

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

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

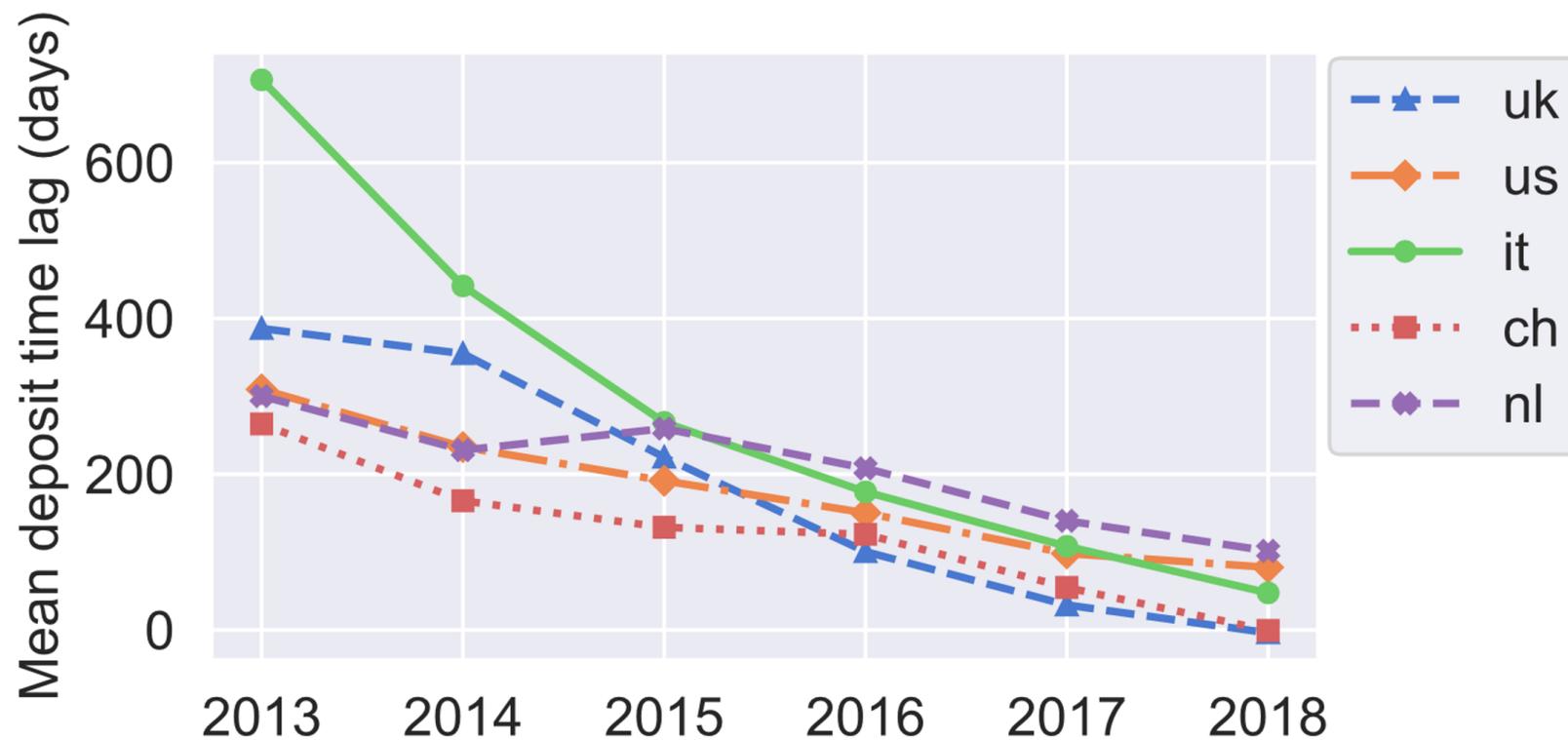


JCDL 2019 Best Paper Award

<https://physicstoday.scitation.org/doi/10.1063/PT.6.2.20190418a/full/>

Papers OA sooner and sooner

The delay between publication and OA availability decreasing globally



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.

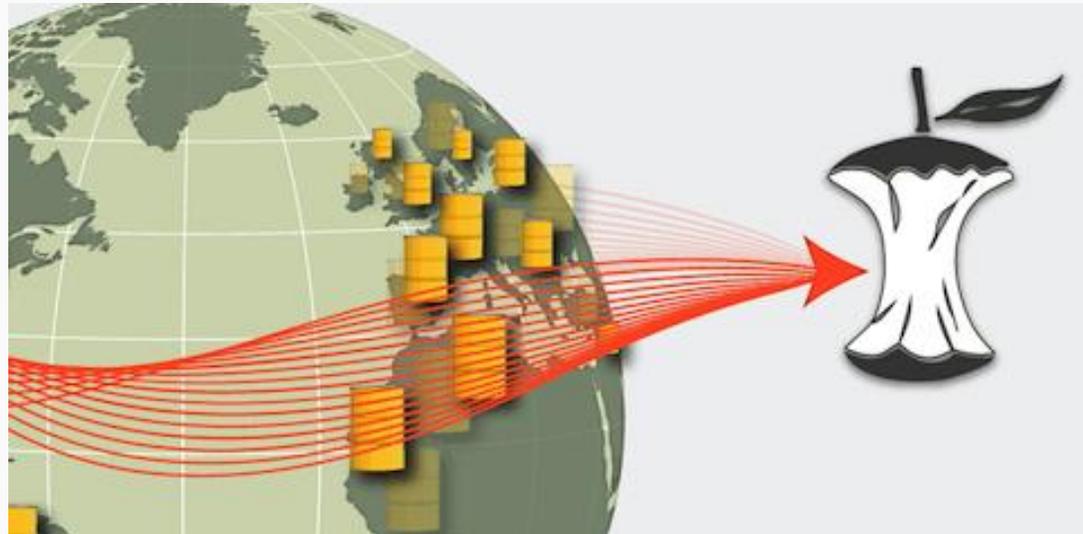
- Self-Archiving: First, scholars need the tools and assistance to deposit their refereed journal articles in open electronic archives, a practice commonly called, self-archiving. When these archives conform to standards created by the Open Archives Initiative, then search engines and other tools can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

