BSDTAg

Big Scientific Data and Text Analytics group :
AI for open and responsible research

CORE

2023

Applications of AI in Academic Libraries and Archives:

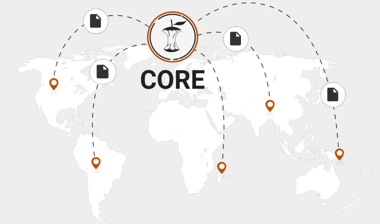
Machine learning from and for open research.

Prof. Petr Knoth

The Open University

BSDTAg

Big Scientific Data and Text Analytics group : AI for open and responsible research



CORE is the world's most used aggregator of **Open Access** papers, collating and enriching content from over **11,000 repositories**.

- **>20 Million** monthly active users (MAU)
- **34 Million** full-text research papers hosted by CORE.
- **260 Million** metadata records

Signatory of Principles of Open Scholarly Infrastructure **(POSI)**

CORE delivers **services** for HEIs, researchers, funders and commercial partners, offering seamless access to research.

Content discovery	Raw data services	Managing content
Search	API	Repository Dashboard
Recommender	Dataset	Identifiers
Discovery	FastSync	OAI Resolver

Providing seamless access to open research for humans and machines.

Commercial Partners

ontochem IT SOLUTIONS

CAKTUS

cypris

EZASSi Technology to innovate faster.

turnitin

- Innovation and trends analysis
- Plagiarism detection
- Fact checking
- Finance
- Health

Institutional Members

University of OXFORD

Lancaster University

University of HUDDERSFIELD Inspiring global professionals

University of BRISTOL

QUEEN'S UNIVERSITY BELFAST

THE UNIVERSITY OF CHICAGO

University of St Andrews FOUNDED 1413

a.r.u.

UNIVERSITY OF BIRMINGHAM

White Rose university consortium

UNIVERSITY of York

UNIVERSITY OF LEEDS

University of Sheffield

32 supporting or sustaining members

Research areas

➔ AI Applications in Research Evaluation (e.g. citation type classification, bibliometrics, impact assessment)

➔ Automatic Expert Finder systems (e.g. for peer-review and grant applications)

➔ Deduplication, document classification, rapid systematic reviews

➔ Research graphs: entity extraction (affiliation, author, etc.)

➔ Research recommender systems and academic search

Dr. Petr Knoth : Senior Research Fellow in Text and Data Mining petr.knoth@open.ac.uk

The Open University

CORE and the OA landscape

CORE's mission is

to index all open access research worldwide and deliver unrestricted access for all.

We are here to support and advance the Open Access / Open Research movement

WE ARE

the world's **most used** collection of open access research papers from repositories

WE ARE

a **not-for-profit** scholarly infrastructure dedicated to the open access mission, **adopters of POSI** principles.

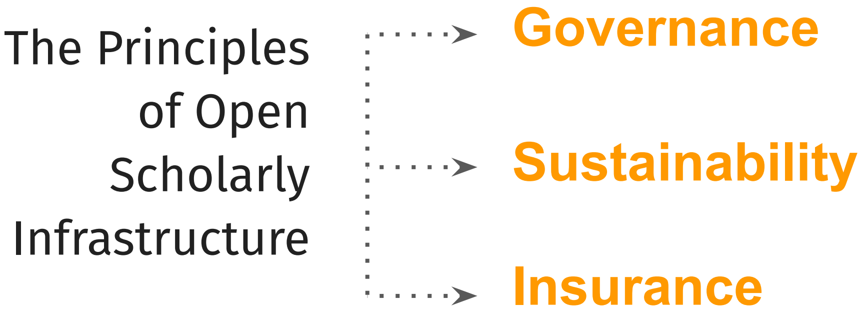
WE

provide solutions for content management, discovery and scalable machine access to research.

WE

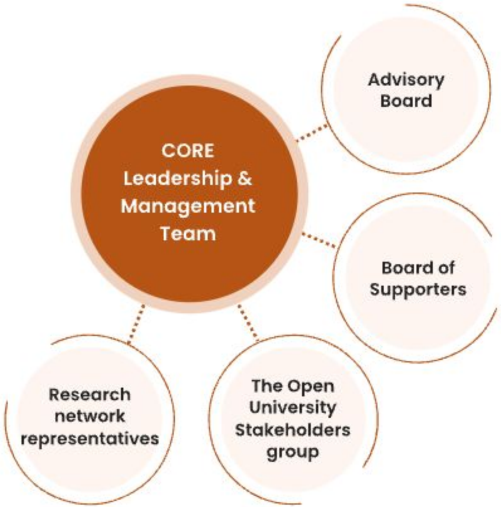
serve the global network of repositories and journals by increasing discoverability and reuse of open access content.

CORE - An adopter of POSI



CORE is a mission-driven not-for-profit endeavour and a signatory of the **Principles of Open Scholarly Infrastructure.**

CORE Community Governance


CORE Leadership & Management Team
Advisory Board
Board of Supporters
Research network representatives
The Open University Stakeholders group

Advisory Board

- Advises on strategic directions
- Ensures mission alignment with the needs of the open research community

Board of Supporters

- Helping to identify requirements and prioritise the development roadmap
- Represents the interests of the global open repositories and journals network.

The Open University Stakeholder Group

- Assumes overall financial and legal responsibility for CORE's obligations.
- Provides institutional support and resources for CORE (e.g. HR, financial, legal, infrastructure).

Research network representatives

- Ensures relevance and provides guidance on effectively supporting the open research community.

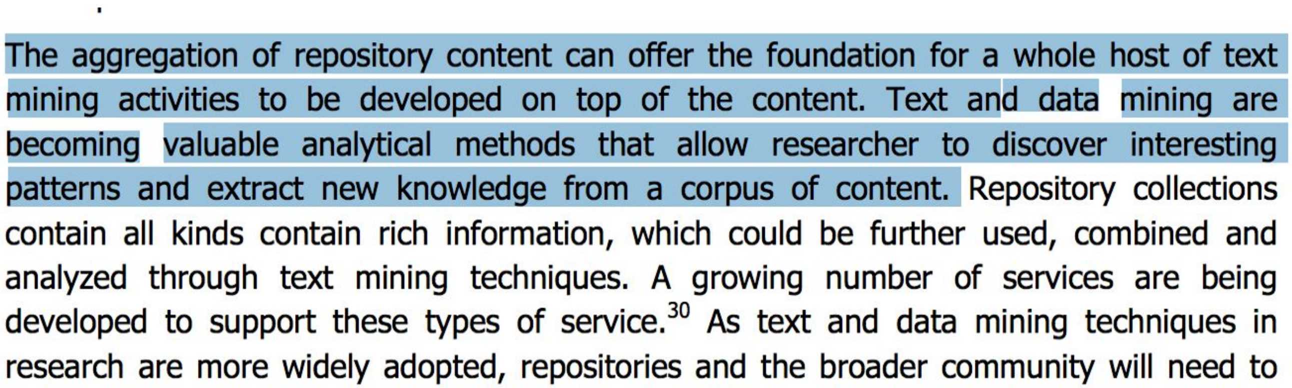
CORE Leadership & Management Team

- Is responsible for the day to day operation of CORE.
- Takes operational decisions with guidance from the governance groups

Outline

1. How can Artificial Intelligence and Machine Learning (**AI/ML**) applied to research papers benefit and **transform research**
2. The crucial role of repositories in providing **machine access** to research content.
3. Using AI/ML for **research intelligence** and improving repository workflows

Outline



The aggregation of repository content can offer the foundation for a whole host of text mining activities to be developed on top of the content. Text and data mining are becoming valuable analytical methods that allow researcher to discover interesting patterns and extract new knowledge from a corpus of content. Repository collections contain all kinds contain rich information, which could be further used, combined and analyzed through text mining techniques. A growing number of services are being developed to support these types of service.[30] As text and data mining techniques in research are more widely adopted, repositories and the broader community will need to

COAR

Confederation of Open Access Repositories

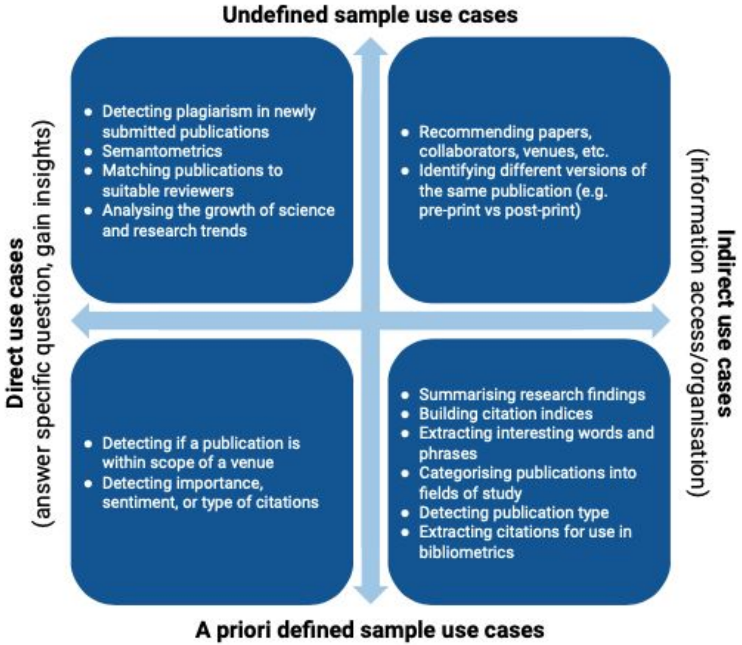
How can AI/ML transform research

The importance of open research literature

Research literature documents the knowledge we have assembled as human species.

The wide variety of use cases over research literature

- A limited number of use cases can be satisfied with a sample of scholarly content.
- Many use cases require machine access to all existing research from everywhere and always up-to-date.
- High cost when a repository does not participate in the open network, by not providing machine access. Some use cases significantly affected.



Undefined sample use cases

- Detecting plagiarism in newly submitted publications
- Semantometrics
- Matching publications to suitable reviewers
- Analysing the growth of science and research trends

- Recommending papers, collaborators, venues, etc.
- Identifying different versions of the same publication (e.g. pre-print vs post-print)

- Detecting if a publication is within scope of a venue
- Detecting importance, sentiment, or type of citations

- Summarising research findings
- Building citation indices
- Extracting interesting words and phrases
- Categorising publications into fields of study
- Detecting publication type
- Extracting citations for use in bibliometrics

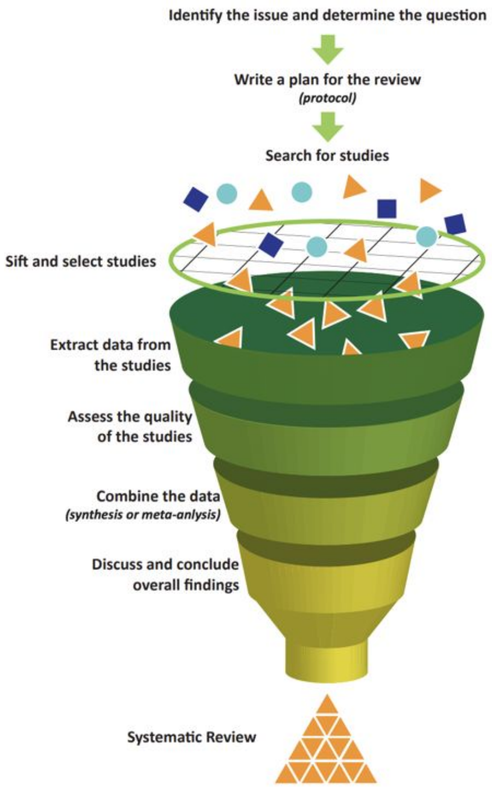
A priori defined sample use cases

Direct use cases (answer specific question, gain insights)

Indirect use cases (information access/organisation)

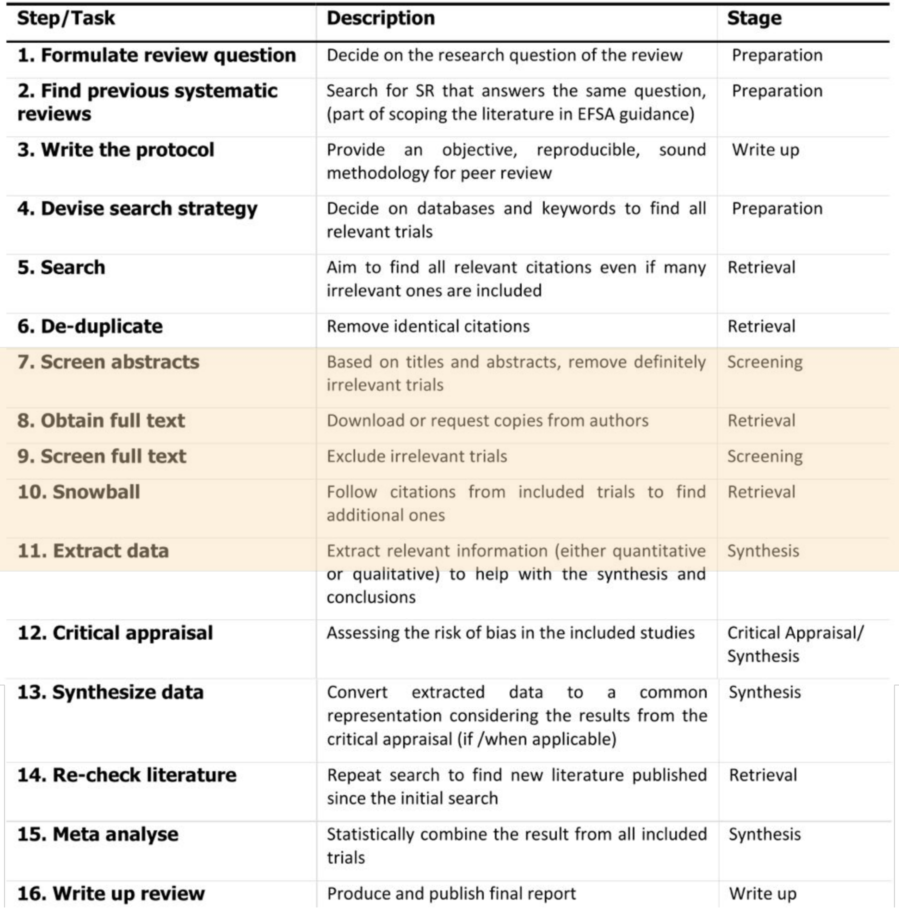
AI for systematic reviews

- **Systematic reviews**
  - Time consuming
- **Rapid reviews**
  - Limitation on the number of outcomes, interventions and comparators
- **Living reviews**
  - Live updates to historic systematic reviews with the help of recommender system


Identify the issue and determine the question
Write a plan for the review
(protocol)
Search for studies
Sift and select studies
Extract data from the studies
Assess the quality of the studies
Combine the data
(synthesis or meta-anlysis)
Discuss and conclude overall findings
Systematic Review

