



AI for the Open Scholarly Web: Opportunities and Challenges.

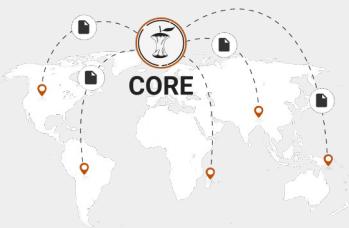
Prof. Petr Knoth



Outline

1. Introducing COnnecting REpositories (CORE)
2. 3 examples of how can **AI / ML transform research workflows**
3. Challenges in gathering research content and the need for **machine access** to this content.
4. **AI Needs Open Datasets and Open Infrastructures**





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

- **>30 Million** monthly active users (MAU)
- **34 Million** full-text research papers hosted by CORE.
- **290 Million** metadata records

Signatory of Principles of Open Scholarly Infrastructure (POSI)

Commercial Partners



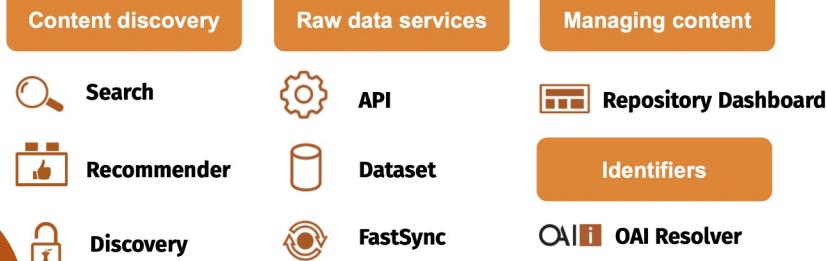
- Innovation and trends analysis
- Plagiarism detection
- Fact checking
- Finance
- Health

Institutional Members



32 supporting or sustaining members

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



Providing
seamless access
to open research
for humans and
machines.

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

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

one of the world's **most used** collections 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.



How can AI/ML transform research workflows

AI for credible trustworthy question answering (CORE-GPT)

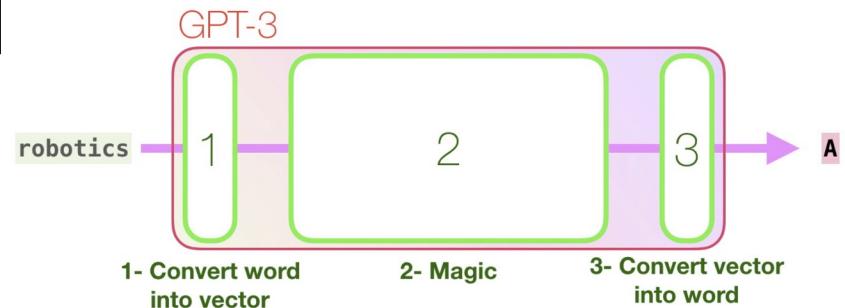
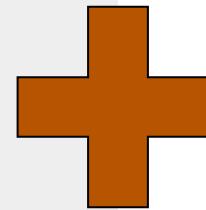
CORE is the world's largest collection of Open Access papers, collating and enriching content from over **12,463** data providers.



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

GPT large language models*

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



*Other large language models are available

@JayAlammar

CORE-GPT Results

What is the effect of changing the composition of metal alloys on their mechanical properties?

[Ask again](#)