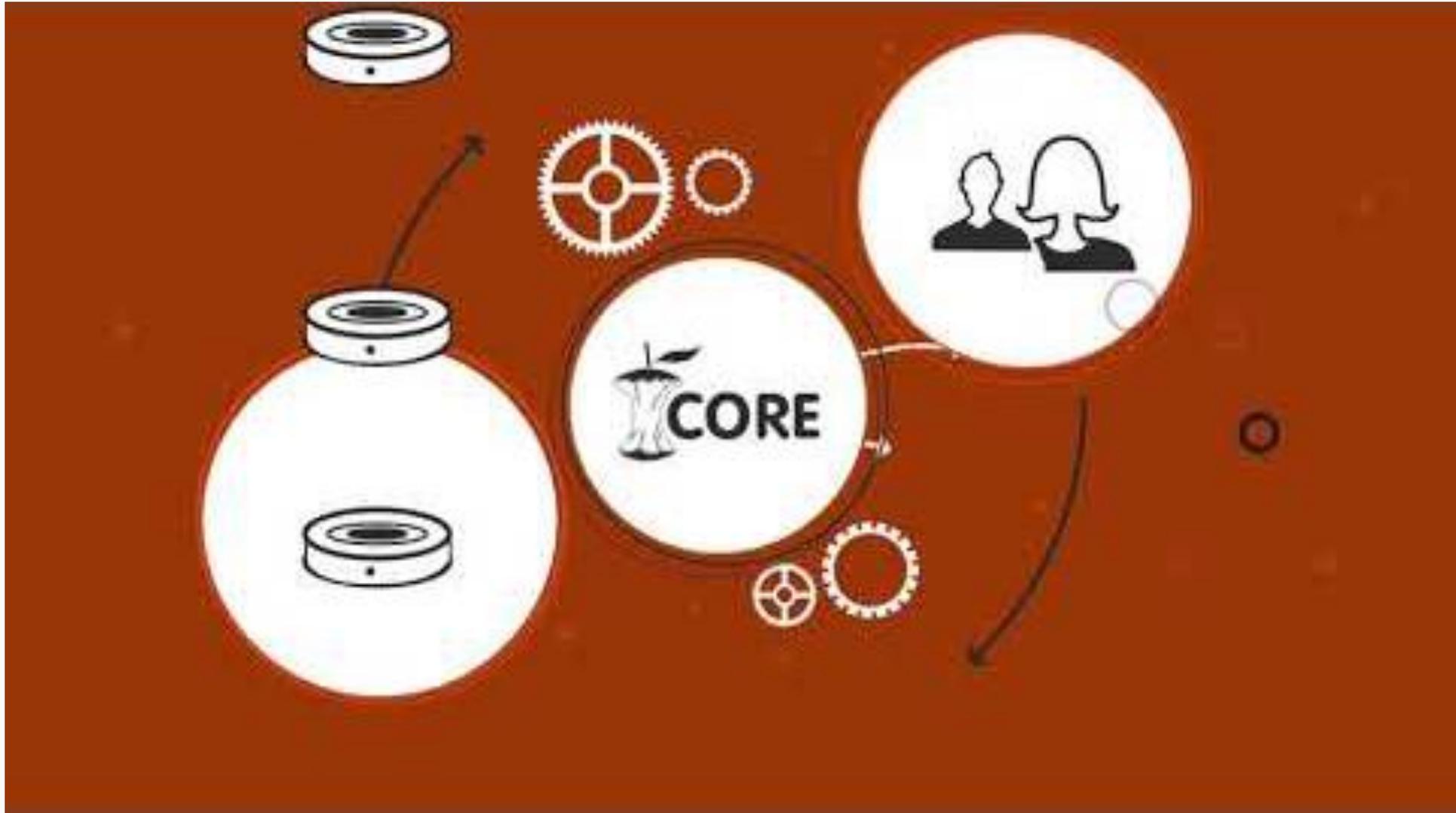
CORE's mission

Aggregate all open access research articles worldwide ...