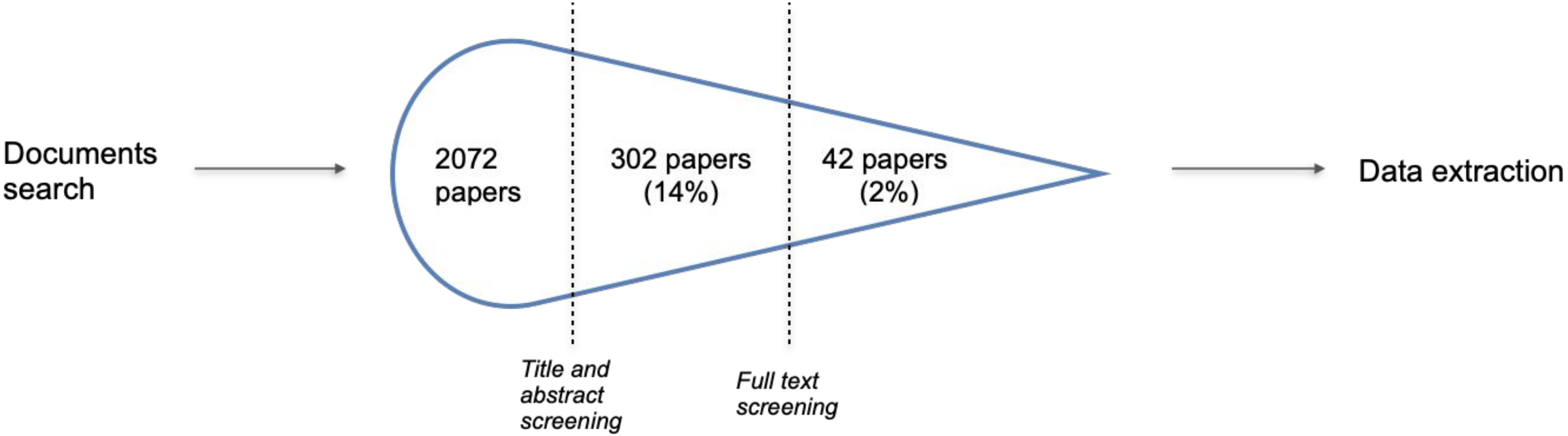
AI for systematic reviews

- → Involves many steps
- → Some of teh most-time consuming can be automated



Step/Task	Description	Stage
1. Formulate review question	Decide on the research question of the review	Preparation
2. Find previous systematic reviews	Search for SR that answers the same question, (part of scoping the literature in EFSA guidance)	Preparation
3. Write the protocol	Provide an objective, reproducible, sound methodology for peer review	Write up
4. Devise search strategy	Decide on databases and keywords to find all relevant trials	Preparation
5. Search	Aim to find all relevant citations even if many irrelevant ones are included	Retrieval
6. De-duplicate	Remove identical citations	Retrieval
7. Screen abstracts	Based on titles and abstracts, remove definitely irrelevant trials	Screening
8. Obtain full text	Download or request copies from authors	Retrieval
9. Screen full text	Exclude irrelevant trials	Screening
10. Snowball	Follow citations from included trials to find additional ones	Retrieval
11. Extract data	Extract relevant information (either quantitative or qualitative) to help with the synthesis and conclusions	Synthesis
12. Critical appraisal	Assessing the risk of bias in the included studies	Critical Appraisal/ Synthesis
13. Synthesize data	Convert extracted data to a common representation considering the results from the critical appraisal (if /when applicable)	Synthesis
14. Re-check literature	Repeat search to find new literature published since the initial search	Retrieval
15. Meta analyse	Statistically combine the result from all included trials	Synthesis
16. Write up review	Produce and publish final report	Write up

AI for systematic reviews



Documents search

2072 papers

302 papers (14%)

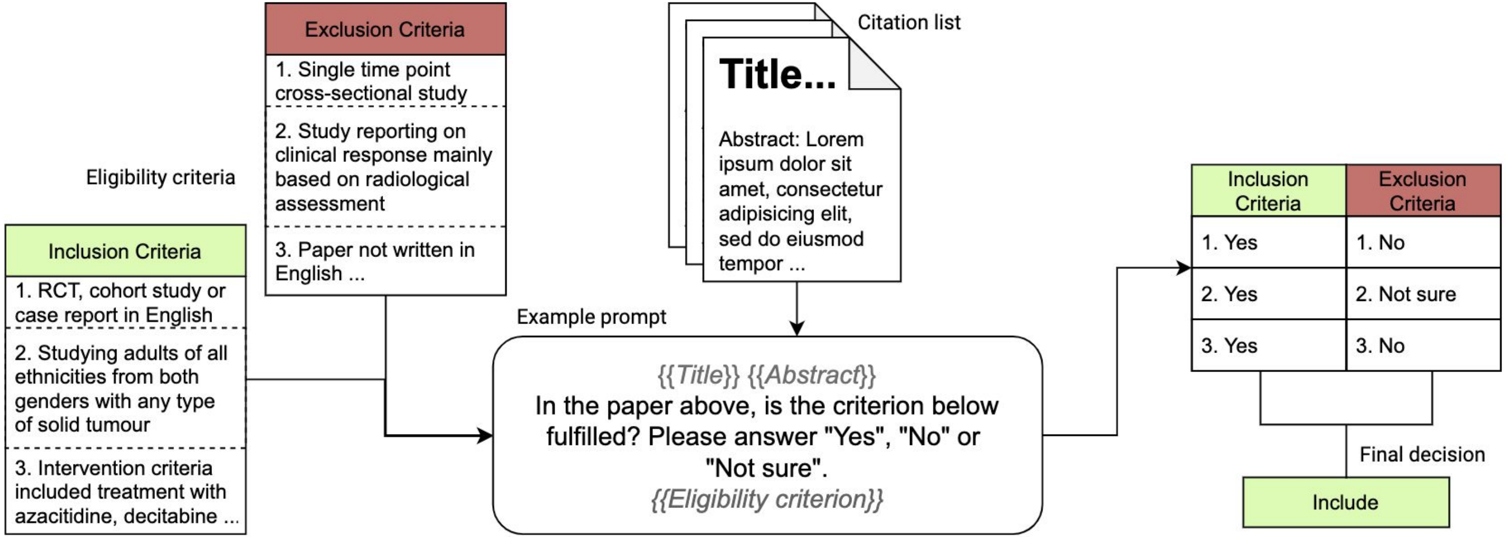
42 papers (2%)

Data extraction

Title and abstract screening

Full text screening

AI for systematic reviews



Eligibility criteria

Inclusion Criteria

1. RCT, cohort study or case report in English
2. Studying adults of all ethnicities from both genders with any type of solid tumour
3. Intervention criteria included treatment with azacitidine, decitabine ...

Exclusion Criteria

1. Single time point cross-sectional study
2. Study reporting on clinical response mainly based on radiological assessment
3. Paper not written in English ...

Citation list

Title...

Abstract: Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor ...

Example prompt

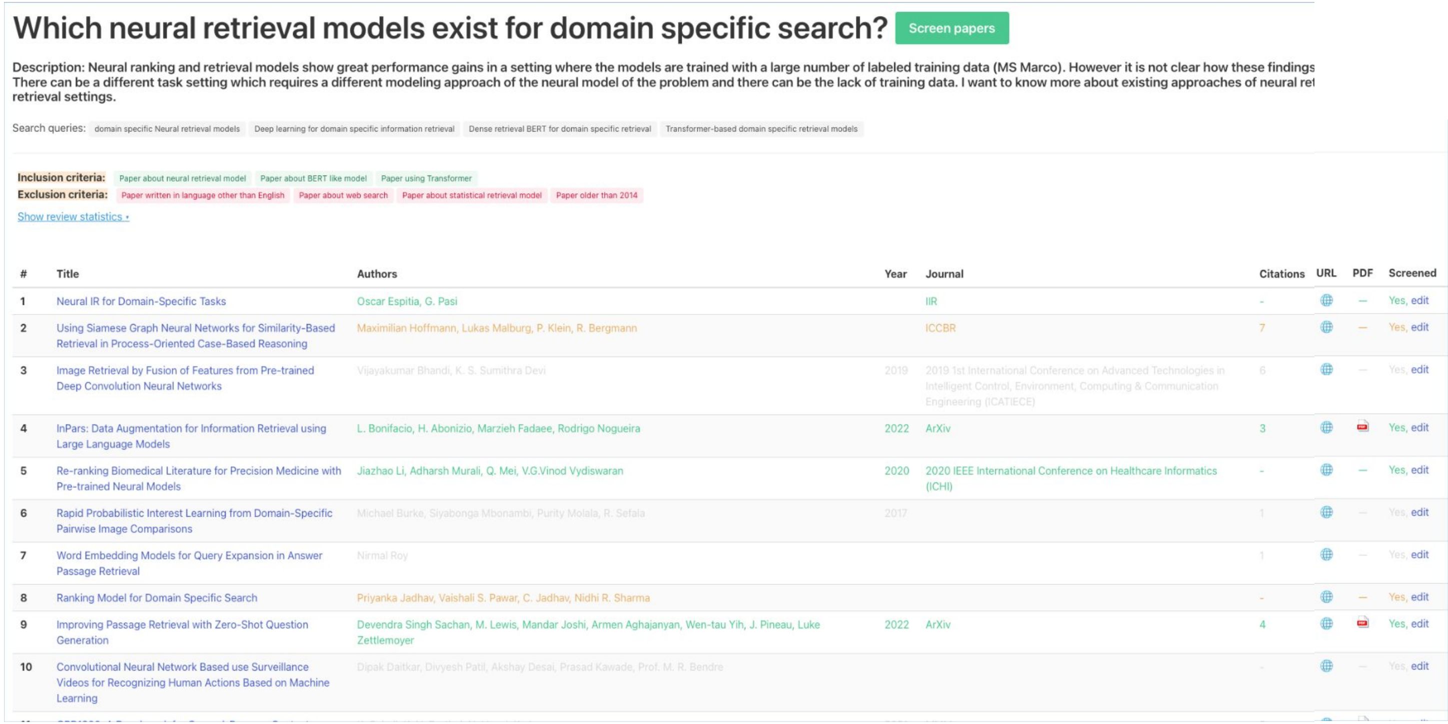
{{Title}} {{Abstract}}
In the paper above, is the criterion below fulfilled? Please answer "Yes", "No" or "Not sure".
{{Eligibility criterion}}

Inclusion Criteria	Exclusion Criteria
1. Yes	1. No
2. Yes	2. Not sure
3. Yes	3. No

Final decision

Include

AI for systematic reviews



Which neural retrieval models exist for domain specific search?

Description: Neural ranking and retrieval models show great performance gains in a setting where the models are trained with a large number of labeled training data (MS Marco). However it is not clear how these findings
There can be a different task setting which requires a different modeling approach of the neural model of the problem and there can be the lack of training data. I want to know more about existing approaches of neural ret
retrieval settings.

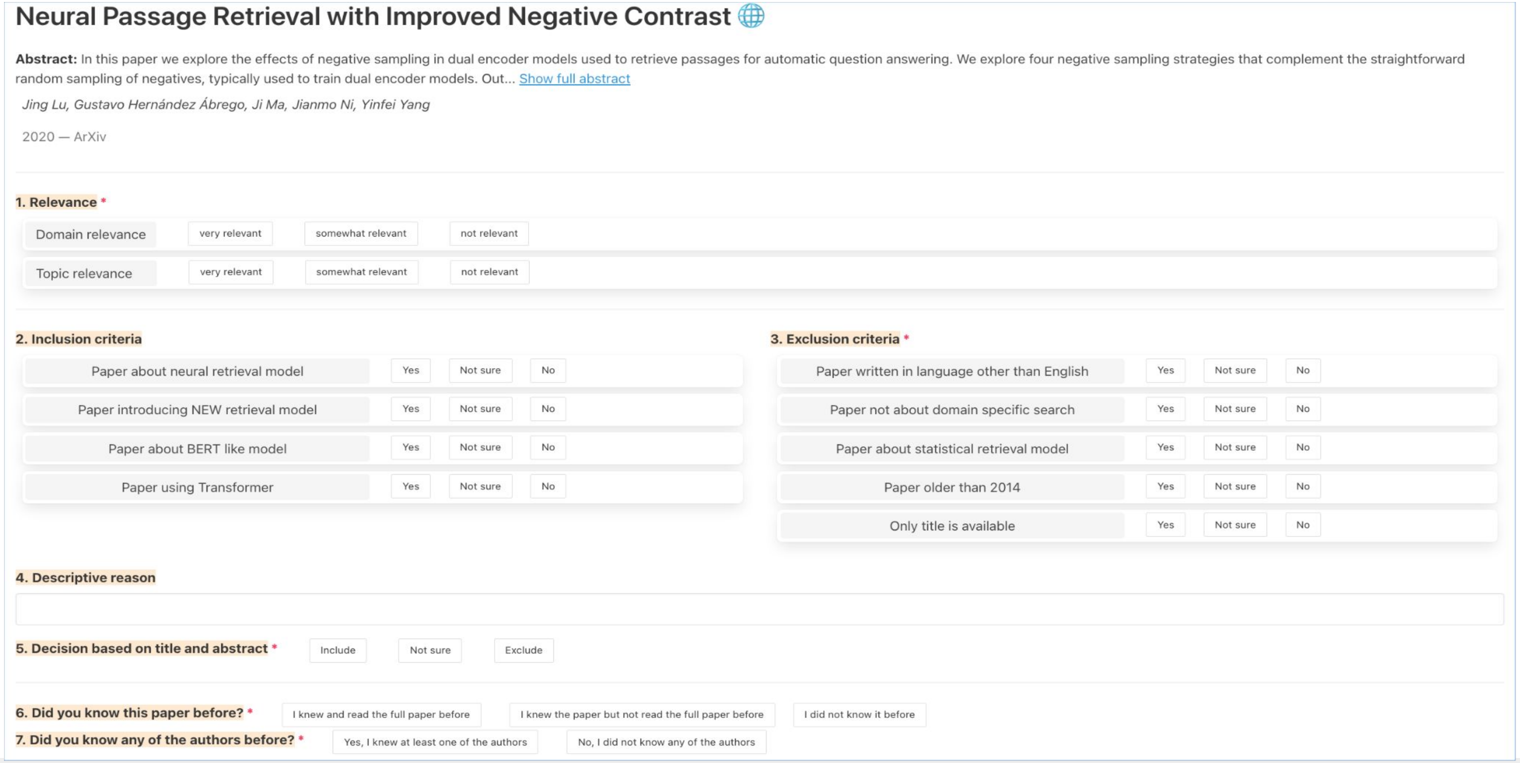
Search queries: domain specific Neural retrieval models | Deep learning for domain specific information retrieval | Dense retrieval BERT for domain specific retrieval | Transformer-based domain specific retrieval models

Inclusion criteria: Paper about neural retrieval model | Paper about BERT like model | Paper using Transformer

Exclusion criteria: Paper written in language other than English | Paper about web search | Paper about statistical retrieval model | Paper older than 2014

#	Title	Authors	Year	Journal	Citations	URL	PDF	Screened
1	Neural IR for Domain-Specific Tasks	Oscar Espitia, G. Pasi		IIR	-	🌐	—	Yes, edit
2	Using Siamese Graph Neural Networks for Similarity-Based Retrieval in Process-Oriented Case-Based Reasoning	Maximilian Hoffmann, Lukas Malburg, P. Klein, R. Bergmann		ICCBR	7	🌐	—	Yes, edit
3	Image Retrieval by Fusion of Features from Pre-trained Deep Convolution Neural Networks	Vijayakumar Bhandi, K. S. Sumithra Devi	2019	2019 1st International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE)	6	🌐	—	Yes, edit
4	InPars: Data Augmentation for Information Retrieval using Large Language Models	L. Bonifacio, H. Abonizio, Marzieh Fadaee, Rodrigo Nogueira	2022	ArXiv	3	🌐	📄	Yes, edit
5	Re-ranking Biomedical Literature for Precision Medicine with Pre-trained Neural Models	Jiazhao Li, Adharsh Murali, Q. Mei, V.G.Vinod Vydiswaran	2020	2020 IEEE International Conference on Healthcare Informatics (ICHI)	-	🌐	—	Yes, edit
6	Rapid Probabilistic Interest Learning from Domain-Specific Pairwise Image Comparisons	Michael Burke, Siyabonga Mbonambi, Purity Molala, R. Sefala	2017		1	🌐	—	Yes, edit
7	Word Embedding Models for Query Expansion in Answer Passage Retrieval	Nirmal Roy			1	🌐	—	Yes, edit
8	Ranking Model for Domain Specific Search	Priyanka Jadhav, Vaishali S. Pawar, C. Jadhav, Nidhi R. Sharma			-	🌐	—	Yes, edit
9	Improving Passage Retrieval with Zero-Shot Question Generation	Devendra Singh Sachan, M. Lewis, Mandar Joshi, Armen Aghajanyan, Wen-tau Yih, J. Pineau, Luke Zettlemoyer	2022	ArXiv	4	🌐	📄	Yes, edit
10	Convolutional Neural Network Based use Surveillance Videos for Recognizing Human Actions Based on Machine Learning	Dipak Daitkar, Divyesh Patil, Akshay Desai, Prasad Kawade, Prof. M. R. Bendre			-	🌐	—	Yes, edit

AI for systematic reviews



Neural Passage Retrieval with Improved Negative Contrast 🌐

Abstract: In this paper we explore the effects of negative sampling in dual encoder models used to retrieve passages for automatic question answering. We explore four negative sampling strategies that complement the straightforward random sampling of negatives, typically used to train dual encoder models. Out... Show full abstract

Jing Lu, Gustavo Hernández Ábrego, Ji Ma, Jianmo Ni, Yinfei Yang

2020 — ArXiv

1. Relevance *

Domain relevance	very relevant	somewhat relevant	not relevant
Topic relevance	very relevant	somewhat relevant	not relevant

2. Inclusion criteria

Paper about neural retrieval model	Yes	Not sure	No
Paper introducing NEW retrieval model	Yes	Not sure	No
Paper about BERT like model	Yes	Not sure	No
Paper using Transformer	Yes	Not sure	No

3. Exclusion criteria *

Paper written in language other than English	Yes	Not sure	No
Paper not about domain specific search	Yes	Not sure	No
Paper about statistical retrieval model	Yes	Not sure	No
Paper older than 2014	Yes	Not sure	No
Only title is available	Yes	Not sure	No

4. Descriptive reason

5. Decision based on title and abstract * Include Not sure Exclude

6. Did you know this paper before? * I knew and read the full paper before I knew the paper but not read the full paper before I did not know it before

7. Did you know any of the authors before? * Yes, I knew at least one of the authors No, I did not know any of the authors

AI for citation typing and research assessment

11 of 34 **REF2014** Peer Review Panels used citation data to 'inform' their decisions

REF GPA results highly correlated with citation data in these domains

Addition of citation type information can allow for better modelling of how research is being used.

