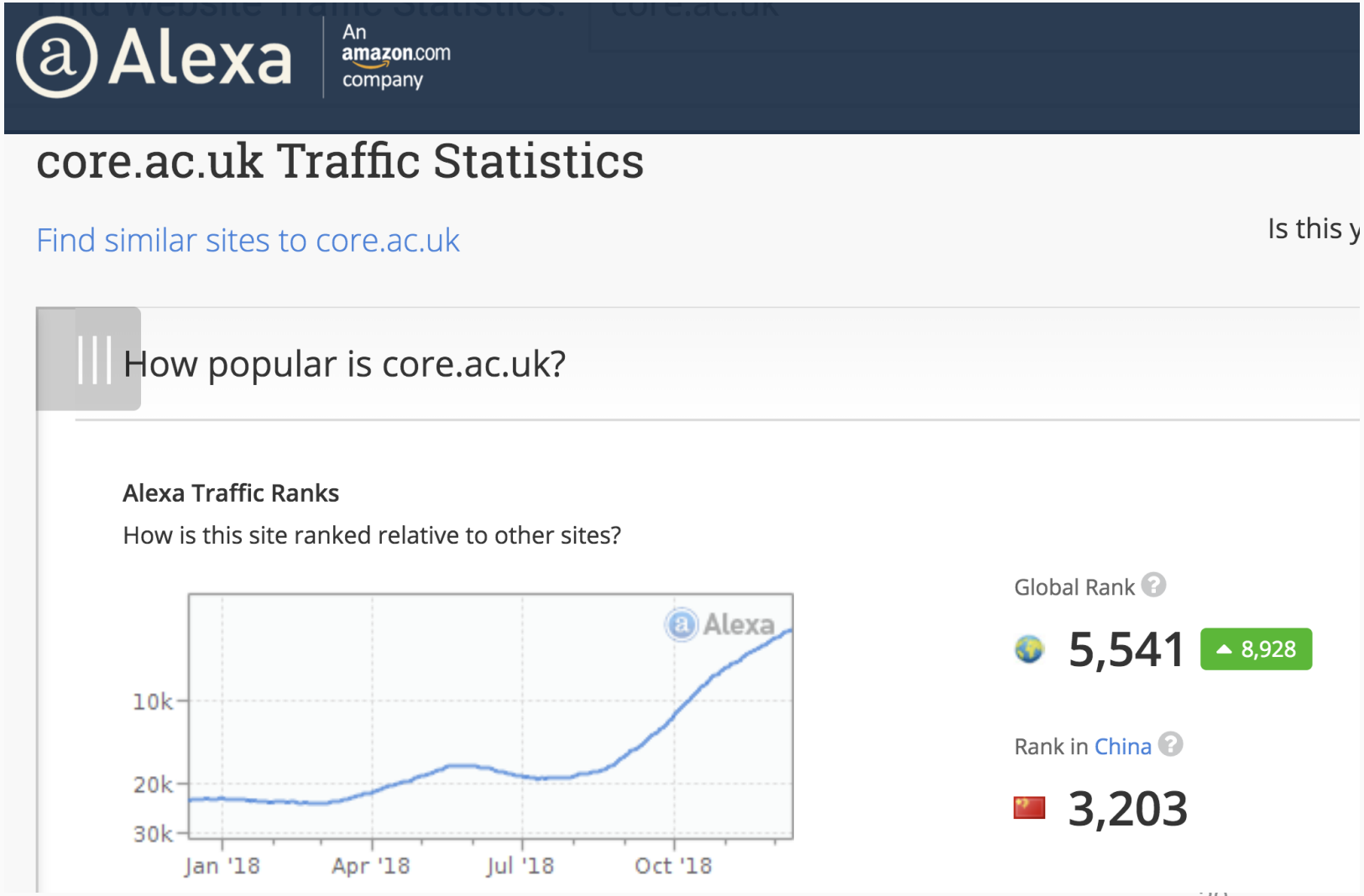
CORE usage



- January 2019 – CORE reached **over 10M monthly active users** for the first time
- 571% increase from January 2018