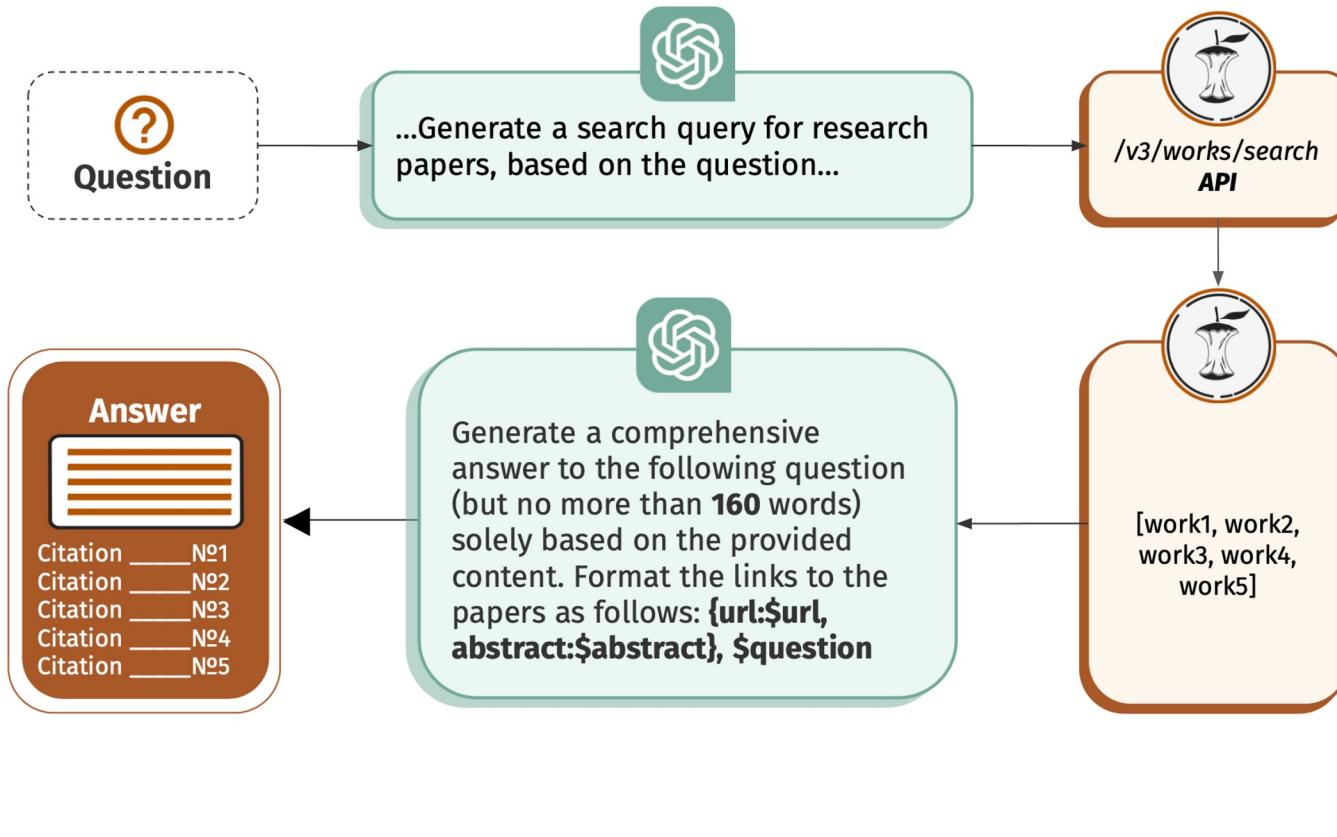
Changing the composition of metal alloys has significant effects on their mechanical properties. For instance, in the 15-5 PH alloy, exposure to intermediate temperatures (300-400°C) can cause embrittlement through the martensitic decomposition, resulting in a drastic drop of the impact strength-energy and an increase in ductile-to-brittle transition temperature [1]. The mechanical properties of Al-Fe-Si alloy are influenced by iron content and plastic deformation, with ultimate tensile strength increasing when iron content is between 0.708% and 0.934% [2]. An innovative AlSi3Cr alloy showed improved tensile strength and ductility when Chromium and Manganese were added [3]. In Aluminium-based alloys, liquid stir casting technique can be used to investigate and improve the structural properties by modifying the alloying elements [4]. Similarly, adding Cr, Nb, and V to TiAlN coatings enhances thermal stability and fracture toughness [5].