Potential for development of new metrics that leverage enhanced citation information

> *'The pilot exercise concluded that citation information is not sufficiently robust to be used formulaically or as a primary indicator of quality in the REF'*

HEFCE. Report on the pilot exercise to develop bibliometric indicators for the Research Excellence Framework;

	UoA	mn2017	med2017	mn2014	med2014
1	Chemistry	0.663	0.802	0.637	0.738
2	Biological Sciences	0.782	0.797	0.688	0.785
3	Aero. Mech. Chem. Engineering	0.771	0.758	0.745	0.760
4	Social Work and Policy	0.697	0.752	0.629	0.635
5	Computer Science and Informatics	0.715	0.743	0.720	0.678

AI for citation typing and research assessment

- Knowing not only that something was cited, but WHY it was cited.

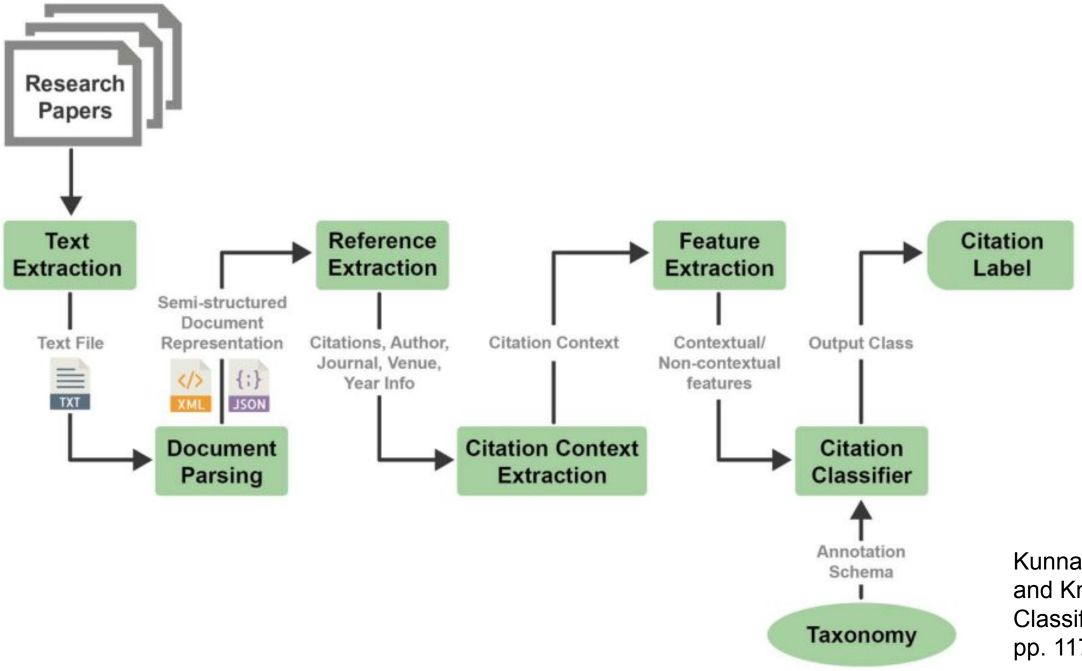
- Built ACT Dataset of >11,000 citations annotated by authors according to classification schema

- Ran 2 Shared Tasks to establish benchmarks for SoA classification models using ACT and extended ACT2 datasets

- Currently investigating extended / dynamic citation contexts to improve model performance

Citation Function	Examples
BACKGROUND	Most of the participatory models to design educational games are founded on educational theories and game design (see for example: Amory, 2007; #CITATION_TAG).
COMPARES_CONTRASTS	Similar observations have been made in the past [30] [31] [32] [33] [34], although others have reported either no relationship or a negative association with SES [#CITATION_TAG].
EXTENSION	This database is the result of a mandatory questionnaire about the home to work displacements and the mobility management measures at large workplaces in Belgium (#CITATION_TAG).
FUTURE	We are thus exploring the option of using datasets such as CrossRef 12, Dimensions 13, OpenCitations [11], and Core [#CITATION_TAG].
MOTIVATION	To illustrate, consider the motivation given by #CITATION_TAG in developing their Bayesian account of word learning.
USES	The diffraction patterns from single crystal measurements were indexed with a home-made program based on the Fit2D software [#CITATION_TAG].

A prototypical citation intent classification system



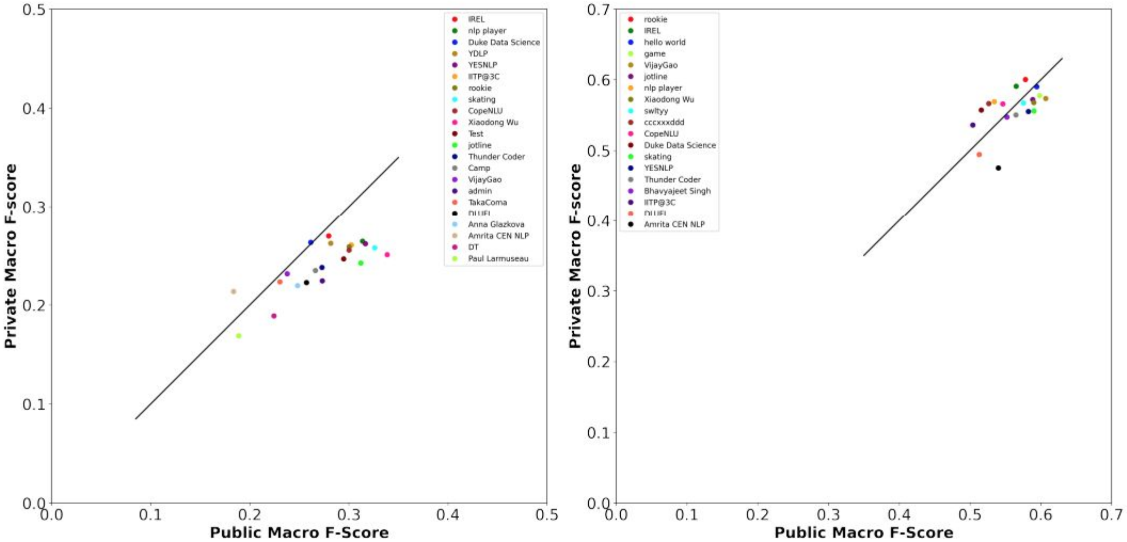
Kunnath, Suchetha N.; Herrmannova, Drahomira; Pride, David and Knoth, Petr (2022). A Meta-analysis of Semantic Classification of Citations. Quantitative Science Studies, 2(4) pp. 1170–1215. DOI: https://doi.org/10.1162/qss_a_00159

Evaluation / shared tasks for citation classification

- Citation Context Classification (3C) shared task

- ACT 2 dataset

- Same conditions for every team



(a) Subtask A

(b) Subtask B

Figure 2: Public Vs Private Macro F-Score performance on the leaderboard

Kunnath, Suchetha N.; Pride, David; Herrmannova, Drahomira and Knoth, Petr **Overview of the 2021 SDP 3C Citation Context Classification Shared Task**. In: *Proceedings of the Second Workshop on Scholarly Document Processing, Association for Computational Linguistics*, Stroudsburg, PA, pp. 150–158. URL: https://aclanthology.org/2021.sdp-1.21/

Table 4. Institution-level Pearson correlations between machine learning predictions with 50% used for training and actual scores for articles 2014-18, following Strategy 1 (averaged across 10 iterations) and aggregated by institution for UoAs 1-11 and 16.

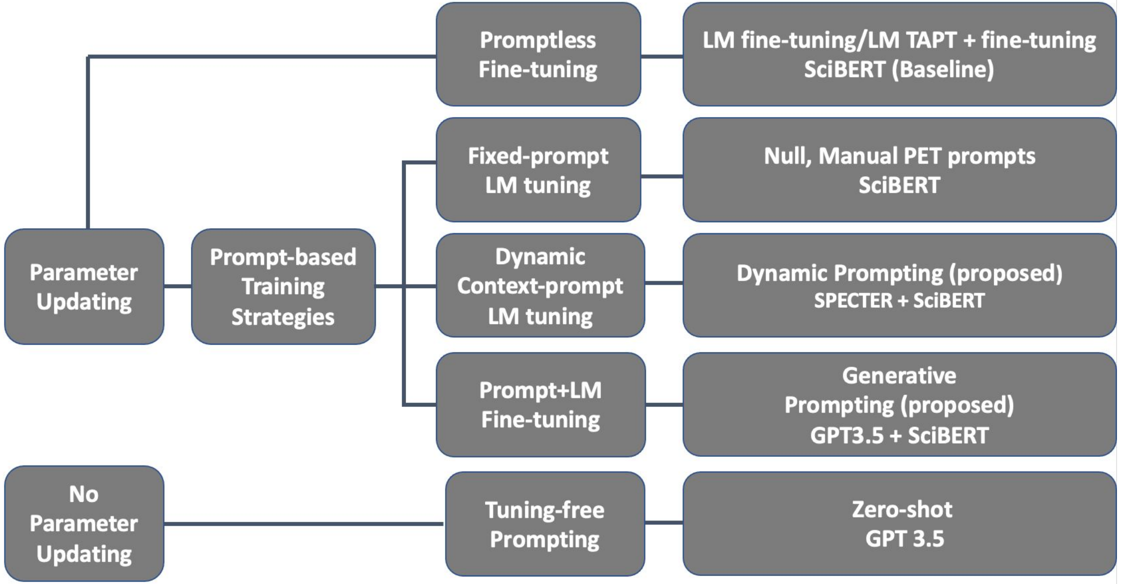
UoA	Actual vs machine learning predicted average score	Actual vs machine learning predicted total score
1:Clinical Medicine	0.895	0.998
2:Public Health, Health Services and Primary Care	0.906	0.995
3:Allied Health Professions, Dentistry, Nursing & Pharmacy	0.747	0.982
4:Psychology, Psychiatry and Neuroscience	0.844	0.995
5:Biological Sciences	0.885	0.995
6:Agriculture, Food and Veterinary Sciences	0.759	0.975
7:Earth Systems and Environmental Sciences	0.840	0.986
8:Chemistry	0.897	0.978
9:Physics	0.855	0.989
10:Mathematical Sciences	0.664	0.984
11:Computer Science and Informatics	0.724	0.945
16:Economics and Econometrics	0.862	0.974

AI for citation typing and research assessment

Significant performance improvement of parameter updating methods across a variety of prompting strategies over promptless fine-tuning.

Dynamic context-based prompts significantly improve model scores for both datasets and surpass the performance on the 3C shared task benchmark.

Kunnath, Suchetha N.; Pride, David and Knoth, Petr (2022) **Prompting Strategies for Citation Classification** In: *CIKM 2023*



Parameter Updating

Prompt-based Training Strategies

Promptless Fine-tuning

LM fine-tuning/LM TAPT + fine-tuning SciBERT (Baseline)

Fixed-prompt LM tuning

Null, Manual PET prompts SciBERT

Dynamic Context-prompt LM tuning

Dynamic Prompting (proposed) SPECTER + SciBERT

Prompt+LM Fine-tuning

Generative Prompting (proposed) GPT3.5 + SciBERT

No Parameter Updating

Tuning-free Prompting

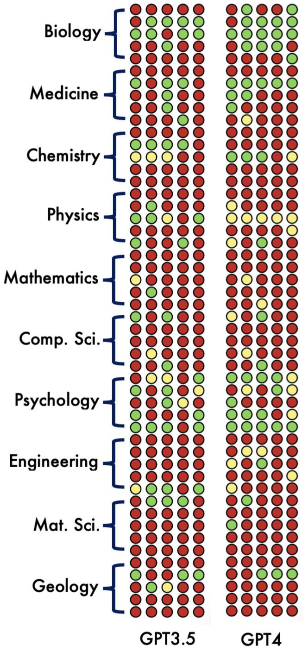
Zero-shot GPT 3.5

GPT Fact vs fiction

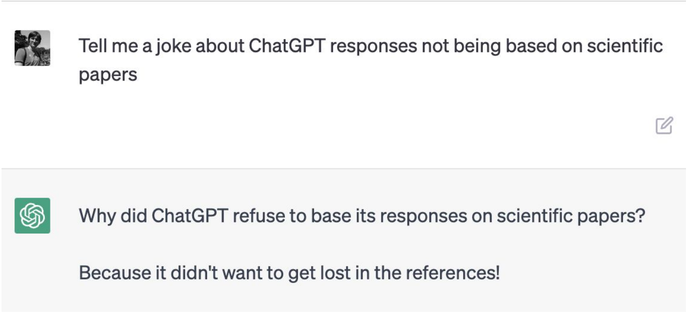
50 questions from across 10 different domains.

Each answer supported by 5 references to papers

72.5% of citations provided by GPT3.5 were fictional.
71.2% of citations provided by GPT4 were fictional.



	GPT3.5	GPT4
🔴 Fictional	72.5%	71.2%
🟡 Conflated	5.5%	9.2%
🟢 Factual	22.0%	19.6%

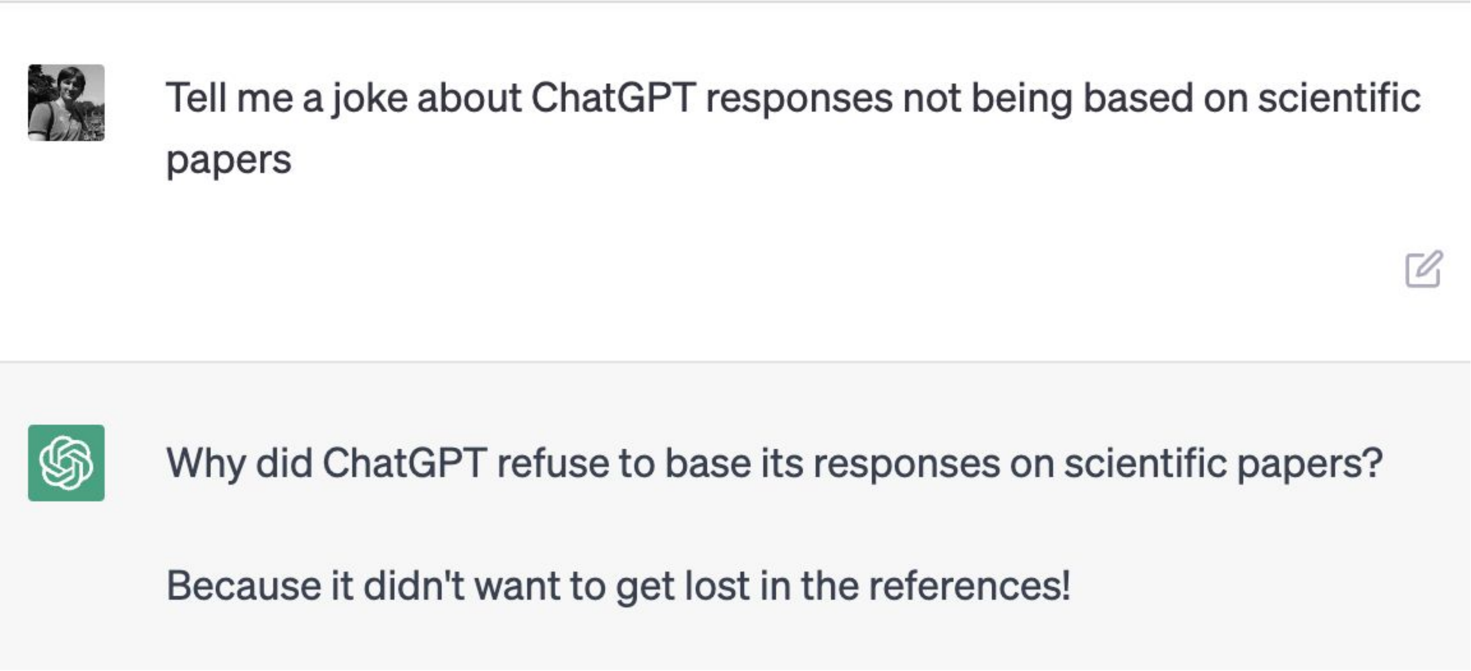


Tell me a joke about ChatGPT responses not being based on scientific papers

Why did ChatGPT refuse to base its responses on scientific papers?