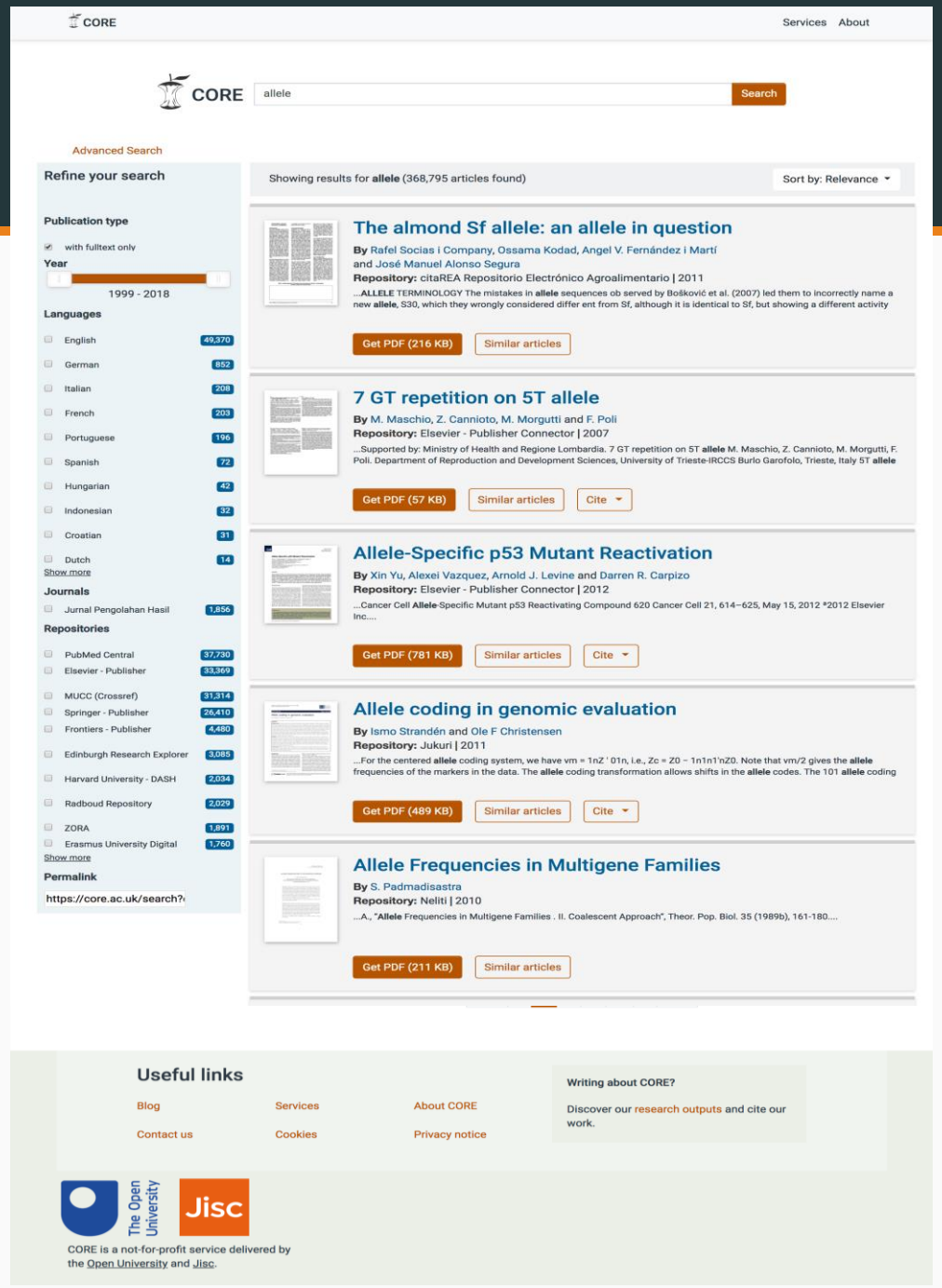
Alexa rank

- Within top 6k global websites: https://www.alexa.com/
- core.ac.uk by usage in the **top 0.0009% of global websites**