1. <https://core.ac.uk/works/5187698> - Evolution of microstructure and impact-strength energy in thermally and thermomechanically aged 15-5 PH
2. <https://core.ac.uk/works/138558590> - INFLUENCE OF IRON CONTENT AND PLASTIC DEFORMATION ON THE MECHANICAL PROPERTIES OF 8011-TYPE Al-Fe-Si ALLOY
3. <https://core.ac.uk/works/50127668> - Investigation of mechanical properties of AlSi3Cr alloy
4. <https://core.ac.uk/works/10956170> - Effect of Alloying Element on the Integrity and Functionality of Aluminium-Based Alloy
5. <https://core.ac.uk/works/51413595> - Enhanced thermal stability and fracture toughness of TiAlN coatings by Cr, Nb and V-alloying

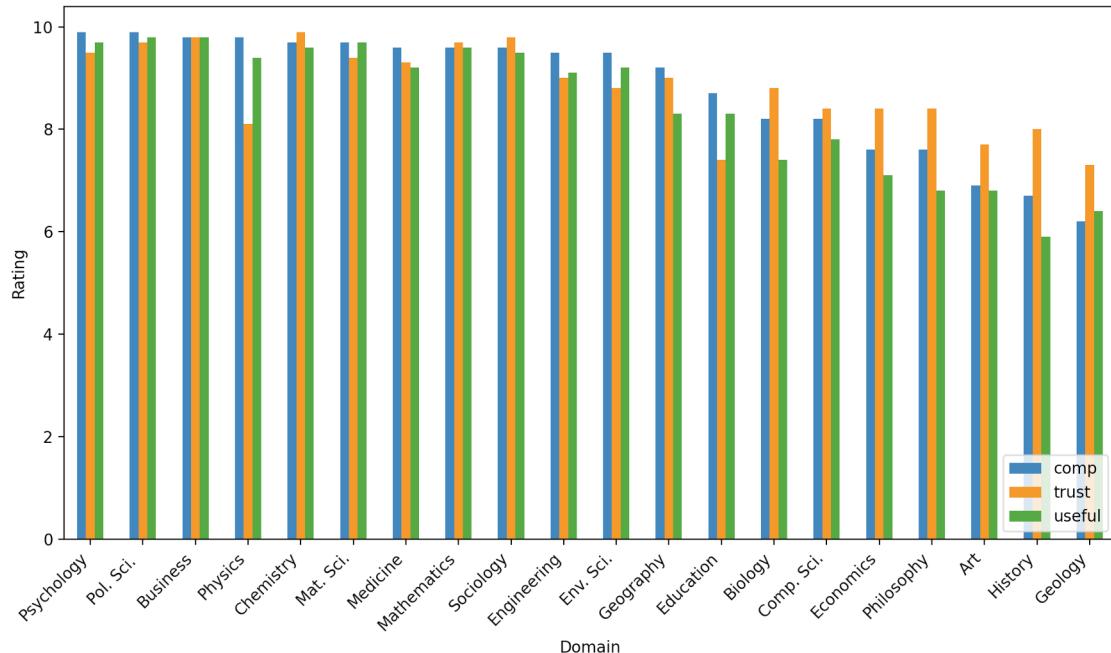
[See more in CORE](#)



CORE-GPT Workflow



CORE-GPT Evaluation



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

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

Pride, David, Petr Knoth, and Jozef Harag. "Act: An annotation platform for citation typing at scale." *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE, 2019.