Because it didn't want to get lost in the references!

ChatGPT



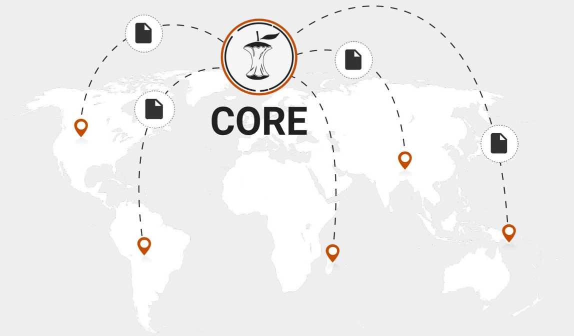
Tell me a joke about ChatGPT responses not being based on scientific papers

Why did ChatGPT refuse to base its responses on scientific papers?

Because it didn't want to get lost in the references!

AI for credible trustworthy question answering (CORE-GPT)

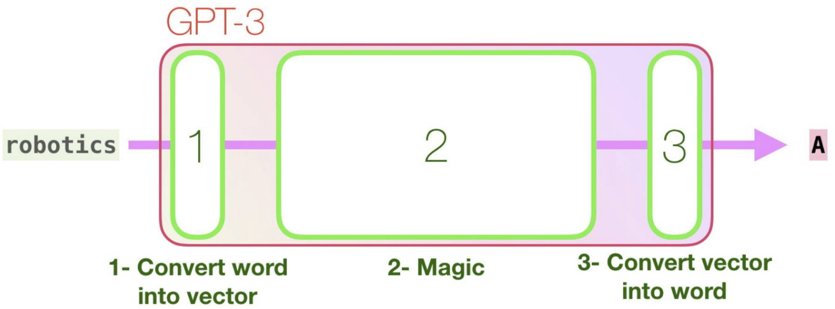
CORE is the world's largest collection of Open Access papers, collating and enriching content from over **11,000** data providers.



- **>20 Million** monthly active users
- **34 Million** full-text research papers hosted by CORE.
- **260 Million** metadata records

GPT large language models*

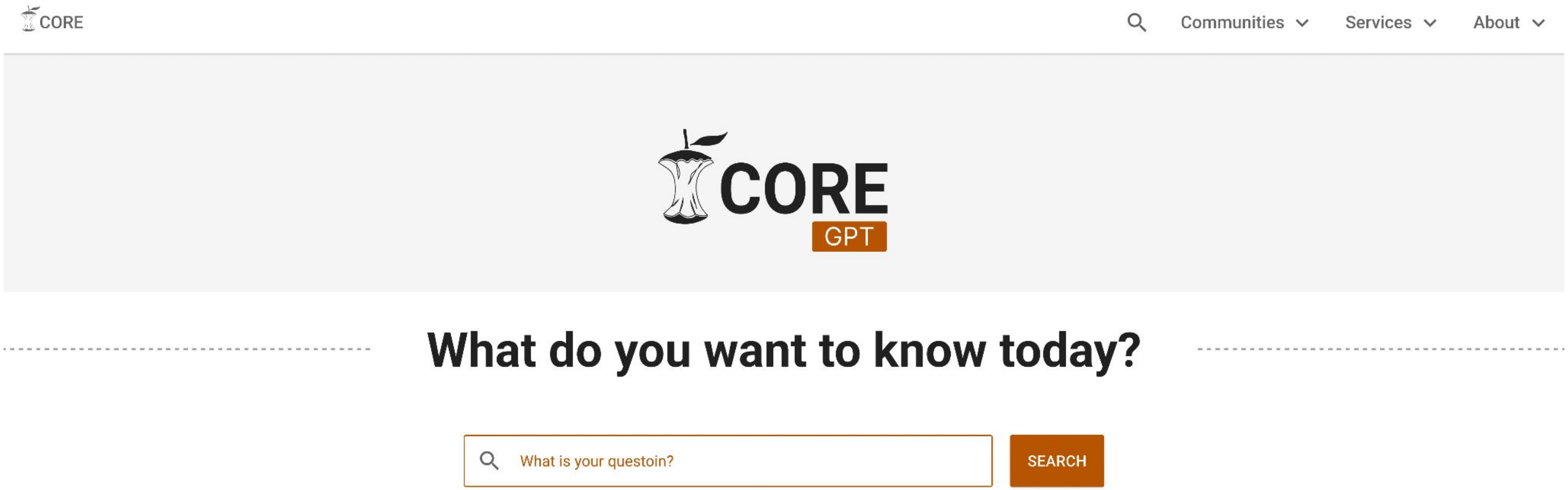
- Can comprehend context and generate human-like text
- Can infer meaning from large-scale data



@JayAlammar

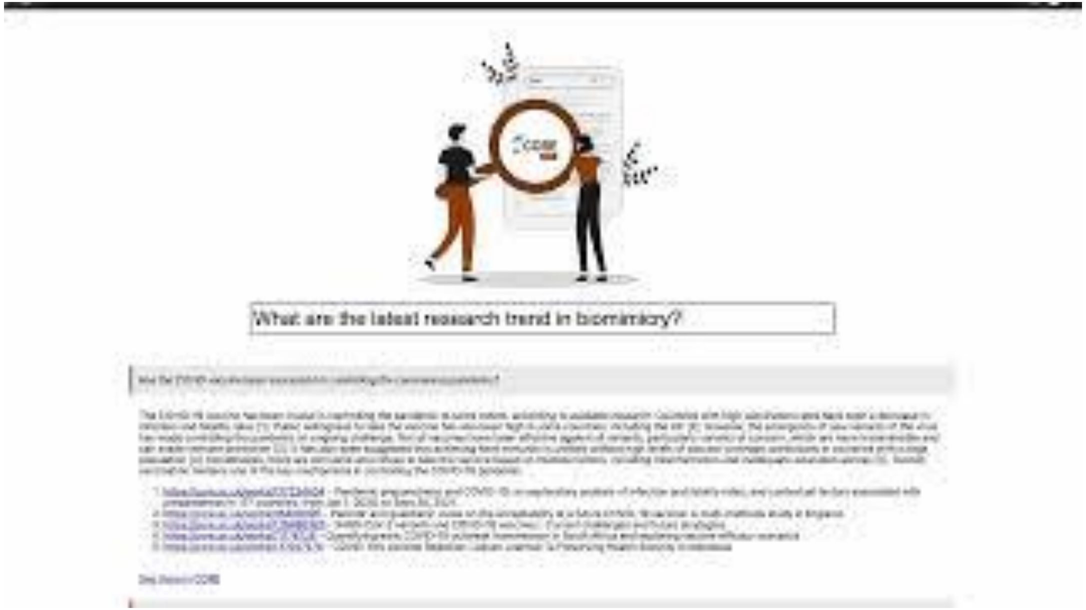
*Other large language models are available

Introducing CORE-GPT



What do you want to know today?

CORE-GPT Results



What are the latest research trend in biomimicry?

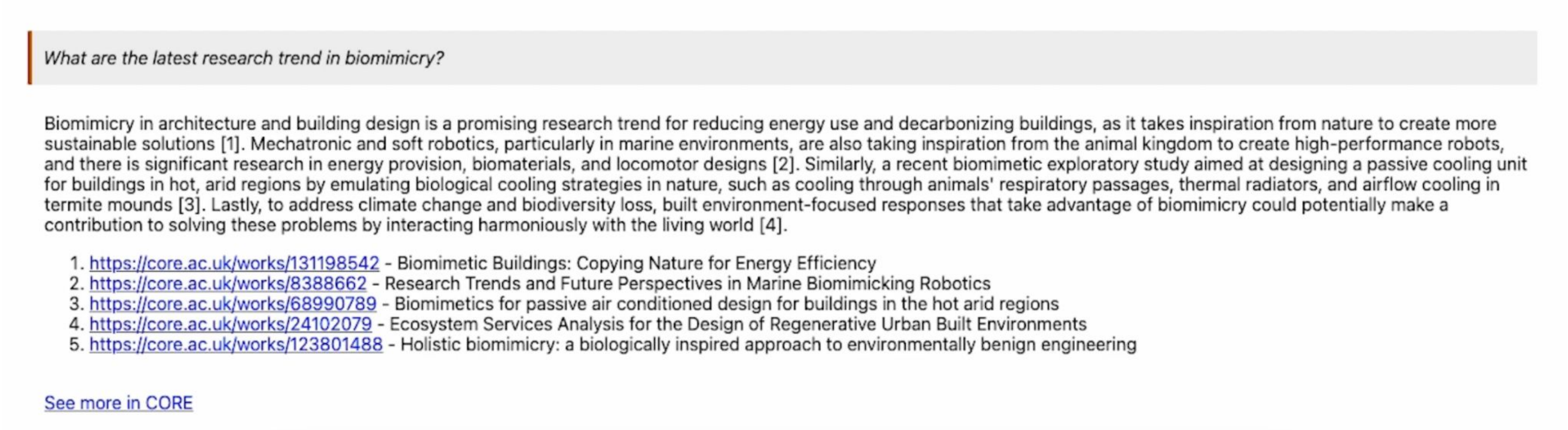
Has the COVID vaccine been successful in controlling the coronavirus pandemic?

The COVID-19 vaccine has been crucial in controlling the pandemic [illegible] [1]. [illegible] [2]. [illegible] [3]. [illegible] [4]. [illegible] [5]. [illegible] controlling the COVID-19 pandemic.

1. [illegible] - Pandemic preparedness and COVID-19: an exploratory analysis of infection and fatality rates, and contextual factors associated with preparedness in 177 countries, from Jan 1, 2020, to Sept 30, 2021.
2. [illegible] - Parental and guardians' views on the acceptability of a future COVID-19 vaccine: a multi-methods study in England
3. [illegible] - SARS-CoV-2 variants and COVID-19 vaccines: Current challenges and future strategies
4. [illegible] - Quantifying early COVID-19 outbreak transmission in South Africa and exploring vaccine efficacy scenarios
5. [illegible] - COVID-19 [illegible] Indonesia

Deep Search in CORE

CORE-GPT Results



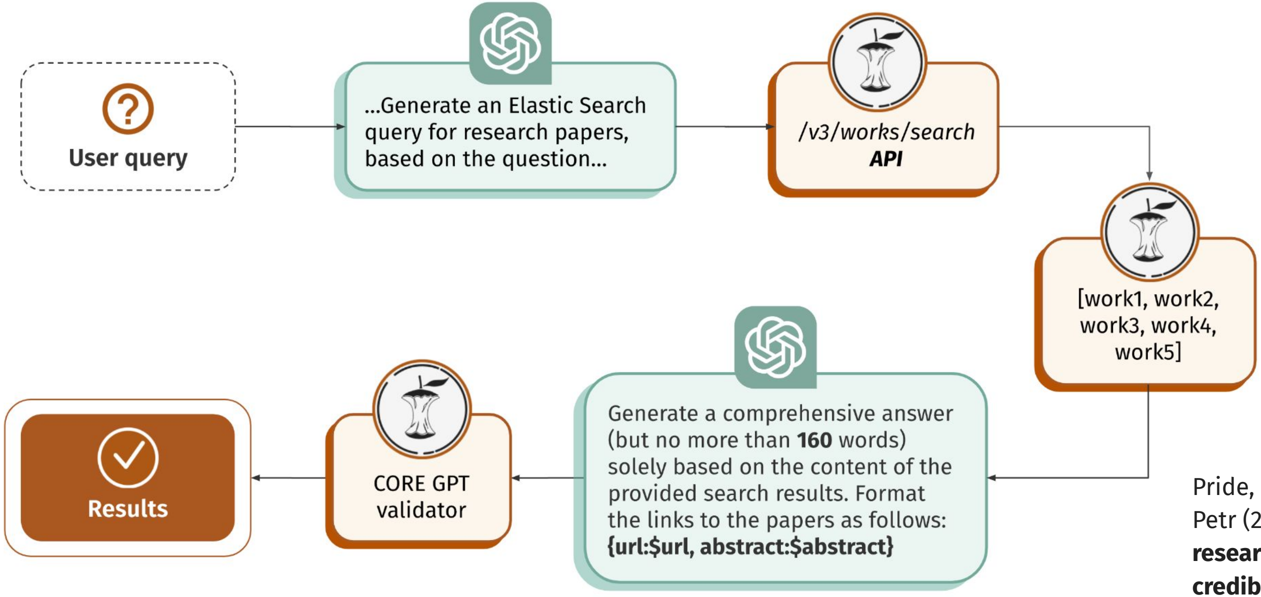
What are the latest research trend in biomimicry?

Biomimicry in architecture and building design is a promising research trend for reducing energy use and decarbonizing buildings, as it takes inspiration from nature to create more sustainable solutions [1]. Mechatronic and soft robotics, particularly in marine environments, are also taking inspiration from the animal kingdom to create high-performance robots, and there is significant research in energy provision, biomaterials, and locomotor designs [2]. Similarly, a recent biomimetic exploratory study aimed at designing a passive cooling unit for buildings in hot, arid regions by emulating biological cooling strategies in nature, such as cooling through animals' respiratory passages, thermal radiators, and airflow cooling in termite mounds [3]. Lastly, to address climate change and biodiversity loss, built environment-focused responses that take advantage of biomimicry could potentially make a contribution to solving these problems by interacting harmoniously with the living world [4].

1. https://core.ac.uk/works/131198542 - Biomimetic Buildings: Copying Nature for Energy Efficiency
2. https://core.ac.uk/works/8388662 - Research Trends and Future Perspectives in Marine Biomimicking Robotics
3. https://core.ac.uk/works/68990789 - Biomimetics for passive air conditioned design for buildings in the hot arid regions
4. https://core.ac.uk/works/24102079 - Ecosystem Services Analysis for the Design of Regenerative Urban Built Environments
5. https://core.ac.uk/works/123801488 - Holistic biomimicry: a biologically inspired approach to environmentally benign engineering

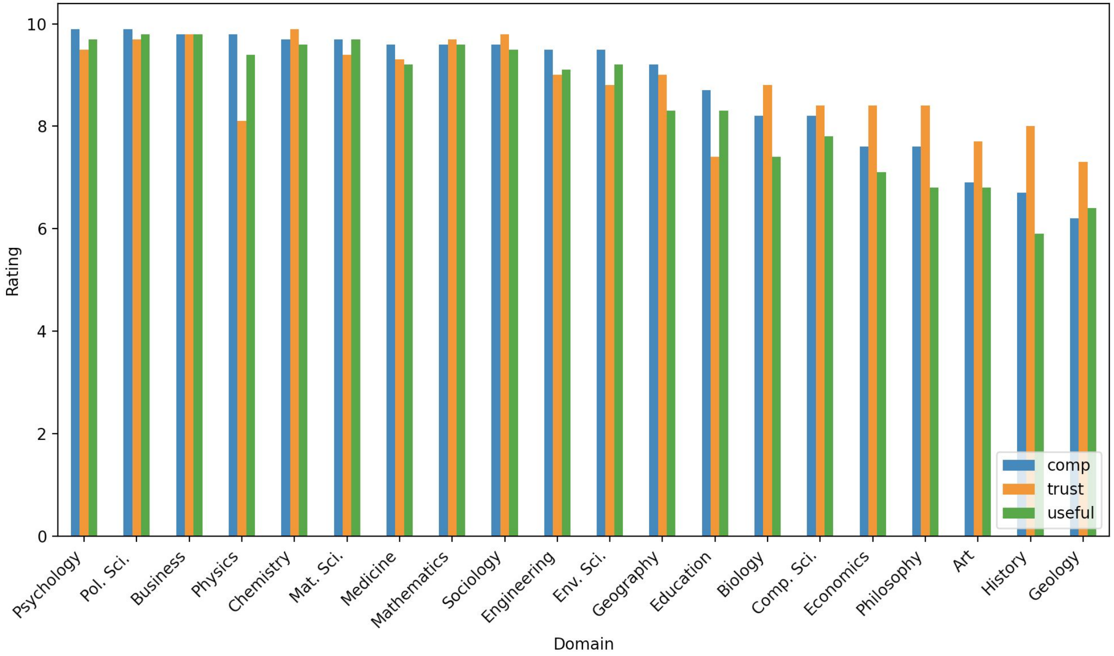
See more in CORE

How does CORE-GPT work?