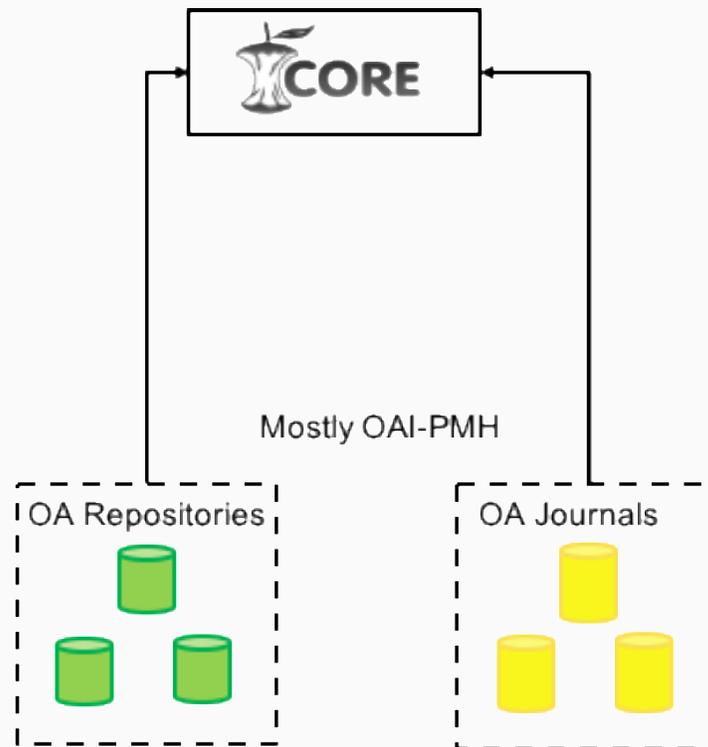
... enrich this content and provide **seamless access** to it through a set of **data services** ...