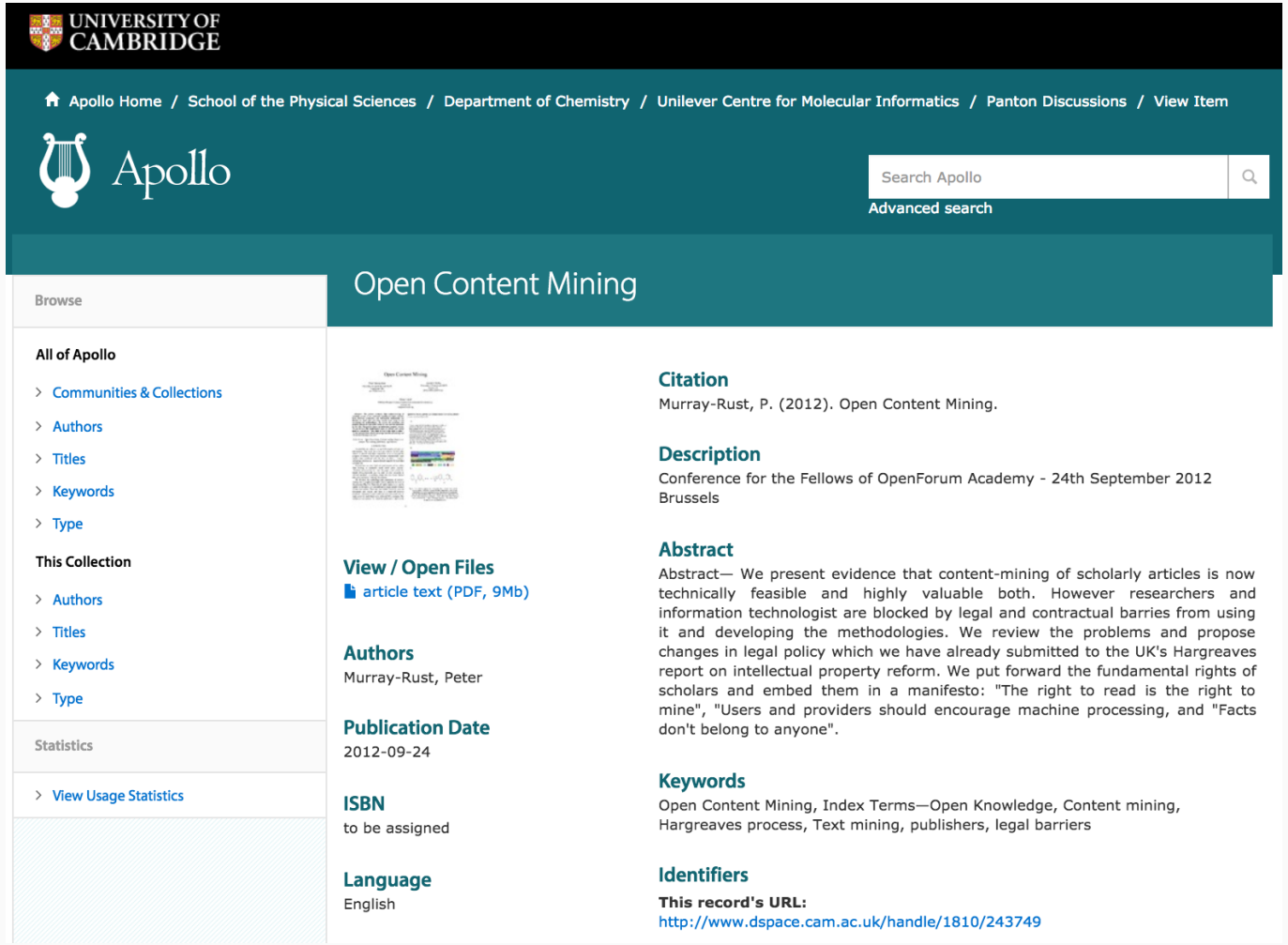


CORE Search

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