User query
...Generate an Elastic Search query for research papers, based on the question...
/v3/works/search
API
[work1, work2, work3, work4, work5]
Generate a comprehensive answer (but no more than 160 words) solely based on the content of the provided search results. Format the links to the papers as follows: {url:$url, abstract:$abstract}
CORE GPT validator
Results

Pride, David; Cancellieri, Matteo and Knoth, Petr (2022) **CORE-GPT: Combining Open Access research and large language models for credible, trustworthy question answering**. In: *TPDL 2023*

How well does CORE-GPT work?



Pride, David; Cancellieri, Matteo and Knoth, Petr (2022) **CORE-GPT: Combining Open Access research and large language models for credible, trustworthy question answering**. In: *TPDL 2023*

Reflections / limitations ...

ChatGPT

- Can get confused (esp. when answers are ambiguous) mixing content from entirely semantically different uses of a concept
- Can be made to argue your way producing biased text
- It can start inventing things / hallucinate ...

CORE-GPT

- Answers need to be anchored to research papers.
- More honest about what it doesn't know => fewer hallucinations
- References make it easier to assess the trustworthiness of the answer.

Both

- Powerful at synthesizing content and creating summaries
- Able to compare and contrast
- Can get confused (esp. when answers are ambiguous) mixing content from entirely semantically different areas / uses of a concept
- Can be made to argue your way producing biased text
- Critical thinking and judgement needs to be exercised

CORE - AI Expert Finder

Prototype tool to automatically identify domain experts based on publications in >34m research papers

Applications in:

Peer review | Proposal review | Consultant/Expert recruitment

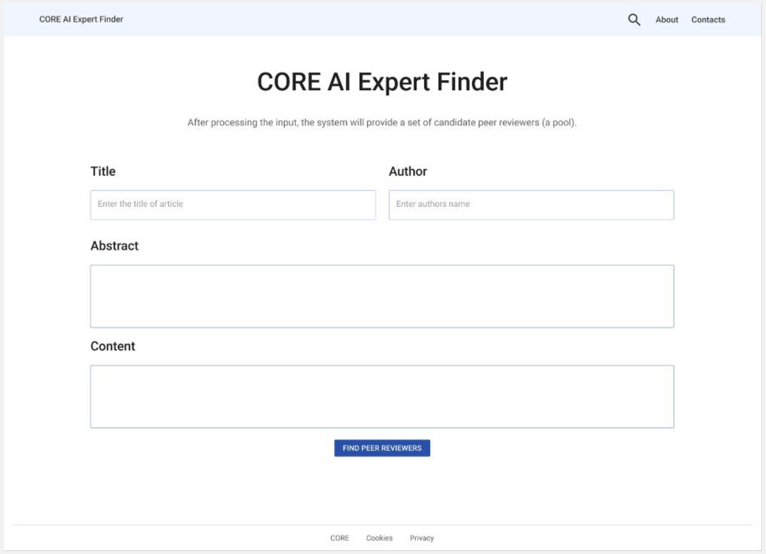
Evaluation:

- **Relevancy** - was the suggested candidate a suitable match?
- **Prior Knowledge** - was the suggested candidate previously known to the enquirer?
- **Conflict** - are there any conflicts of interest with the proposed candidate?

Results

74% of suggested candidates were suitable

34% of suggested candidates were not known to enquirer


CORE AI Expert Finder

About Contacts

CORE AI Expert Finder

After processing the input, the system will provide a set of candidate peer reviewers (a pool).

Title: Enter the title of article

Author: Enter authors name

Abstract

Content

FIND PEER REVIEWERS

CORE Cookies Privacy

CORE - AI Expert Finder

Prototype tool to automatically identify domain experts based on publications in >34m research papers

Applications in:

Peer review | Proposal review | Consultant/Expert recruitment

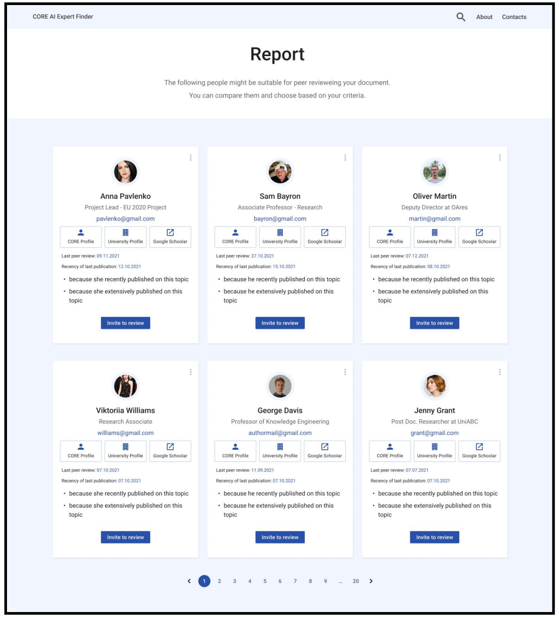
Evaluation:

- **Relevancy** - was the suggested candidate a suitable match?
- **Prior Knowledge** - was the suggested candidate previously known to the enquirer?
- **Conflict** - are there any conflicts of interest with the proposed candidate?

Results

74% of suggested candidates were suitable

34% of suggested candidates were not known to enquirer



Report

The following people might be suitable for peer revieweing your document.
You can compare them and choose based on your criteria.

Anna Pavlenko
Project Lead - EU 2020 Project
pavlenko@gmail.com
CORE Profile | University Profile | Google Schoolar
Last peer review: 09.11.2021
Recency of last publication: 12.10.2021
- because she recently published on this topic
- because she extensively published on this topic

Invite to review

Sam Bayron
Associate Professor - Research
bayron@gmail.com
CORE Profile | University Profile | Google Schoolar
Last peer review: 27.10.2021
Recency of last publication: 15.10.2021
- because he recently published on this topic
- because he extensively published on this topic

Invite to review

Oliver Martin
Deputy Director at OAres
martin@gmail.com
CORE Profile | University Profile | Google Schoolar
Last peer review: 07.12.2021
Recency of last publication: 08.10.2021
- because he recently published on this topic
- because he extensively published on this topic

Invite to review

Viktoriia Williams
Research Associate
williams@gmail.com
CORE Profile | University Profile | Google Schoolar
Last peer review: 07.10.2021
Recency of last publication: 07.10.2021
- because she recently published on this topic
- because she extensively published on this topic

Invite to review

George Davis
Professor of Knowledge Engineering
authormail@gmail.com
CORE Profile | University Profile | Google Schoolar
Last peer review: 11.09.2021
Recency of last publication: 07.10.2021
- because he recently published on this topic
- because he extensively published on this topic

Invite to review

Jenny Grant
Post Doc. Researcher at UniABC
grant@gmail.com
CORE Profile | University Profile | Google Schoolar
Last peer review: 07.07.2021
Recency of last publication: 07.10.2021
- because she recently published on this topic
- because she extensively published on this topic

Invite to review

CORE

The crucial role of repositories in providing machine access to research content.

The Open University

Principle 1

Repositories should always establish a link from the metadata record to the item the metadata record describes using a dereferencable identifier pointing to the version held locally in the repository (if applicable). The dereferencable identifier should be provided in the appropriate metadata element in the used metadata format.

Principle 2

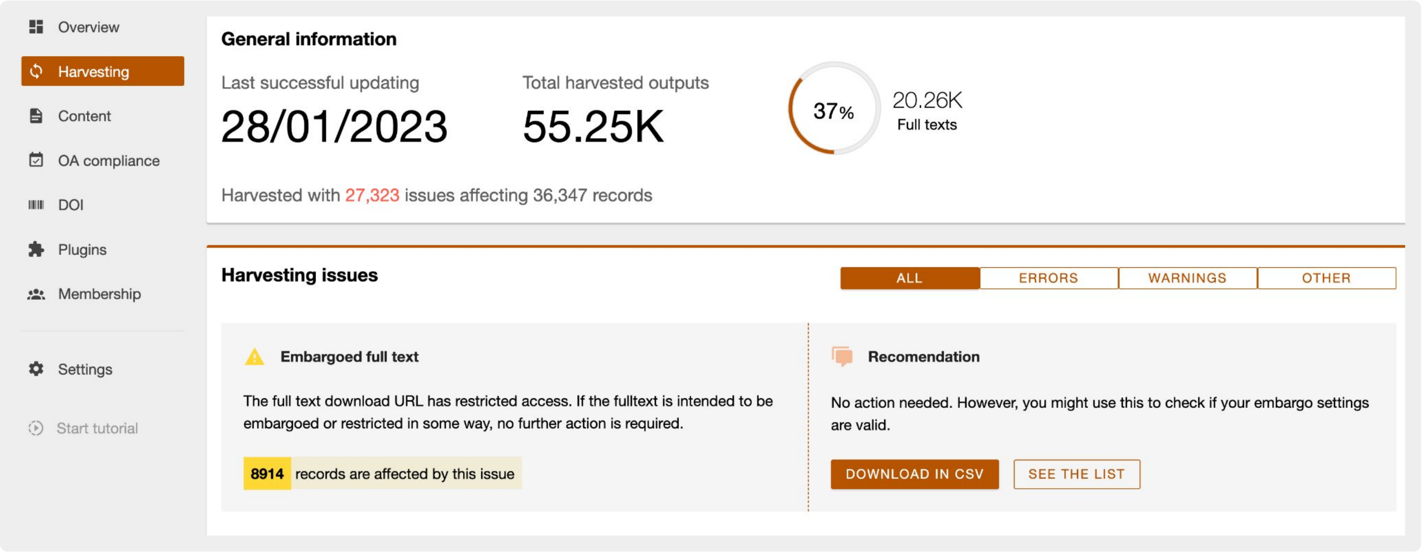
Repositories should provide universal access to machines with the same level of access as humans have. It should be possible for machines to harvest the entire content of the repository in a reasonable time to enable a machine to maintain up-to-date information about the content held in the repository.

Functional OAI-PMH endpoint

Use an external system to see how your repository is seen from the outside of your organisation.

Test, don't take that it works for granted

Monitor: the fact that it works now doesn't mean it can't go wrong when you least expect it



General information

Last successful updating
28/01/2023

Total harvested outputs
55.25K

37%
20.26K
Full texts

Harvested with 27,323 issues affecting 36,347 records

Harvesting issues

Embargoed full text

The full text download URL has restricted access. If the fulltext is intended to be embargoed or restricted in some way, no further action is required.

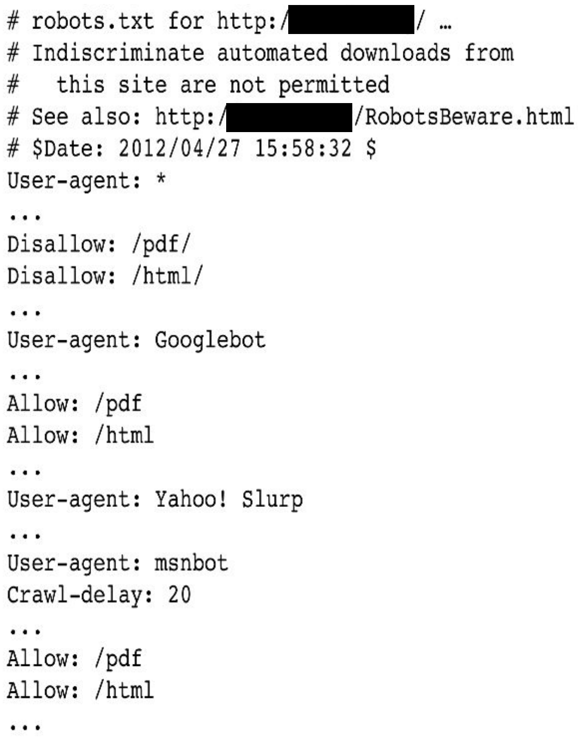
8914 records are affected by this issue

Recomendation

No action needed. However, you might use this to check if your embargo settings are valid.

Robots.txt

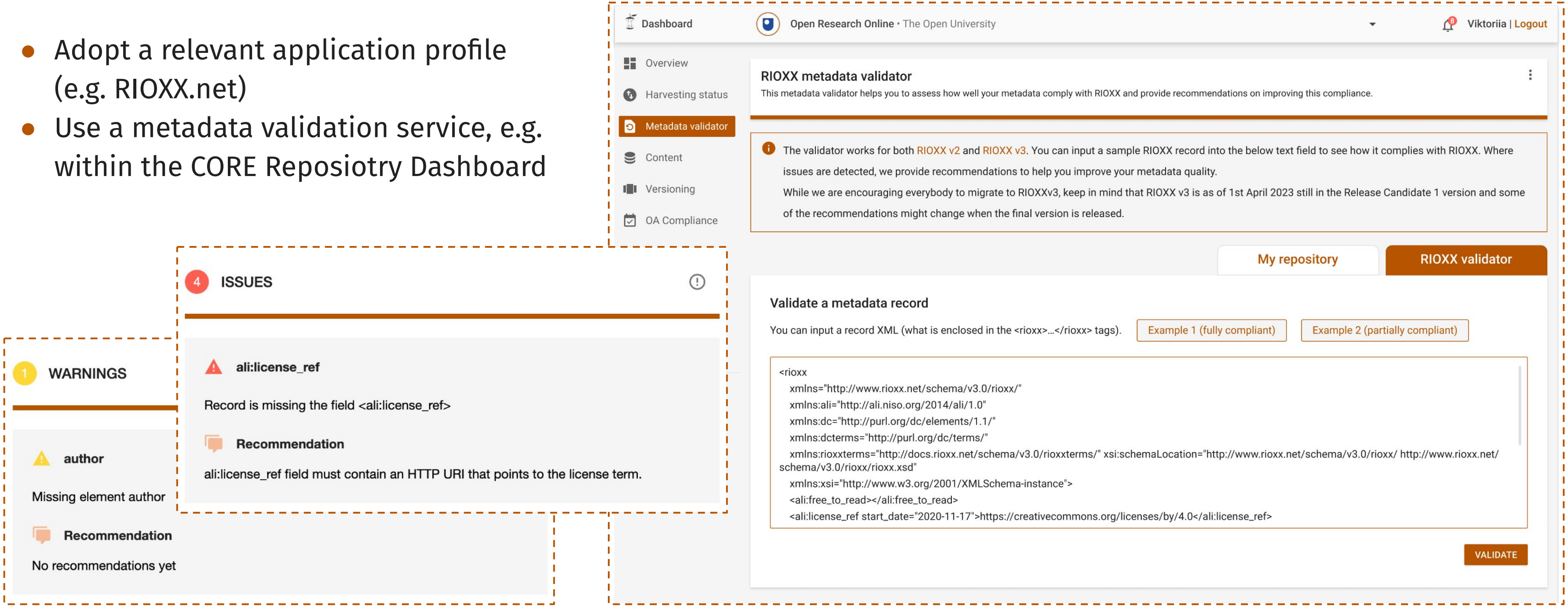
- Be careful not to block robots
- Don’t give preferential treatment



```
# robots.txt for http:/[redacted]/ …
# Indiscriminate automated downloads from
#   this site are not permitted
# See also: http:/[redacted]/RobotsBeware.html
# $Date: 2012/04/27 15:58:32 $
User-agent: *
...
Disallow: /pdf/
Disallow: /html/
...
User-agent: Googlebot
...
Allow: /pdf
Allow: /html
...
User-agent: Yahoo! Slurp
...
User-agent: msnbot
Crawl-delay: 20
...
Allow: /pdf
Allow: /html
...
```

Validate metadata

- Adopt a relevant application profile (e.g. RIOXX.net)
- Use a metadata validation service, e.g. within the CORE Repsiotry Dashboard



RIOXX metadata validator

This metadata validator helps you to assess how well your metadata comply with RIOXX and provide recommendations on improving this compliance.

The validator works for both RIOXX v2 and RIOXX v3. You can input a sample RIOXX record into the below text field to see how it complies with RIOXX. Where issues are detected, we provide recommendations to help you improve your metadata quality.

While we are encouraging everybody to migrate to RIOXXv3, keep in mind that RIOXX v3 is as of 1st April 2023 still in the Release Candidate 1 version and some of the recommendations might change when the final version is released.