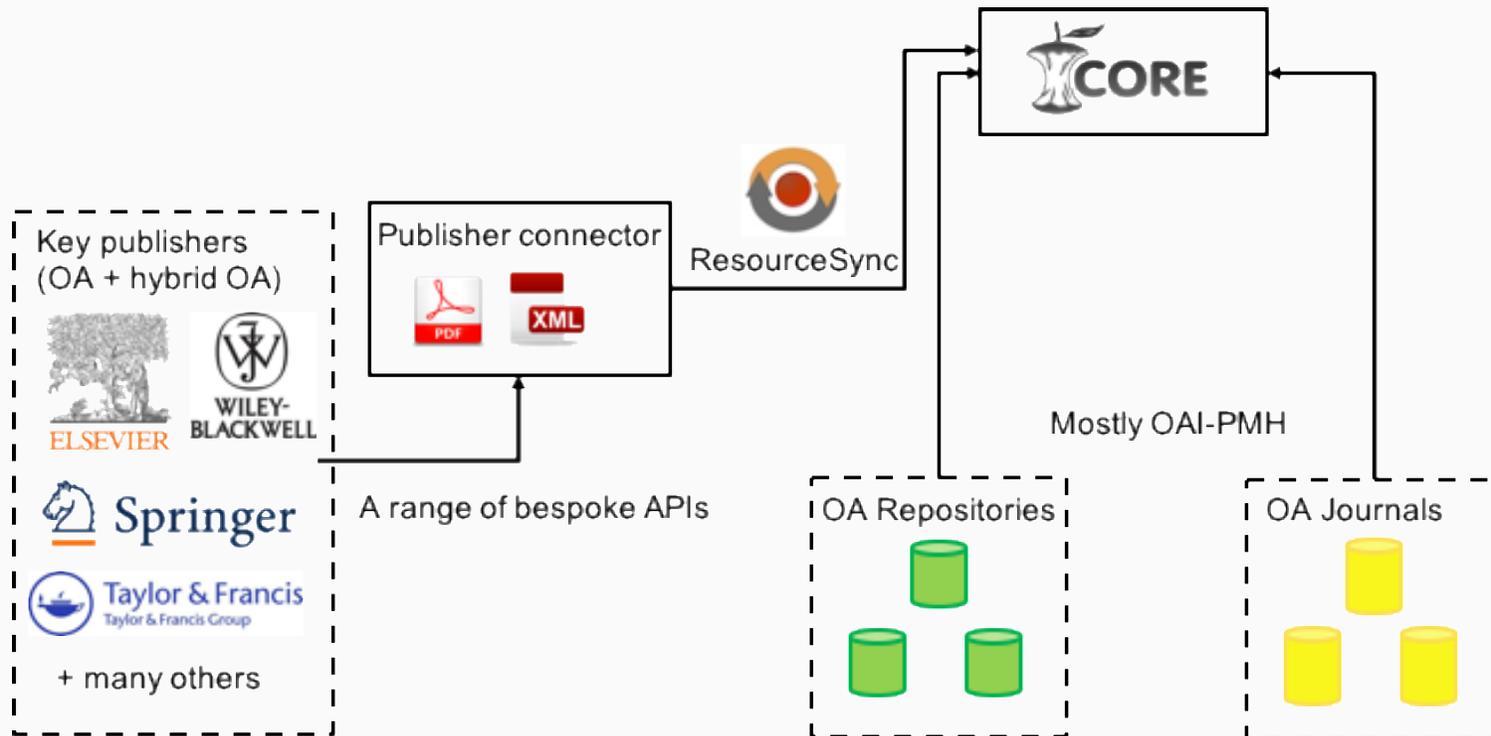
Introducing CORE



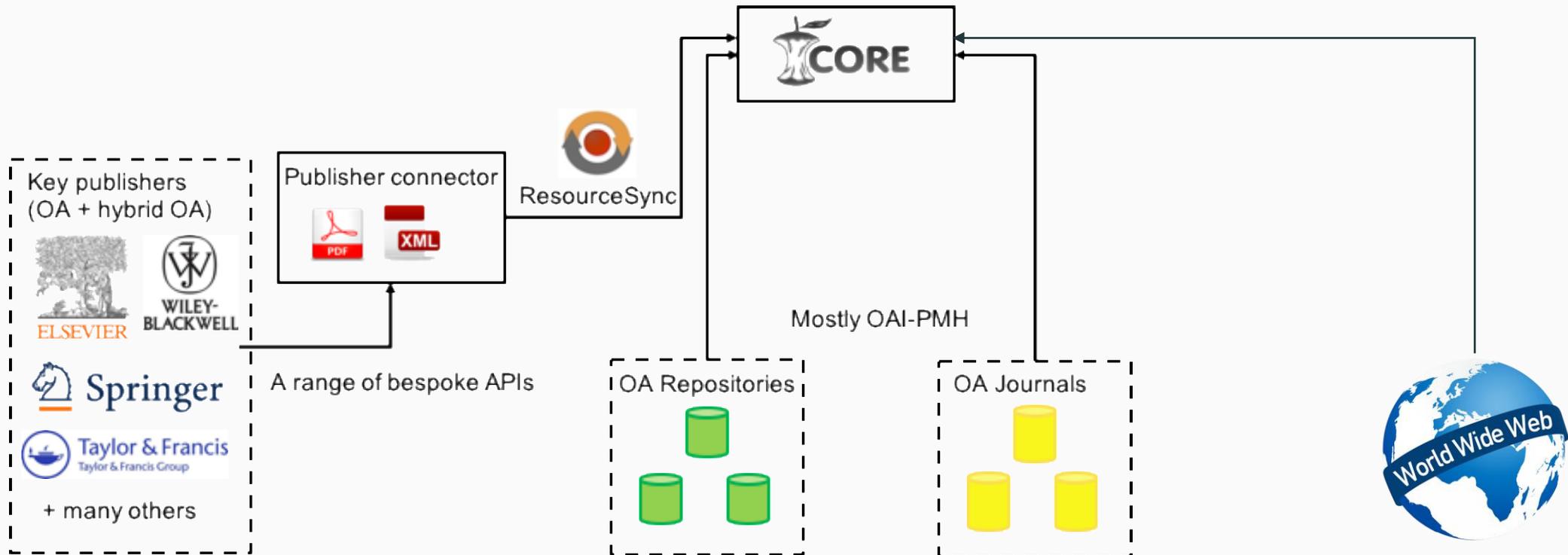
Harvesting data is challenging



Harvesting data is challenging



Harvesting data is challenging



Need for more interoperability across systems

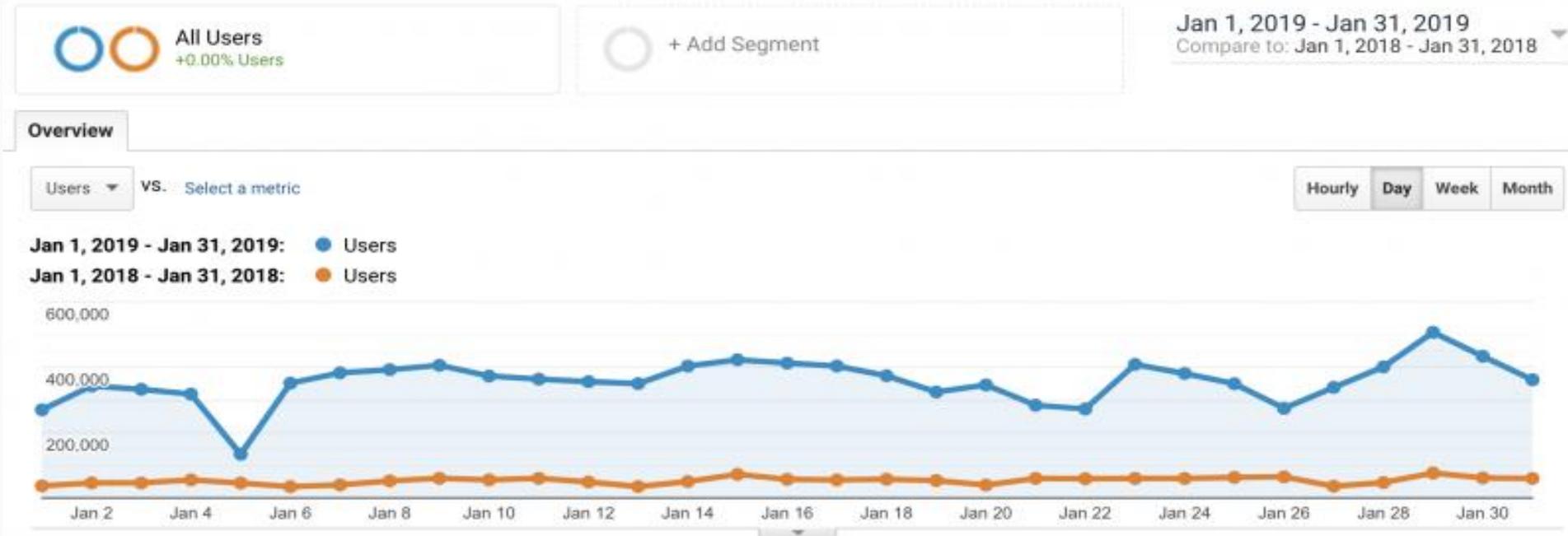
World's largest dataset of Open Access full texts

- **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.alexacom/>
- core.ac.uk by usage in the **top 0.0009% of global websites**



core.ac.uk Traffic Statistics

[Find similar sites to core.ac.uk](#)

Is this y

How popular is core.ac.uk?

Alexa Traffic Ranks

How is this site ranked relative to other sites?



Global Rank ?

5,541 ▲ 8,928

Rank in [China](#) ?

3,203

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

The screenshot displays the CORE Search website interface. At the top, the CORE logo is visible on the left, and 'Services About' links are on the right. A search bar contains the term 'allele' and a 'Search' button. Below the search bar, the page is divided into a left sidebar for refining search and a main content area for search results.