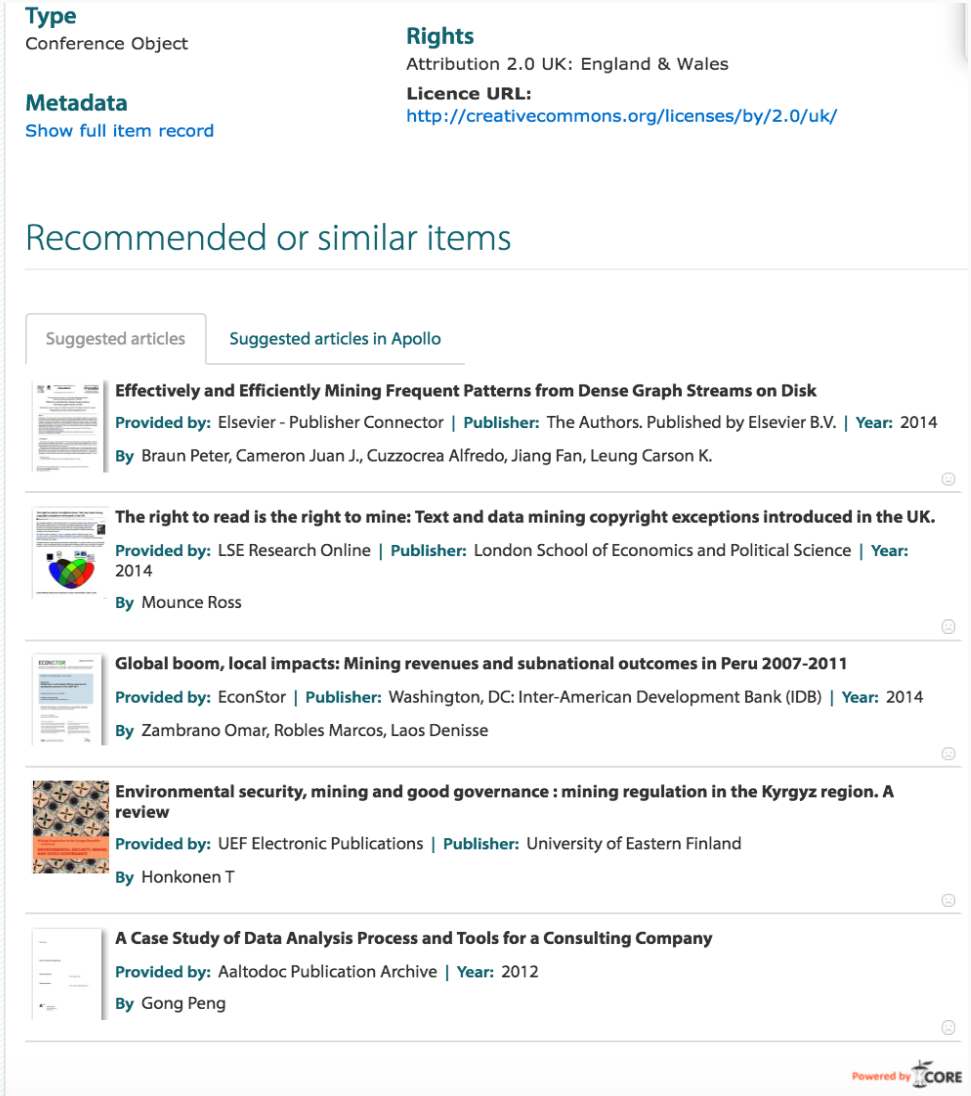


CORE Recommender



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