Validate a metadata record

You can input a record XML (what is enclosed in the <rioxx>...</rioxx> tags).

```
<rioxx
   xmlns="http://www.rioxx.net/schema/v3.0/rioxx/"
   xmlns:ali="http://ali.niso.org/2014/ali/1.0"
   xmlns:dc="http://purl.org/dc/elements/1.1/"
   xmlns:dcterms="http://purl.org/dc/terms/"
   xmlns:rioxxterms="http://docs.rioxx.net/schema/v3.0/rioxxterms/" xsi:schemaLocation="http://www.rioxx.net/schema/v3.0/rioxx/ http://www.rioxx.net/
schema/v3.0/rioxx/rioxx.xsd"
   xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
   <ali:free_to_read></ali:free_to_read>
   <ali:license_ref start_date="2020-11-17">https://creativecommons.org/licenses/by/4.0</ali:license_ref>
```

4 ISSUES

ali:license_ref

Record is missing the field <ali:license_ref>

Recommendation

ali:license_ref field must contain an HTTP URI that points to the license term.

1 WARNINGS

author

Missing element author

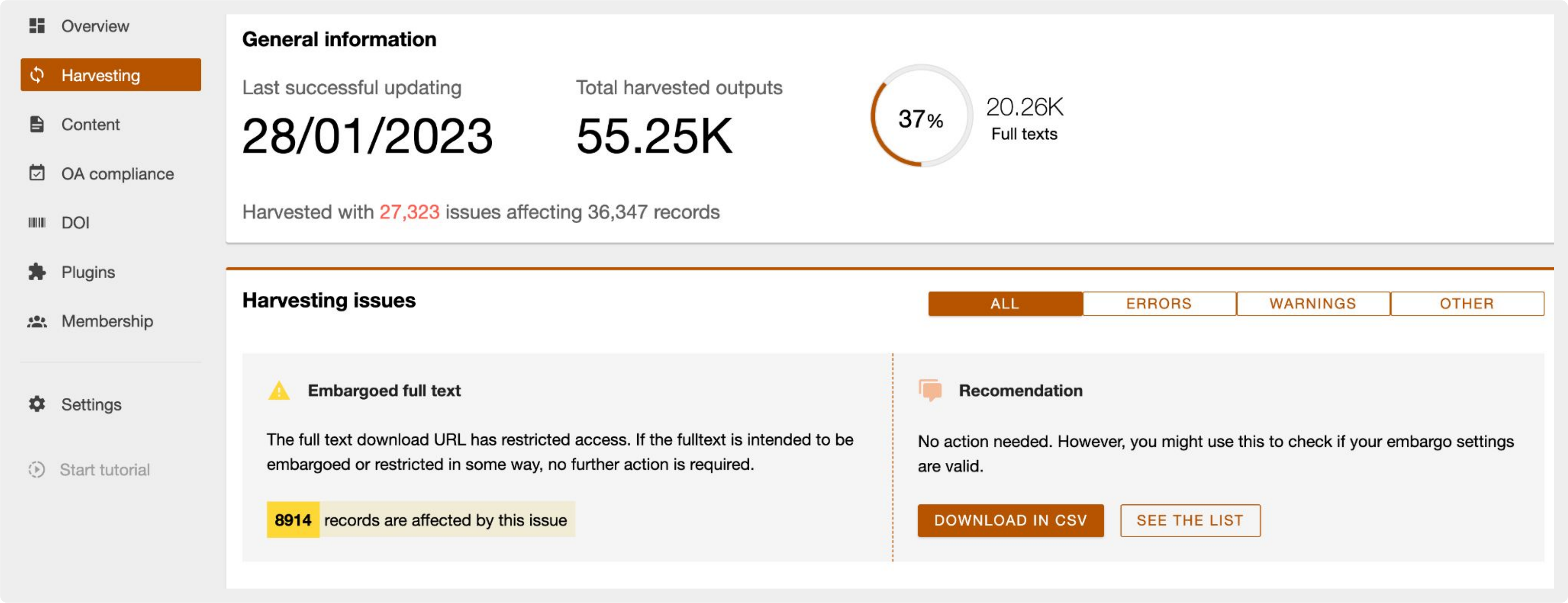
Recommendation

No recommendations yet

Validate

Validate, don't take it works for granted

Monitor: the fact that it works now doesn't mean it can't go wrong when you least expect



General information

Last successful updating

28/01/2023

Total harvested outputs

55.25K

37%

20.26K
Full texts

Harvested with 27,323 issues affecting 36,347 records

Harvesting issues

Embargoed full text

The full text download URL has restricted access. If the fulltext is intended to be embargoed or restricted in some way, no further action is required.

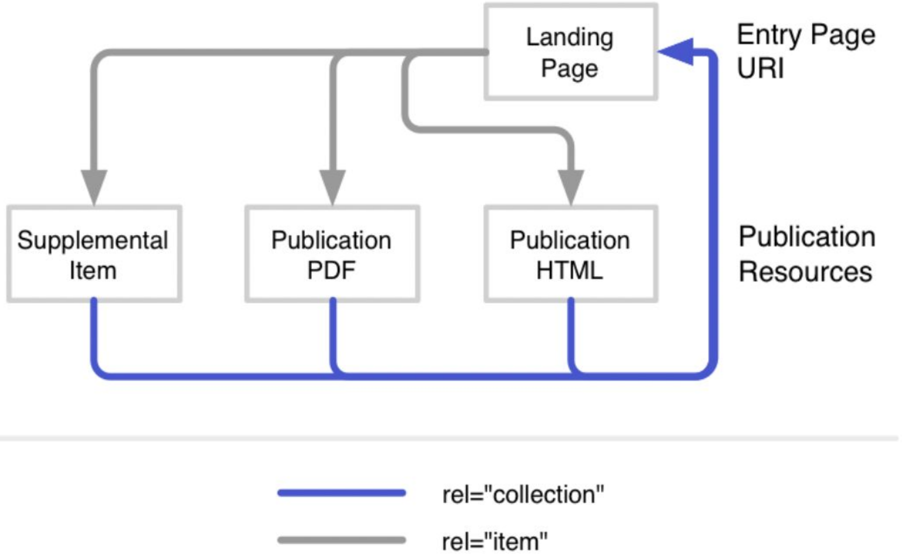
8914 records are affected by this issue

Recomendation

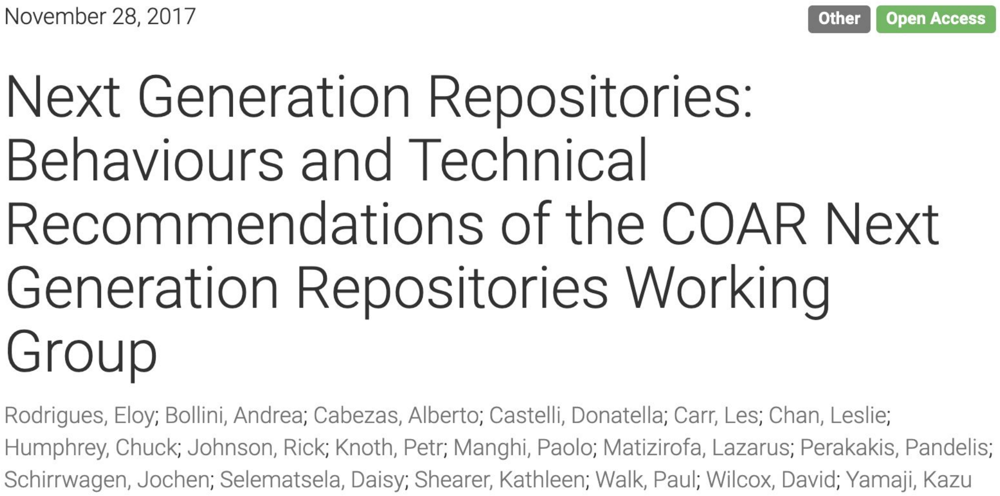
No action needed. However, you might use this to check if your embargo settings are valid.

Support Signposting

Helping machines to navigate repositories in order to locate the content.



COAR Next Generation Repositories Working Group



November 28, 2017

Other Open Access

Next Generation Repositories: Behaviours and Technical Recommendations of the COAR Next Generation Repositories Working Group

Rodrigues, Eloy; Bollini, Andrea; Cabezas, Alberto; Castelli, Donatella; Carr, Les; Chan, Leslie; Humphrey, Chuck; Johnson, Rick; Knoth, Petr; Manghi, Paolo; Matizirofa, Lazarus; Perakakis, Pandelis; Schirrwagen, Jochen; Selematsela, Daisy; Shearer, Kathleen; Walk, Paul; Wilcox, David; Yamaji, Kazu

Why is CORE important?

Increase your contents' discoverability and prevent its misuse

Search, Recommender, Discovery, PMC Linkout

Make your papers uniquely identifiable and resolvable with PIDs

OAI Resolver

Assess and contribute to Open Access compliance and FAIRness

Indexed by CORE badge

Make your content machine readable

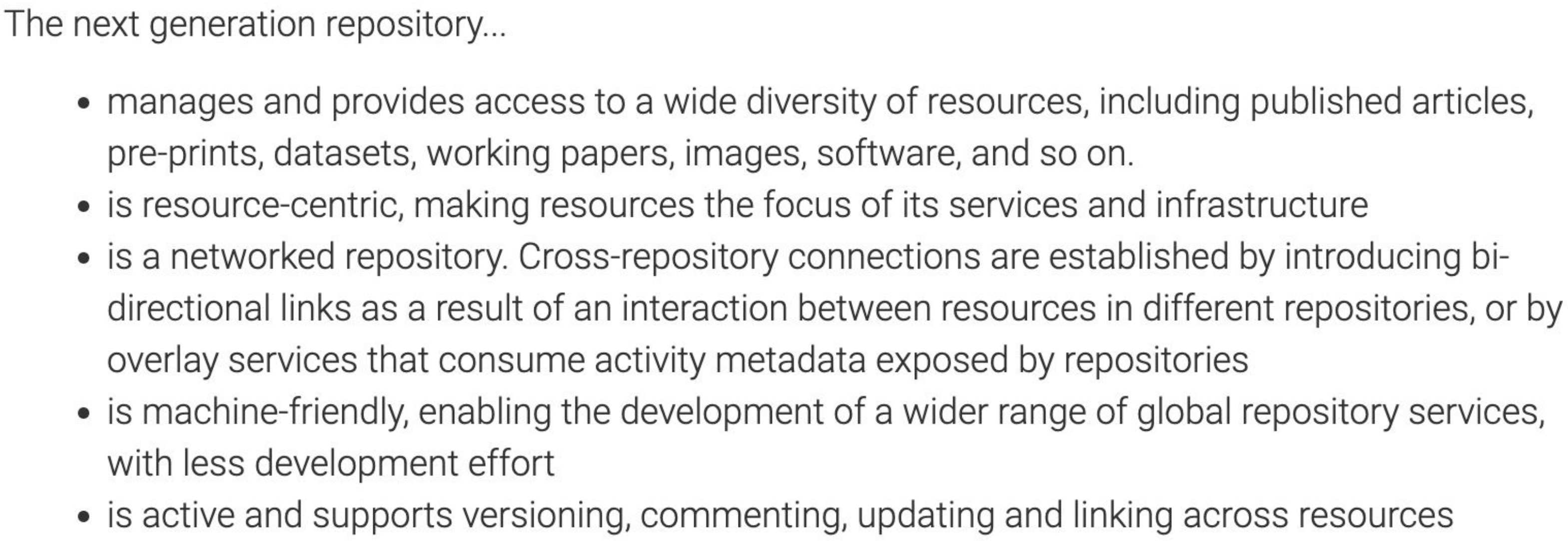
Repository Health Check, CORE API, CORE Dataset, CORE FastSYnc,

Become a CORE Member and benefit from lots more

Dashboard: Metadata validation and monitoring

>30M monthly active users

Next Generation Repositories: Behaviours



The next generation repository...

- manages and provides access to a wide diversity of resources, including published articles, pre-prints, datasets, working papers, images, software, and so on.
- is resource-centric, making resources the focus of its services and infrastructure
- is a networked repository. Cross-repository connections are established by introducing bi-directional links as a result of an interaction between resources in different repositories, or by overlay services that consume activity metadata exposed by repositories
- is machine-friendly, enabling the development of a wider range of global repository services, with less development effort
- is active and supports versioning, commenting, updating and linking across resources

CORE

AI/ML for research intelligence and for improving repository workflows

The Open University

Affiliation extraction

1. Problem

Many metadata records do not have Some text ...

Show an example how affiliations can be extracted. Show Grobid output ...

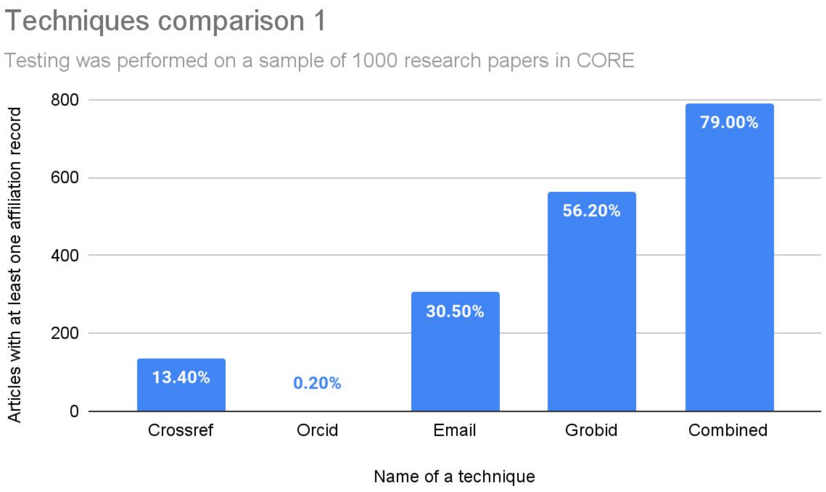
How does this correspond with ROR

This is a problem we are currently working on

2. Publication footprint

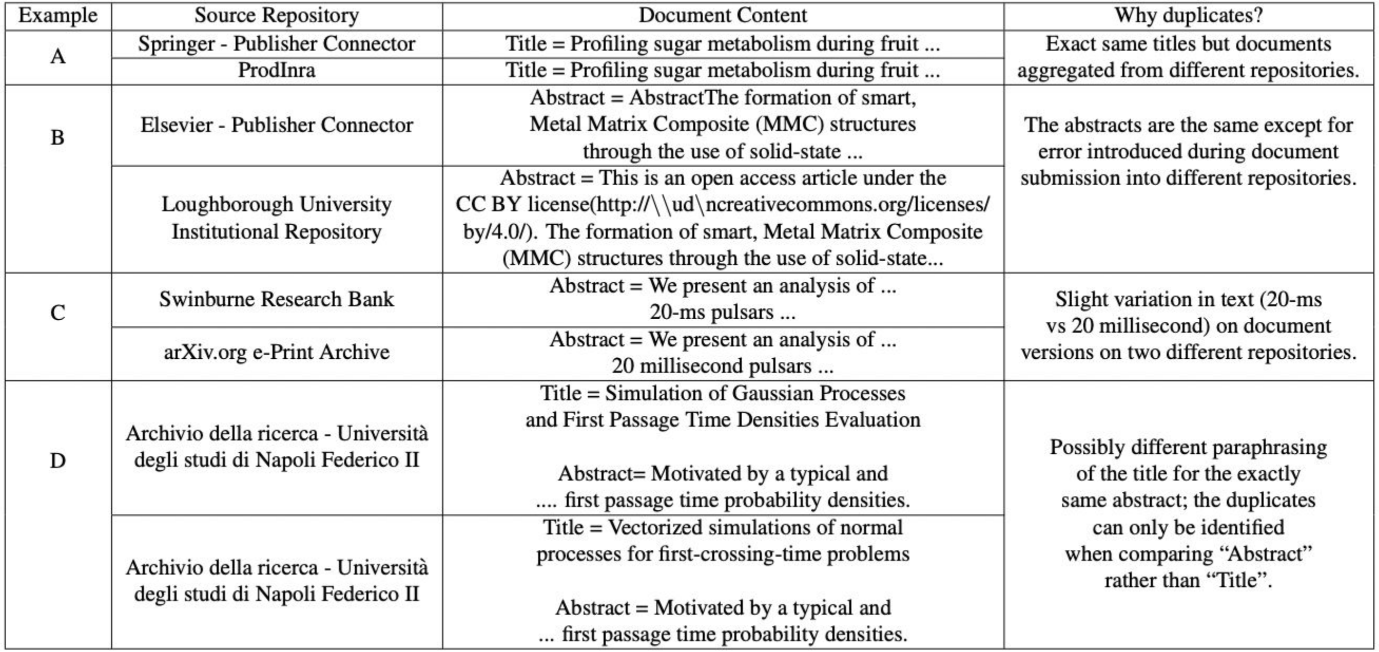
Affiliation extraction

- Many metadata records do not have affiliation data
- Affiliation is important for a range of use cases, including **publication footprint**
- At CORE, we developed a method to extract affiliation information from papers using a supervised ML model.
- Will propagate to the CORE API and Dashboard.