Refine your search

- Publication type**
 - with fulltext only
- Year**
 - 1999 - 2018
- Languages**
 - English 49,370
 - German 852
 - Italian 208
 - French 203
 - Portuguese 196
 - Spanish 72
 - Hungarian 42
 - Indonesian 32
 - Croatian 31
 - Dutch 14
- Journals**
 - Jurnal Pengolahan Hasil 1,856
- Repositories**
 - PubMed Central 37,730
 - Elsevier - Publisher 33,369
 - MUCC (Crossref) 31,314
 - Springer - Publisher 28,410
 - Frontiers - Publisher 4,480
 - Edinburgh Research Explorer 3,085
 - Harvard University - DASH 2,034
 - Radboud Repository 2,029
 - ZORA 1,891
 - Erasmus University Digital 1,760
- Permalink**
 - <https://core.ac.uk/search/>

Showing results for allele (368,795 articles found) Sort by: Relevance ▾

The almond Sf allele: an allele in question
By Rafel Socias | Company, Osama Kodad, Angel V. Fernández | Martí and José Manuel Alonso Segura
Repository: citaREA Repositorio Electrónico Agroalimentario | 2011
...ALLELE TERMINOLOGY The mistakes in allele sequences observed by Bošković et al. (2007) led them to incorrectly name a new allele, S30, which they wrongly considered different from Sf, although it is identical to Sf, but showing a different activity
Get PDF (216 KB) Similar articles

7 GT repetition on 5T allele
By M. Maschio, Z. Cannioto, M. Morgutti and F. Poli
Repository: Elsevier - Publisher Connector | 2007
...Supported by: Ministry of Health and Regione Lombardia. 7 GT repetition on 5T allele M. Maschio, Z. Cannioto, M. Morgutti, F. Poli. Department of Reproduction and Development Sciences, University of Trieste-IRCCS Burlo Garofolo, Trieste, Italy 5T allele
Get PDF (57 KB) Similar articles Cite ▾

Allele-Specific p53 Mutant Reactivation
By Xin Yu, Alexei Vazquez, Arnold J. Levine and Darren R. Carpizo
Repository: Elsevier - Publisher Connector | 2012
...Cancer Cell Allele-Specific Mutant p53 Reactivating Compound 620 Cancer Cell 21, 614-625, May 15, 2012 *2012 Elsevier Inc...
Get PDF (781 KB) Similar articles Cite ▾

Allele coding in genomic evaluation
By Ismo Strandén and Ole F Christensen
Repository: Jukuri | 2011
...For the centered allele coding system, we have $vm = 1nZ' 01n$, i.e., $Z_0 = Z_0 - 1n1n1'n2_0$. Note that $vm/2$ gives the allele frequencies of the markers in the data. The allele coding transformation allows shifts in the allele codes. The 101 allele coding
Get PDF (489 KB) Similar articles Cite ▾

Allele Frequencies in Multigene Families
By S. Padmadisastra
Repository: Neliti | 2010
...A., "Allele Frequencies in Multigene Families. II. Coalescent Approach", Theor. Pop. Biol. 35 (1989b), 161-180...
Get PDF (211 KB) Similar articles

Useful links

- Blog
- Contact us
- Services
- Cookies
- About CORE
- Privacy notice

Writing about CORE?
Discover our research outputs and cite our work.

The Open University **Jisc**

CORE is a not-for-profit service delivered by the Open University and Jisc.

CORE Recommender

The screenshot displays the Apollo repository interface. At the top, the University of Cambridge logo and name are visible. Below this is a navigation breadcrumb: 'Apollo Home / School of the Physical Sciences / Department of Chemistry / Unilever Centre for Molecular Informatics / Panton Discussions / View Item'. The Apollo logo and a search bar are also present. The main content area is titled 'Open Content Mining' and features a sidebar on the left with navigation options like 'All of Apollo', 'Communities & Collections', 'Authors', 'Titles', 'Keywords', 'Type', 'This Collection', 'Authors', 'Titles', 'Keywords', 'Type', 'Statistics', and 'View Usage Statistics'. The main content area includes a thumbnail of the article, a 'View / Open Files' section with a PDF link, and metadata sections for 'Citation', 'Description', 'Abstract', 'Keywords', 'Identifiers', 'Publication Date', 'ISBN', and 'Language'.

UNIVERSITY OF CAMBRIDGE

Apollo Home / School of the Physical Sciences / Department of Chemistry / Unilever Centre for Molecular Informatics / Panton Discussions / View Item

Search Apollo
Advanced search

Open Content Mining

All of Apollo

- > Communities & Collections
- > Authors
- > Titles
- > Keywords
- > Type

This Collection

- > Authors
- > Titles
- > Keywords
- > Type

Statistics

- > View Usage Statistics

View / Open Files

- article text (PDF, 9Mb)

Authors

Murray-Rust, Peter

Publication Date

2012-09-24

ISBN

to be assigned

Language

English

Citation

Murray-Rust, P. (2012). Open Content Mining.

Description

Conference for the Fellows of OpenForum Academy - 24th September 2012 Brussels

Abstract

Abstract— We present evidence that content-mining of scholarly articles is now technically feasible and highly valuable both. However researchers and information technologist are blocked by legal and contractual barriers from using it and developing the methodologies. We review the problems and propose changes in legal policy which we have already submitted to the UK's Hargreaves report on intellectual property reform. We put forward the fundamental rights of scholars and embed them in a manifesto: "The right to read is the right to mine", "Users and providers should encourage machine processing, and "Facts don't belong to anyone".