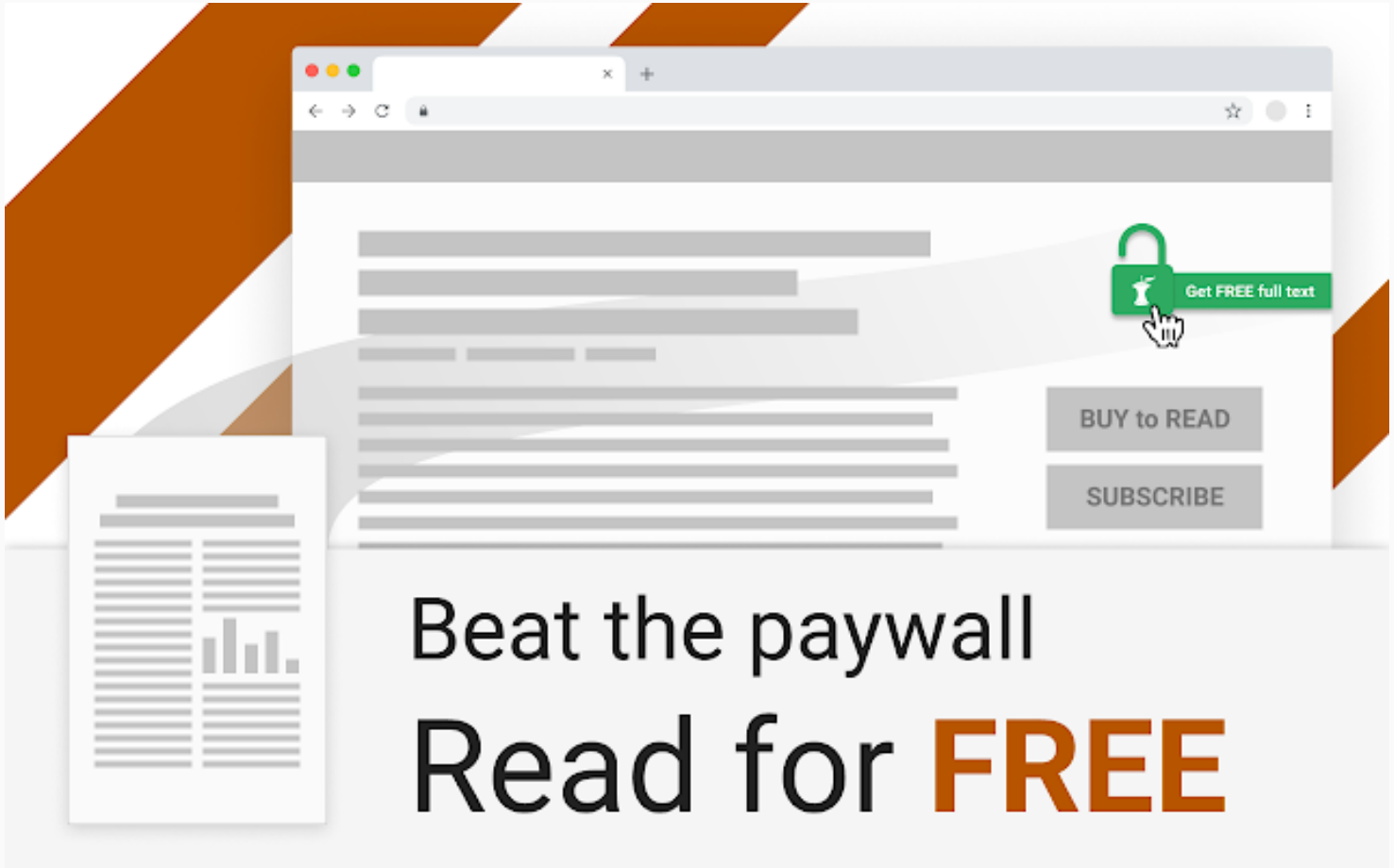
CORE Recommender



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