Deduplication

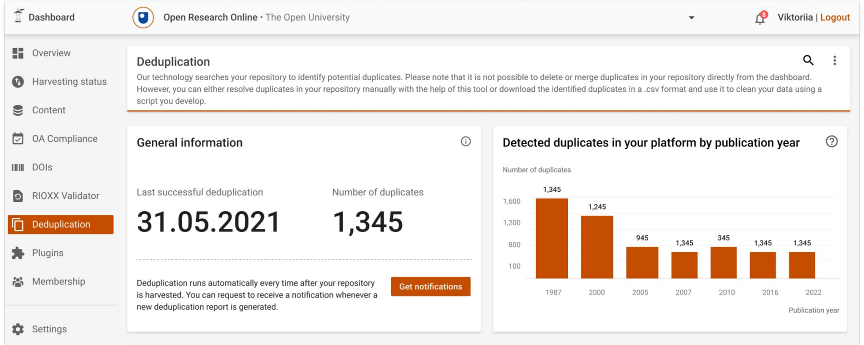
How do duplicates look like and why do they occur in repositories?



Example	Source Repository	Document Content	Why duplicates?
A	Springer - Publisher Connector	Title = Profiling sugar metabolism during fruit ...	Exact same titles but documents aggregated from different repositories.
	ProdInra	Title = Profiling sugar metabolism during fruit ...	
B	Elsevier - Publisher Connector	Abstract = AbstractThe formation of smart, Metal Matrix Composite (MMC) structures through the use of solid-state ...	The abstracts are the same except for error introduced during document submission into different repositories.
	Loughborough University Institutional Repository	Abstract = This is an open access article under the CC BY license(http://\\ud\ncreativecommons.org/licenses/by/4.0/). The formation of smart, Metal Matrix Composite (MMC) structures through the use of solid-state...	
C	Swinburne Research Bank	Abstract = We present an analysis of ... 20-ms pulsars ...	Slight variation in text (20-ms vs 20 millisecond) on document versions on two different repositories.
	arXiv.org e-Print Archive	Abstract = We present an analysis of ... 20 millisecond pulsars ...	
D	Archivio della ricerca - Università degli studi di Napoli Federico II	Title = Simulation of Gaussian Processes and First Passage Time Densities Evaluation Abstract= Motivated by a typical and .... first passage time probability densities.	Possibly different paraphrasing of the title for the exactly same abstract; the duplicates can only be identified when comparing "Abstract" rather than "Title".
	Archivio della ricerca - Università degli studi di Napoli Federico II	Title = Vectorized simulations of normal processes for first-crossing-time problems Abstract = Motivated by a typical and ... first passage time probability densities.	

Deduplication

1. CORE uses an adapted version of locality sensitive hashing (simhash) for deduplication.
2. >90% F1-score performance
3. Deduplication powers our service including in the Dashboard for:
   a. versioning
   b. OA compliance (cross-repository)
   c. with affiliation extraction, this will allow us to warn institutions before outputs become non-compliant



Deduplication

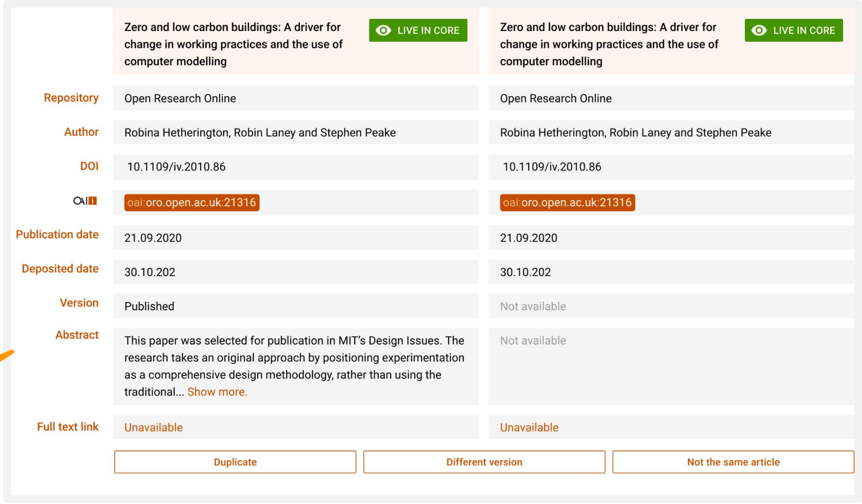
Our technology searches your repository to identify potential duplicates. Please note that it is not possible to delete or merge duplicates in your repository directly from the dashboard. However, you can either resolve duplicates in your repository manually with the help of this tool or download the identified duplicates in a .csv format and use it to clean your data using a script you develop.

General information

Last successful deduplication: 31.05.2021

Number of duplicates: 1,345

Deduplication runs automatically every time after your repository is harvested. You can request to receive a notification whenever a new deduplication report is generated.



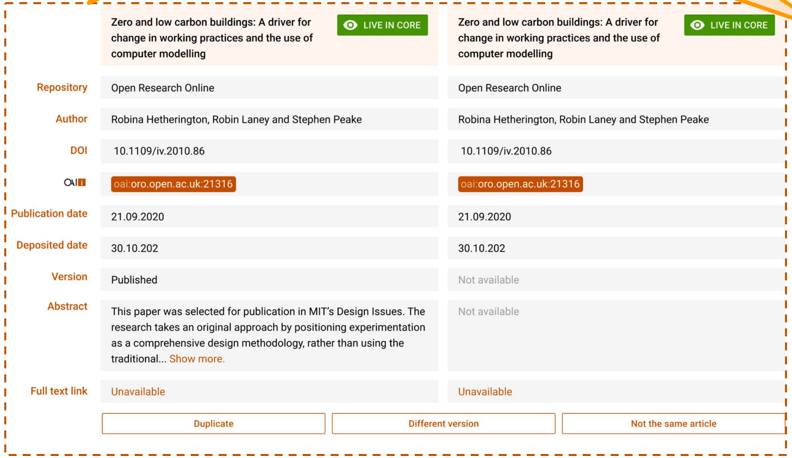
	Zero and low carbon buildings: A driver for change in working practices and the use of computer modelling	Zero and low carbon buildings: A driver for change in working practices and the use of computer modelling
Repository	Open Research Online	Open Research Online
Author	Robina Hetherington, Robin Laney and Stephen Peake	Robina Hetherington, Robin Laney and Stephen Peake
DOI	10.1109/iv.2010.86	10.1109/iv.2010.86
OAI	oai:oro.open.ac.uk:21316	oai:oro.open.ac.uk:21316
Publication date	21.09.2020	21.09.2020
Deposited date	30.10.202	30.10.202
Version	Published	Not available
Abstract	This paper was selected for publication in MIT's Design Issues. The research takes an original approach by positioning experimentation as a comprehensive design methodology, rather than using the traditional... Show more.	Not available
Full text link	Unavailable	Unavailable

Comparison mode

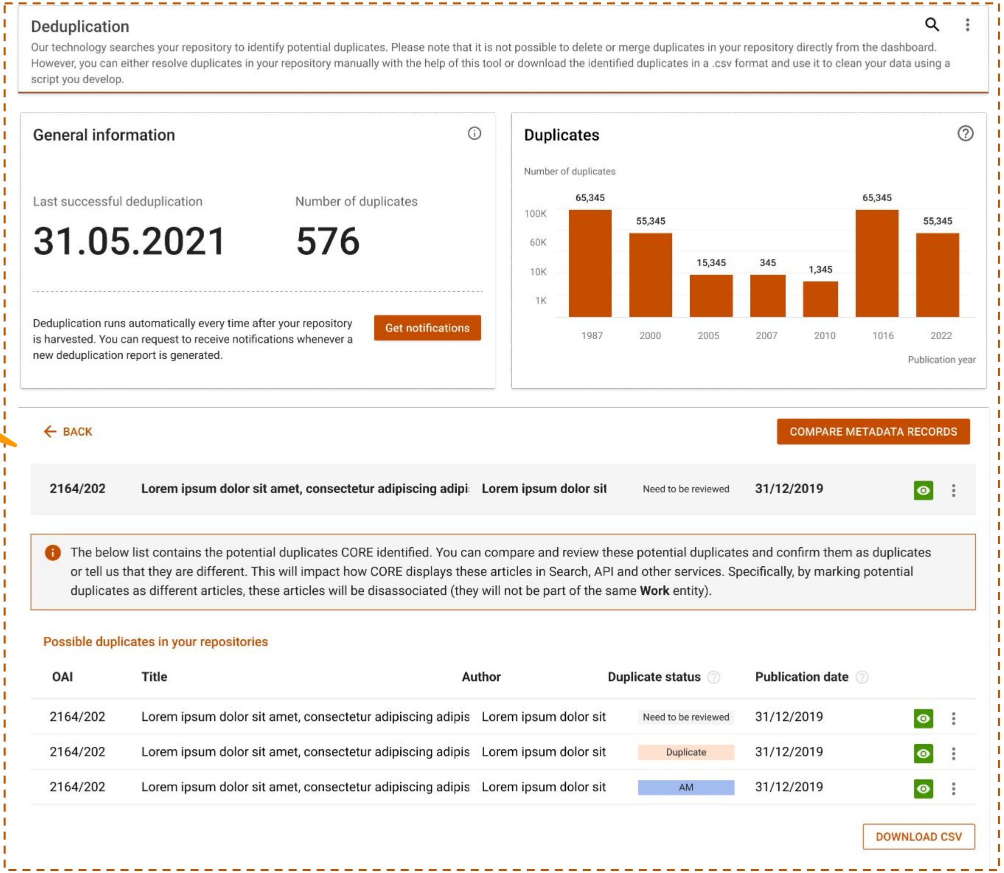
Deduplication

Comparison mode

List of possible duplicates



	Zero and low carbon buildings: A driver for change in working practices and the use of computer modelling LIVE IN CORE	Zero and low carbon buildings: A driver for change in working practices and the use of computer modelling LIVE IN CORE
Repository	Open Research Online	Open Research Online
Author	Robina Hetherington, Robin Laney and Stephen Peake	Robina Hetherington, Robin Laney and Stephen Peake
DOI	10.1109/iv.2010.86	10.1109/iv.2010.86
OAI	oai:oro.open.ac.uk:21316	oai:oro.open.ac.uk:21316
Publication date	21.09.2020	21.09.2020
Deposited date	30.10.202	30.10.202
Version	Published	Not available
Abstract	This paper was selected for publication in MIT's Design Issues. The research takes an original approach by positioning experimentation as a comprehensive design methodology, rather than using the traditional... Show more.	Not available
Full text link	Unavailable	Unavailable



Deduplication

Our technology searches your repository to identify potential duplicates. Please note that it is not possible to delete or merge duplicates in your repository directly from the dashboard. However, you can either resolve duplicates in your repository manually with the help of this tool or download the identified duplicates in a .csv format and use it to clean your data using a script you develop.

General information

Last successful deduplication

31.05.2021

Number of duplicates

576

Deduplication runs automatically every time after your repository is harvested. You can request to receive notifications whenever a new deduplication report is generated.

Duplicates

2164/202	Lorem ipsum dolor sit amet, consectetur adipiscing adipi:	Lorem ipsum dolor sit	Need to be reviewed	31/12/2019

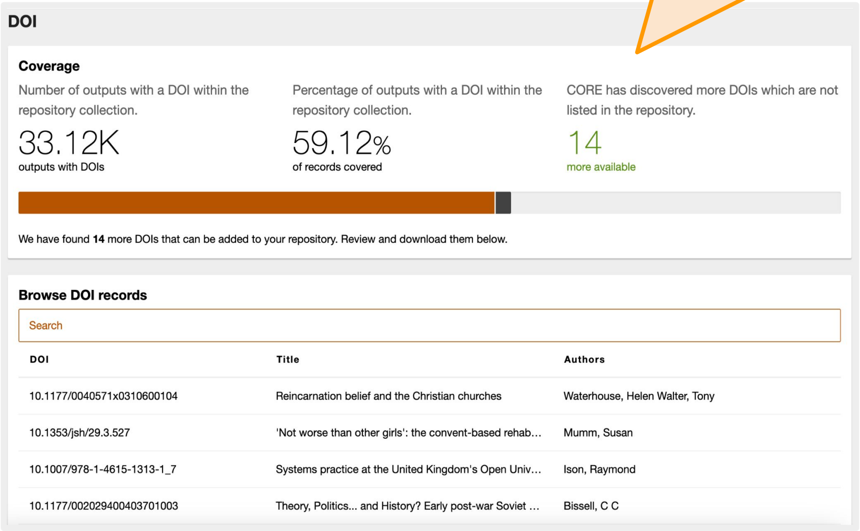
The below list contains the potential duplicates CORE identified. You can compare and review these potential duplicates and confirm them as duplicates or tell us that they are different. This will impact how CORE displays these articles in Search, API and other services. Specifically, by marking potential duplicates as different articles, these articles will be disassociated (they will not be part of the same **Work** entity).

Possible duplicates in your repositories

OAI	Title	Author	Duplicate status	Publication date
2164/202	Lorem ipsum dolor sit amet, consectetur adipiscing adipis	Lorem ipsum dolor sit	Need to be reviewed	31/12/2019
2164/202	Lorem ipsum dolor sit amet, consectetur adipiscing adipis	Lorem ipsum dolor sit	Duplicate	31/12/2019
2164/202	Lorem ipsum dolor sit amet, consectetur adipiscing adipis	Lorem ipsum dolor sit	AM	31/12/2019

Data enrichment

You can enrich data with **DOIs** identified in other repositories



DOI

Coverage

Number of outputs with a DOI within the repository collection.

33.12K

outputs with DOIs

Percentage of outputs with a DOI within the repository collection.

59.12%

of records covered

CORE has discovered more DOIs which are not listed in the repository.

14

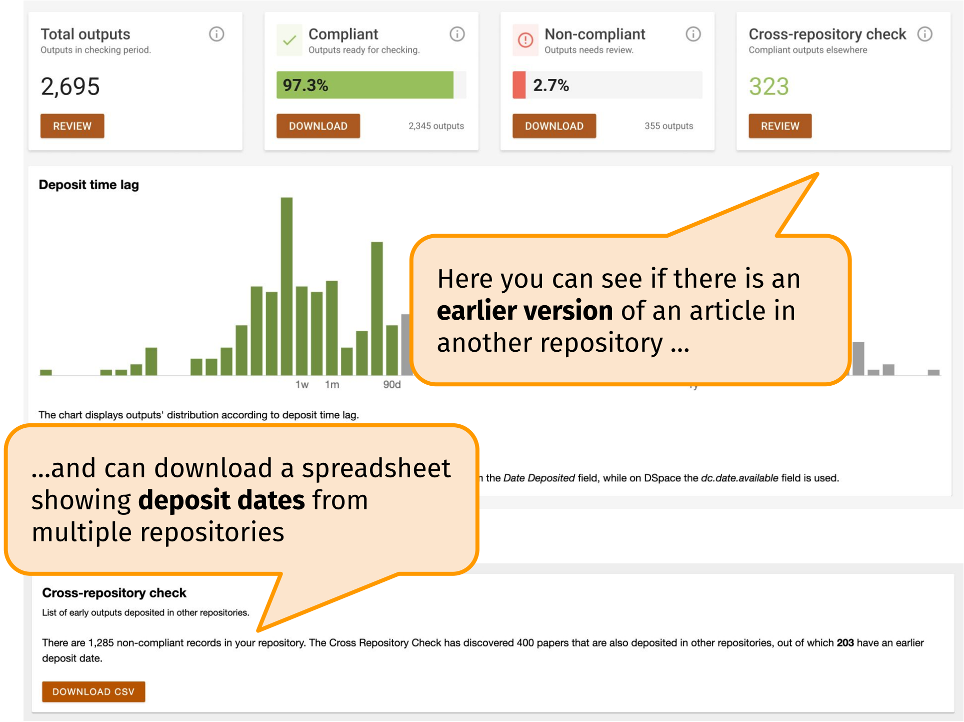
more available

We have found **14** more DOIs that can be added to your repository. Review and download them below.