Keywords

Open Content Mining, Index Terms—Open Knowledge, Content mining, Hargreaves process, Text mining, publishers, legal barriers

Identifiers

This record's URL:
<http://www.dspace.cam.ac.uk/handle/1810/243749>

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

CORE Recommender

Type
Conference Object

Metadata
[Show full item record](#)

Rights
Attribution 2.0 UK: England & Wales

Licence URL:
<http://creativecommons.org/licenses/by/2.0/uk/>

Recommended or similar items

Suggested articles | Suggested articles in Apollo

 **Effectively and Efficiently Mining Frequent Patterns from Dense Graph Streams on Disk**
Provided by: Elsevier - Publisher Connector | **Publisher:** The Authors. Published by Elsevier B.V. | **Year:** 2014
By Braun Peter, Cameron Juan J., Cuzzocrea Alfredo, Jiang Fan, Leung Carson K.

 **The right to read is the right to mine: Text and data mining copyright exceptions introduced in the UK.**
Provided by: LSE Research Online | **Publisher:** London School of Economics and Political Science | **Year:** 2014
By Mounce Ross

 **Global boom, local impacts: Mining revenues and subnational outcomes in Peru 2007-2011**
Provided by: EconStor | **Publisher:** Washington, DC: Inter-American Development Bank (IDB) | **Year:** 2014
By Zambrano Omar, Robles Marcos, Laos Denisse

 **Environmental security, mining and good governance : mining regulation in the Kyrgyz region. A review**
Provided by: UEF Electronic Publications | **Publisher:** University of Eastern Finland
By Honkonen T

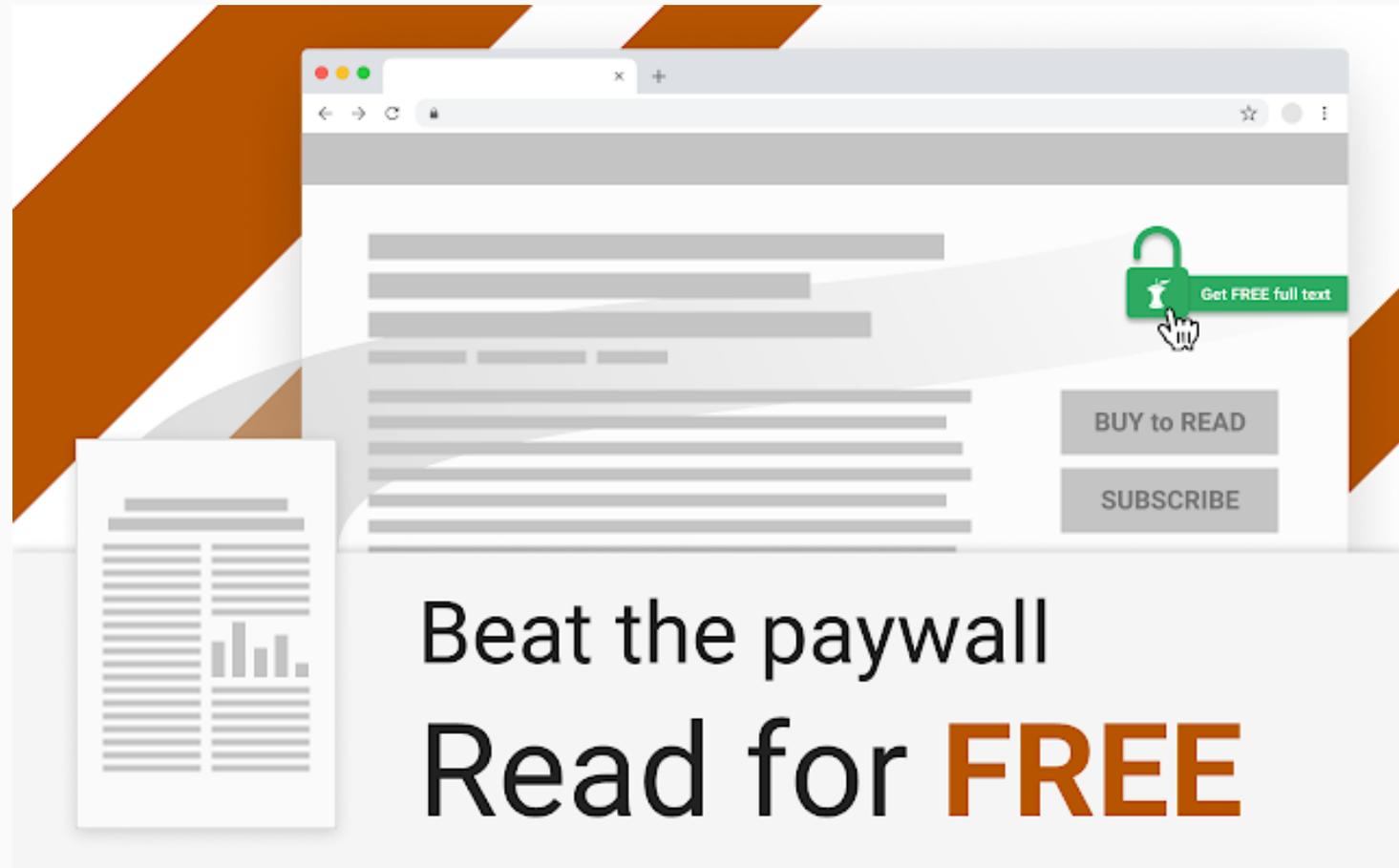
 **A Case Study of Data Analysis Process and Tools for a Consulting Company**
Provided by: Aaltodoc Publication Archive | **Year:** 2012
By Gong Peng