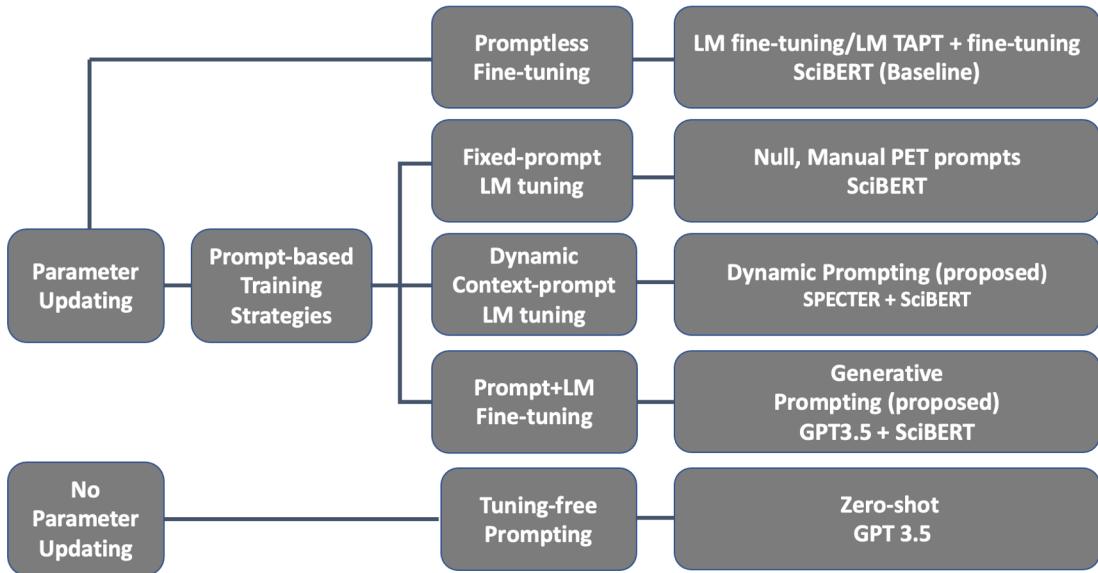
| Citation Function | Examples |
|--------------------|--|
| BACKGROUND | Most of the participatory models to design educational games are founded on educational theories and game design (see for example: Amory, 2007; #CITATION_TAG). |
| COMPARES_CONTRASTS | Similar observations have been made in the 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]. |

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 bench-mark.

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



AI for systematic reviews

→ Systematic reviews

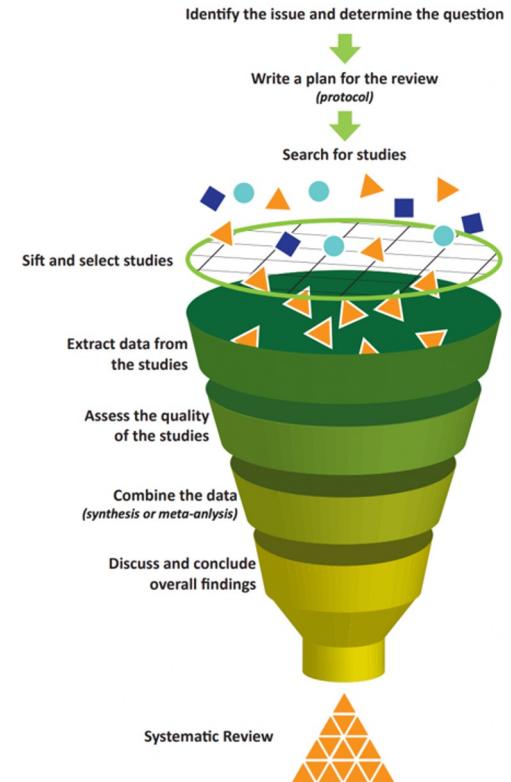
- ◆ Time consuming

→ 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



AI for systematic reviews

→ Involves many steps

→ Some of the most-time consuming can be automated

Kusa, W., Lipani, A., Knoth, P., & Hanbury, A. (2023). An analysis of work saved over sampling in the evaluation of automated citation screening in systematic literature reviews. *Intelligent Systems with Applications*, 18, 200193.