Browse DOI records

Search

DOI	Title	Authors
10.1177/0040571x0310600104	Reincarnation belief and the Christian churches	Waterhouse, Helen Walter, Tony
10.1353/jsh/29.3.527	'Not worse than other girls': the convent-based rehab…	Mumm, Susan
10.1007/978-1-4615-1313-1_7	Systems practice at the United Kingdom's Open Univ…	Ison, Raymond
10.1177/002029400403701003	Theory, Politics... and History? Early post-war Soviet …	Bissell, C C



Total outputs	Compliant	Non-compliant	Cross-repository check
Outputs in checking period.	Outputs ready for checking.	Outputs needs review.	Compliant outputs elsewhere
2,695	97.3%	2.7%	323
REVIEW	DOWNLOAD 2,345 outputs	DOWNLOAD 355 outputs	REVIEW

Deposit time lag

The chart displays outputs' distribution according to deposit time lag.

n the *Date Deposited* field, while on DSpace the *dc.date.available* field is used.

Here you can see if there is an **earlier version** of an article in another repository ...

...and can download a spreadsheet showing **deposit dates** from multiple repositories

Cross-repository check

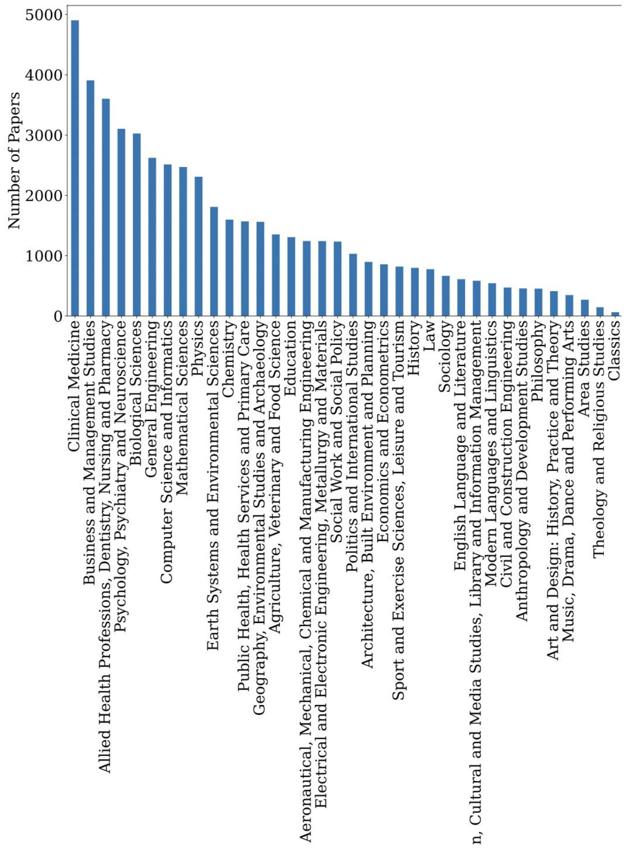
List of early outputs deposited in other repositories.

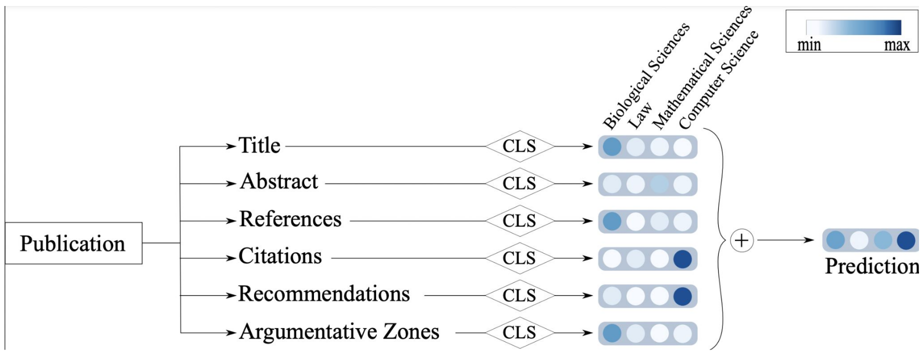
There are 1,285 non-compliant records in your repository. The Cross Repository Check has discovered 400 papers that are also deposited in other repositories, out of which **203** have an earlier deposit date.

DOWNLOAD CSV

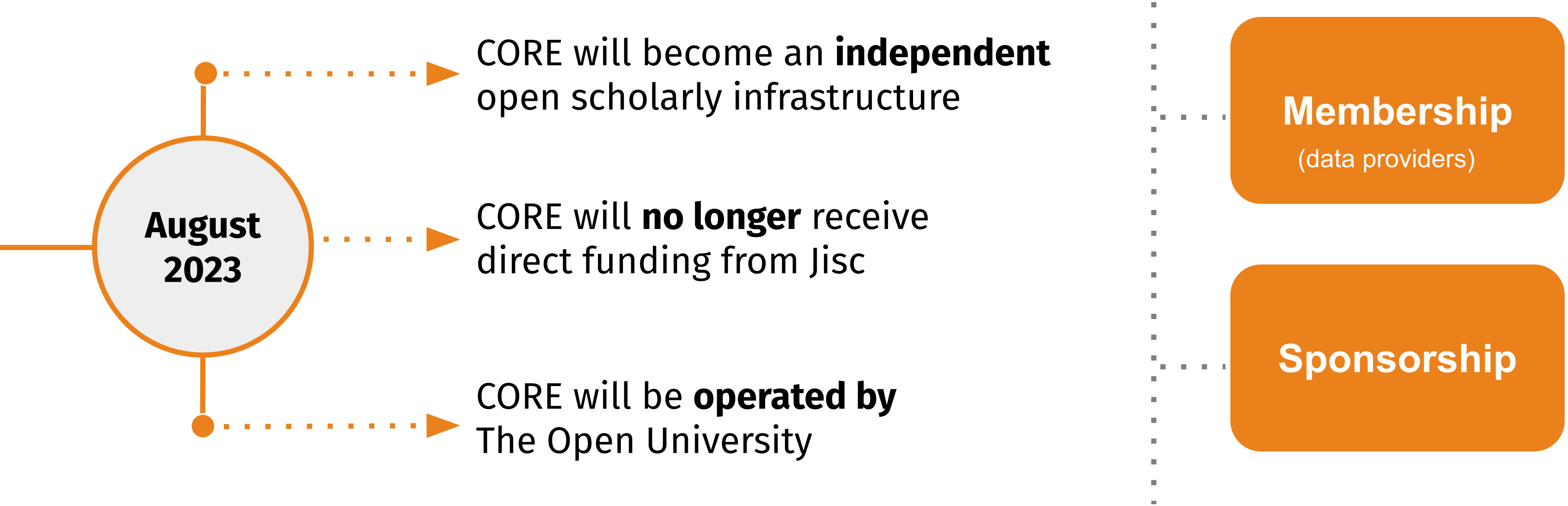
Document classification

- Classification of research papers in a distributed environment is a problem.
- Established a benchmark for research document classification as part of the SDP/COLING conference.
- In the process of bringing themes to the CORE API.





CORE moving to a membership model



August 2023

CORE will become an **independent** open scholarly infrastructure

CORE will **no longer** receive direct funding from Jisc

CORE will be **operated by** The Open University

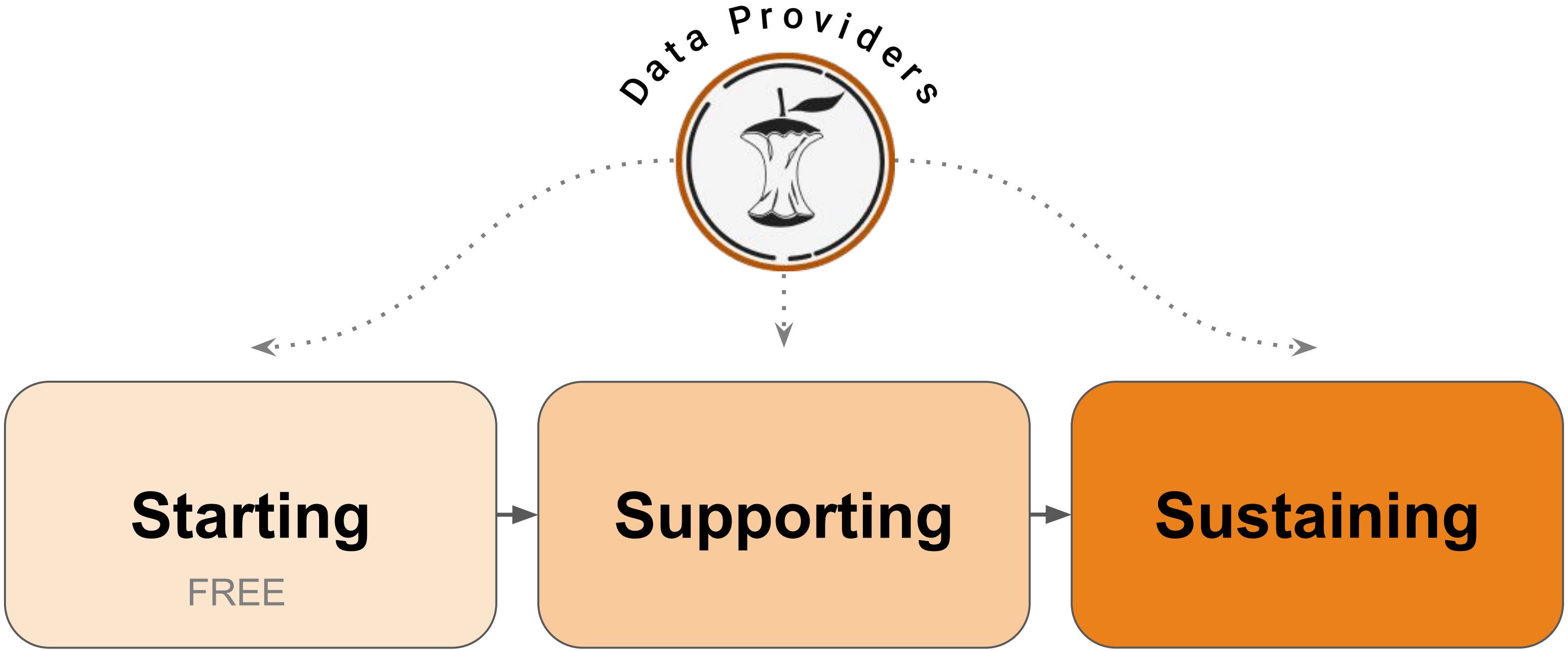
Membership
(data providers)

Sponsorship

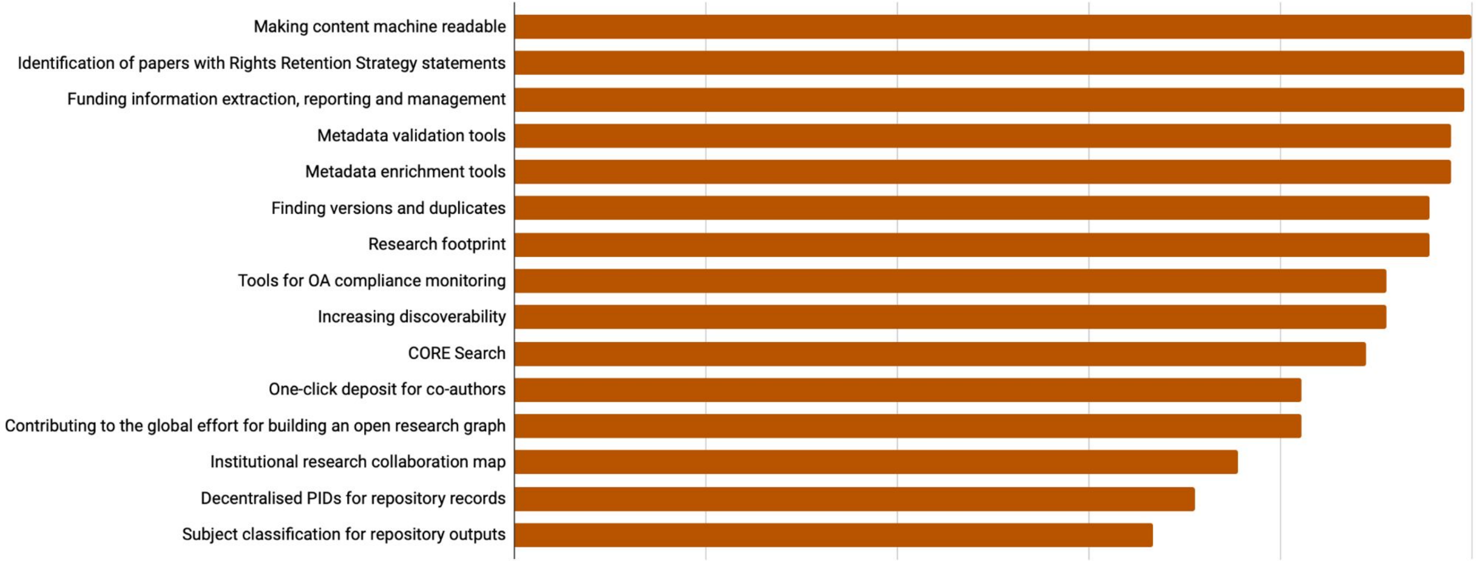
CORE Membership

- A network of data providers who are committed to the ongoing success of the **Open Access movement**
- We provide **tools and benefits** for our members
- All CORE data providers are eligible to become CORE Starting Members **free** of charge
- Supporting and Sustaining Members:
    - help shape our development roadmap
    - support and sustain CORE

Three levels of CORE Membership



What matters to members (Board of Supporters survey)



More reading: references

Knoth, P. (2013). **From open access metadata to open access content: two principles for increased visibility of open access content**. In Open Repositories 2013. Retrieved from http://oro.open.ac.uk/37824/

Pride, D., & Knoth, P. (2020). **An Authoritative Approach to Citation Classification.** Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020. doi:10.1145/3383583.3398617

Kunnath, Suchetha N.; Pride, David; Gyawali, Bikash and Knoth, Petr (2020). **Overview of the 2020 WOSP 3C Citation Context Classification Task**. In: Proceedings of the 8th International Workshop on Mining Scientific Publications, Association for Computational Linguistics pp. 75–83.

Kunnath, Suchetha N.; Herrmannova, Drahomira; Pride, David; Knoth, Petr (2022). **A Meta-analysis of Semantic Classification of Citations** . Quantitative Science Studies, 2 (4), pp. 1170-1215

More reading: references

Kusa, Wojciech; Hanbury, Allan; Knoth, Petr (2022). **Automation of Citation Screening for Systematic Literature Reviews using Neural Networks: A Replicability Study** . In: 44th European Conference on Information Retrieval, 10-14 Apr 2022, Stavanger, Norway Springer , 13185 , pp. 584-598

Nambanoor Kunnath, Suchetha; Pride, David; Knoth, Petr (2022). **Dynamic Context Extraction for Citation Classification**. In: The 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing, 20-23 Nov 2022, Virtual

Gyawali, Bikash; Anastasiou, Lucas; Knoth, Petr (2020). **Deduplication of Scholarly Documents using Locality Sensitive Hashing and Word Embeddings**. In: 12th Language Resources and Evaluation Conference, 11-16 May 2020, Marseille, France European Language Resources Association , pp. 894-903

More reading: references

Óscar E. Mendoza, Wojciech Kusa, Alaa El-Ebshihy, Ronin Wu, David Pride, Petr Knoth, Drahomira Herrmannova, Florina Piroi, Gabriella Pasi, and Allan Hanbury. 2022. **Benchmark for Research Theme Classification of Scholarly Documents.** In Proceedings of the Third Workshop on Scholarly Document Processing, pages 253–262, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Pride, David; Harag, Jozef; Knoth, Petr (2019). **ACT: An Annotation Platform for Citation Typing at Scale**. In: JCDL 2019 - ACM/IEEE-CS Joint Conference on Digital Libraries 2019, 2-6 Jun 2019, Urbana-Champaign, Illinois

Herrmannova, Drahomira; Pontika, Nancy; 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 , pp. 206-216 BEST PAPER AWARD

CORE

2023

Take home ...

- **ML/AI** has the potential to transform all stages of the research process, including how we carry out research, how we assess it and how we organise research knowledge.
- **OA** repositories play a key role in this process by providing machine access to research content.
- **AI/ML** already provides opportunities for improving the ways we use repositories, organise, enrich and curate content in them.

CORE

2023

THANK YOU

The Open University