Powered by 

- Recommending relevant content to users from across all free content
- Recommender plugin for repositories
- <https://core.ac.uk/service/s/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/>

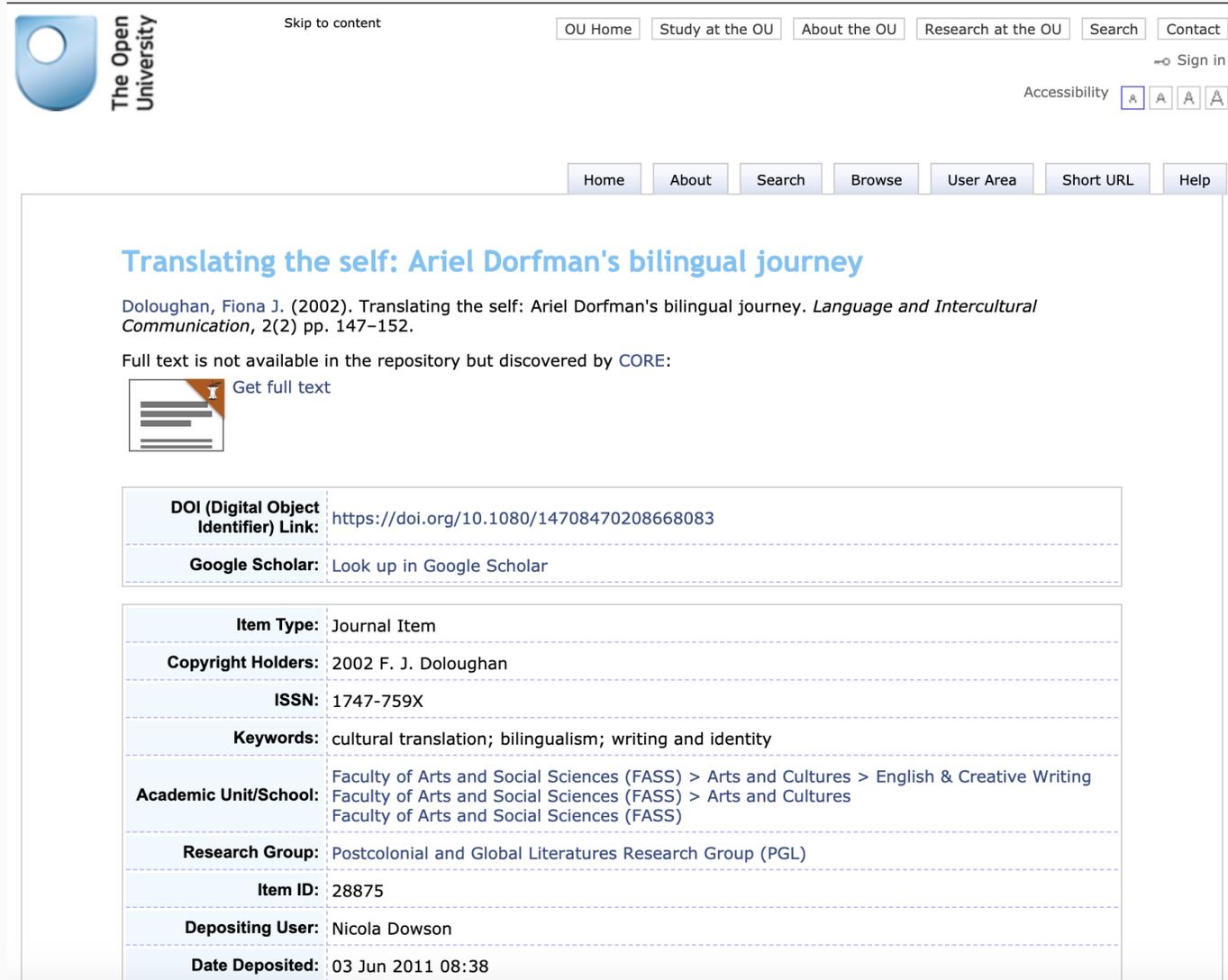
CORE Discovery demonstration

The screenshot displays a web browser window with the following elements:

- Browser Tab:** "Geographical trends in academi x"
- Address Bar:** "content.iospress.com/articles/data-science/ds190015"
- Header:** "IOS Press" logo and "IOS Press Content Library" text. Navigation links: "Help", "About us", "Contact us".
- Menu Bar:** "Home", "Journals", "Cart", "Log in / Register".
- Search Bar:** A text input field with the placeholder "Search" and a "Search" button.
- Filters:** "Published between:" followed by two "YYYY" input boxes and the word "and".
- Help Link:** "Search syntax help".
- Main Content Area:** A large, empty light blue rectangular area.

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.



The screenshot shows the Open University website interface. At the top, there is a navigation bar with links for 'OU Home', 'Study at the OU', 'About the OU', 'Research at the OU', 'Search', and 'Contact'. Below this, there are additional navigation links: 'Home', 'About', 'Search', 'Browse', 'User Area', 'Short URL', and 'Help'. The main content area displays a search result for the article 'Translating the self: Ariel Dorfman's bilingual journey' by Fiona J. Doloughan (2002). The article title is highlighted in blue. Below the title, the citation information is provided: 'Doloughan, Fiona J. (2002). Translating the self: Ariel Dorfman's bilingual journey. *Language and Intercultural Communication*, 2(2) pp. 147-152.' A note indicates that the full text is not available in the repository but was discovered by CORE. A 'Get full text' button is visible next to a document icon. Below this, there are two tables of metadata. The first table contains links for the DOI and Google Scholar. The second table contains detailed metadata including item type, copyright holders, ISSN, keywords, academic unit, research group, item ID, depositing user, and date deposited.