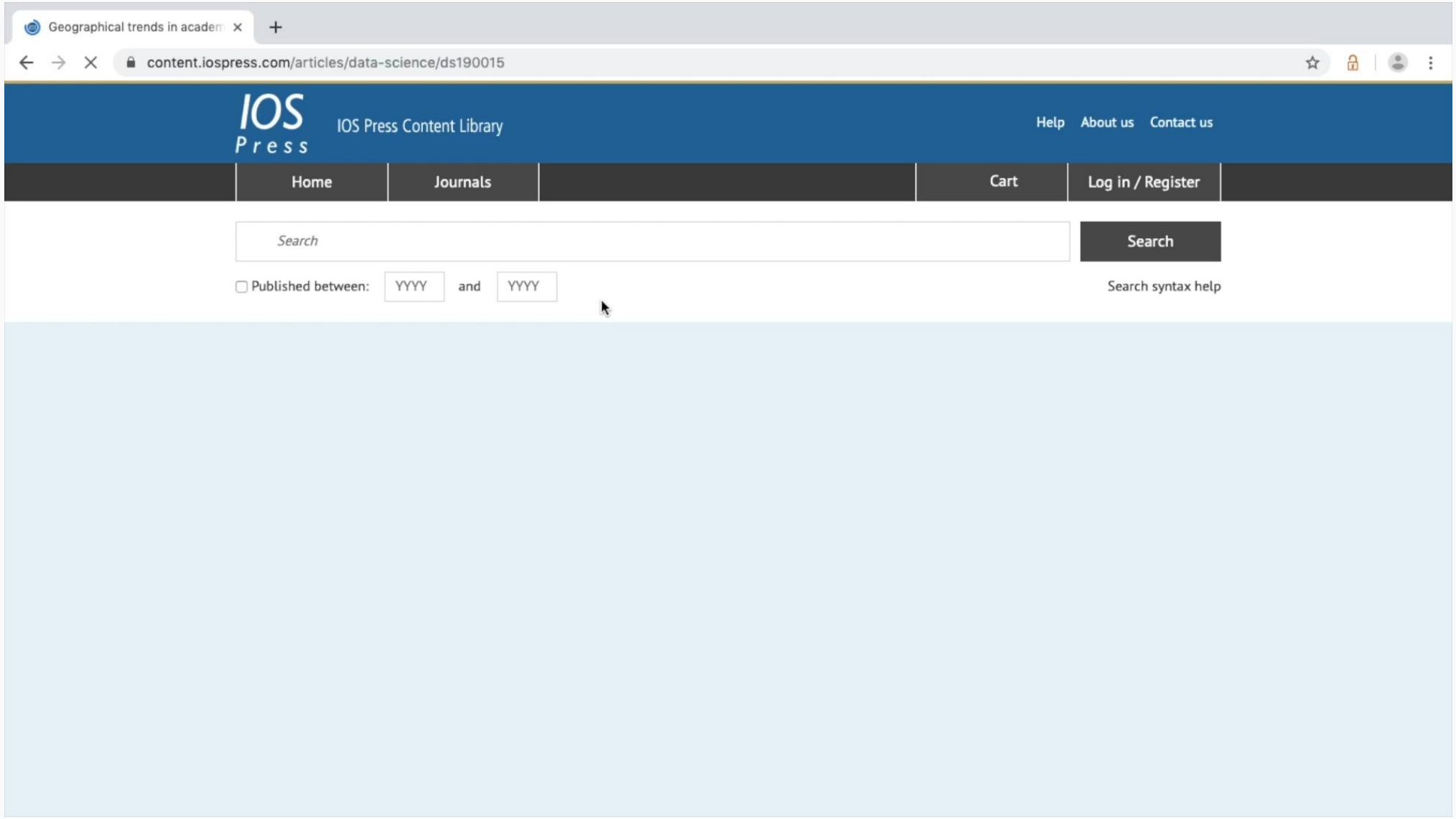
Introducing CORE Discovery

- 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.