Skip to content

OU Home Study at the OU About the OU Research at the OU Search Contact

Sign in

Accessibility

Home About Search Browse User Area Short URL Help

Translating the self: Ariel Dorfman's bilingual journey

Doloughan, Fiona J. (2002). Translating the self: Ariel Dorfman's bilingual journey. *Language and Intercultural Communication*, 2(2) pp. 147-152.

Full text is not available in the repository but discovered by CORE:

 [Get full text](#)

DOI (Digital Object Identifier) Link:	https://doi.org/10.1080/14708470208668083
Google Scholar:	Look up in Google Scholar

Item Type:	Journal Item
Copyright Holders:	2002 F. J. Doloughan
ISSN:	1747-759X
Keywords:	cultural translation; bilingualism; writing and identity
Academic Unit/School:	Faculty of Arts and Social Sciences (FASS) > Arts and Cultures > English & Creative Writing Faculty of Arts and Social Sciences (FASS) > Arts and Cultures Faculty of Arts and Social Sciences (FASS)
Research Group:	Postcolonial and Global Literatures Research Group (PGL)
Item ID:	28875
Depositing User:	Nicola Dowson
Date Deposited:	03 Jun 2011 08:38

CORE Search

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

The screenshot displays the CORE Search website interface. At the top, the CORE logo is visible on the left, and 'Services About' links are on the right. A search bar contains the term 'allele' and a 'Search' button. Below the search bar, the page is divided into a left sidebar for refining search results and a main content area showing search results.

Refine your search

- Publication type**
 - with fulltext only
- Year**
 - 1999 - 2018
- Languages**
 - English 49,370
 - German 852
 - Italian 208
 - French 203
 - Portuguese 196
 - Spanish 72
 - Hungarian 42
 - Indonesian 32
 - Croatian 31
 - Dutch 14
- Journals**
 - Jurnal Pengolahan Hasil 1,856
- Repositories**
 - PubMed Central 37,730
 - Elsevier - Publisher 33,369
 - MUCC (Crossref) 31,314
 - Springer - Publisher 28,410
 - Frontiers - Publisher 4,480
 - Edinburgh Research Explorer 3,085
 - Harvard University - DASH 2,034
 - Radboud Repository 2,029
 - ZORA 1,891
 - Erasmus University Digital 1,760
- Permalink**
 - <https://core.ac.uk/search/>

Showing results for allele (368,795 articles found) Sort by: Relevance ▾

The almond Sf allele: an allele in question
By Rafel Socias | Company, Osama Kodad, Angel V. Fernández | Martí and José Manuel Alonso Segura
Repository: citaREA Repositorio Electrónico Agroalimentario | 2011
...ALLELE TERMINOLOGY The mistakes in allele sequences observed by Bošković et al. (2007) led them to incorrectly name a new allele, S30, which they wrongly considered different from Sf, although it is identical to Sf, but showing a different activity
Get PDF (216 KB) Similar articles

7 GT repetition on 5T allele
By M. Maschio, Z. Cannioto, M. Morgutti and F. Poli
Repository: Elsevier - Publisher Connector | 2007
...Supported by: Ministry of Health and Regione Lombardia. 7 GT repetition on 5T allele M. Maschio, Z. Cannioto, M. Morgutti, F. Poli. Department of Reproduction and Development Sciences, University of Trieste-IRCCS Burlo Garofolo, Trieste, Italy 5T allele
Get PDF (57 KB) Similar articles Cite ▾

Allele-Specific p53 Mutant Reactivation
By Xin Yu, Alexei Vazquez, Arnold J. Levine and Darren R. Carpizo
Repository: Elsevier - Publisher Connector | 2012
...Cancer Cell Allele-Specific Mutant p53 Reactivating Compound 620 Cancer Cell 21, 614-625, May 15, 2012 *2012 Elsevier Inc...
Get PDF (781 KB) Similar articles Cite ▾

Allele coding in genomic evaluation
By Ismo Strandén and Ole F Christensen
Repository: Jukuri | 2011
...For the centered allele coding system, we have $vm = 1n2' 01n$, i.e., $Zc = Z0 - 1n1n1'n20$. Note that $vm/2$ gives the allele frequencies of the markers in the data. The allele coding transformation allows shifts in the allele codes. The 101 allele coding
Get PDF (489 KB) Similar articles Cite ▾

Allele Frequencies in Multigene Families
By S. Padmadisastra
Repository: Neliti | 2010
...A., "Allele Frequencies in Multigene Families. II. Coalescent Approach", Theor. Pop. Biol. 35 (1989b), 161-180...
Get PDF (211 KB) Similar articles

Useful links

- Blog
- Contact us
- Services
- Cookies
- About CORE
- Privacy notice

Writing about CORE?
Discover our research outputs and cite our work.

The Open University **Jisc**

CORE is a not-for-profit service delivered by the Open University and Jisc.

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



**UNIVERSITY OF
CAMBRIDGE**



turnitin®



I R I S . A I



**Open
Access
Button**



**The
University
Of
Sheffield.**



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 corpus**

Thank you!

<https://core.ac.uk>