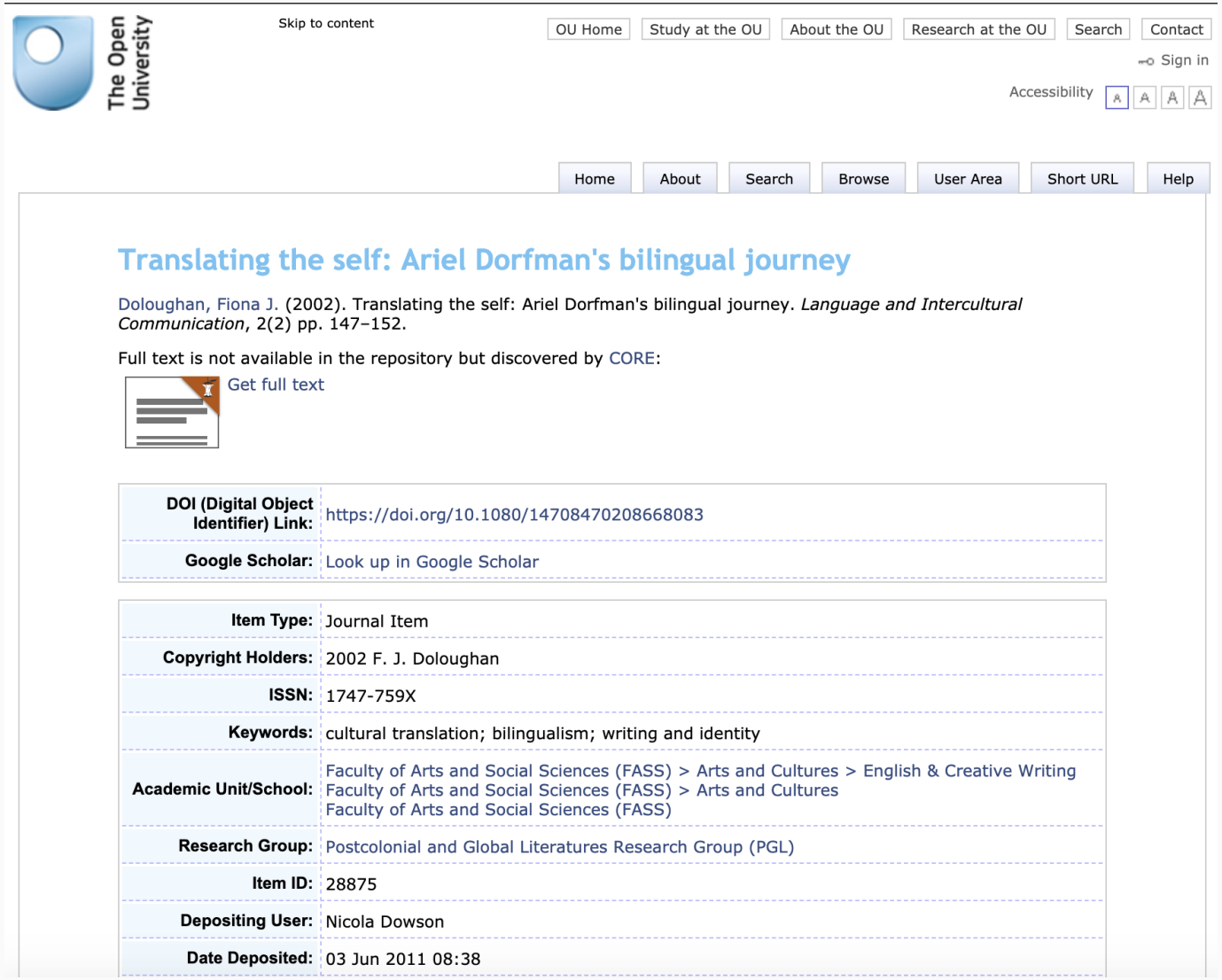
https://core.ac.uk/services/discovery/

CORE Discovery demonstration



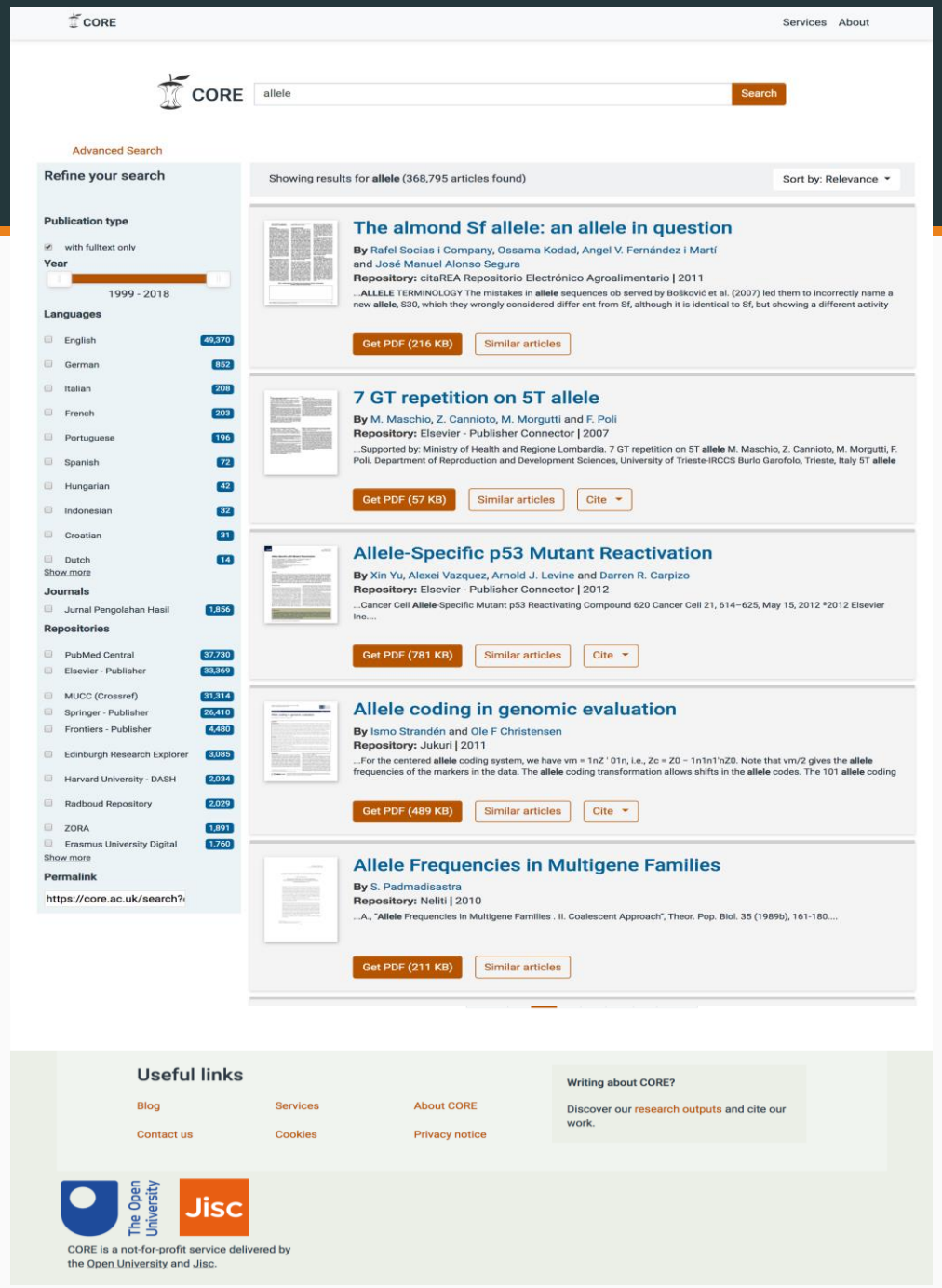
CORE Discovery Repository integration

- 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.



CORE Search

- Full text search for OA content
- Faceted searching
- What you find is what you get



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
  - …

Working with partners

NAVER

IRIS.AI

ontochem
IT SOLUTIONS

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 courpus

Thank you!

https://core.ac.uk