| 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 |



Gathering scholarly content and the challenges of it

COAR Manifesto

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

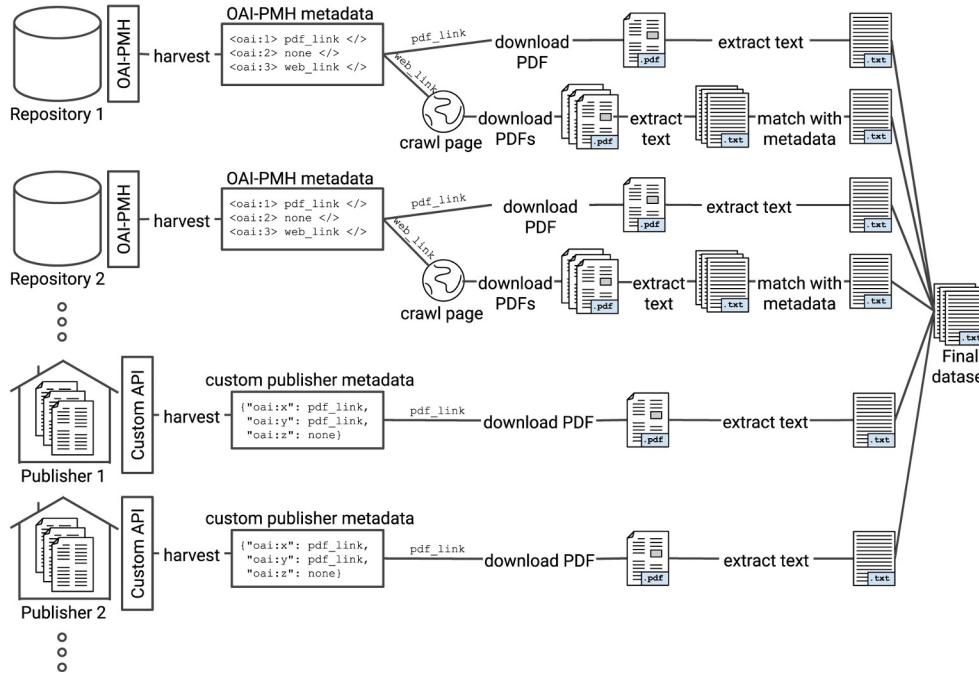


Aggregating research literature is no small task

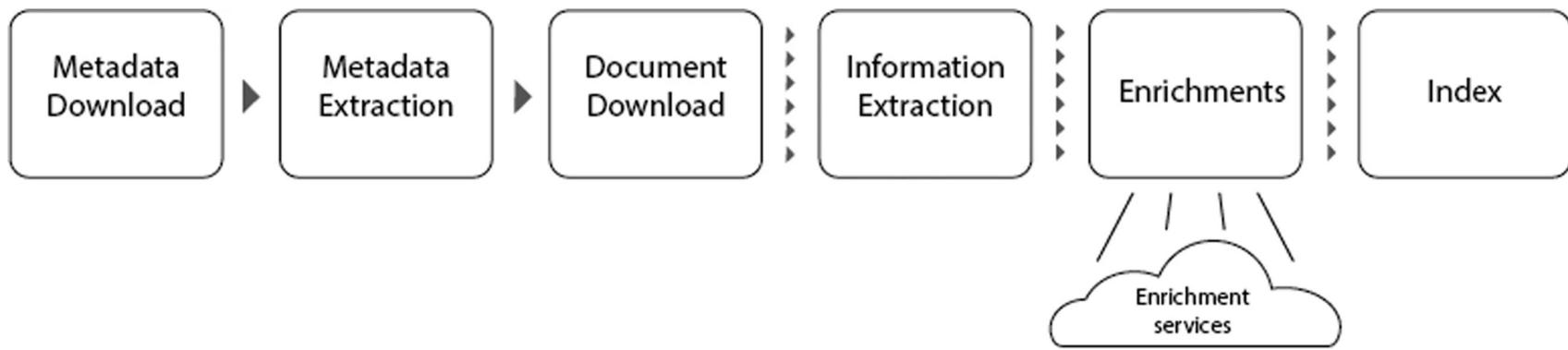
- Closed content
- Interoperability issues
- Widely distributed network
- Not just metadata - fulltext



Minimum steps for content aggregation from repositories



CORE ingestion pipeline



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**

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

Recommendation

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

ALL **ERRORS** **WARNINGS** **OTHER**

DOWNLOAD IN CSV **SEE THE LIST**

Validate metadata

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

The screenshot shows the CORE Repository Dashboard with validation results. A dashed orange box highlights the 'WARNINGS' and 'ISSUES' sections.

WARNINGS (1):

- author**: Missing element author

ISSUES (4):

- 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.

The screenshot shows the RIOXX metadata validator service. A dashed orange box highlights the left sidebar and the main content area.

Dashboard (selected)

RIOXX metadata validator

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

INFO: 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.

My repository | **RIOXX validator**

Validate a metadata record

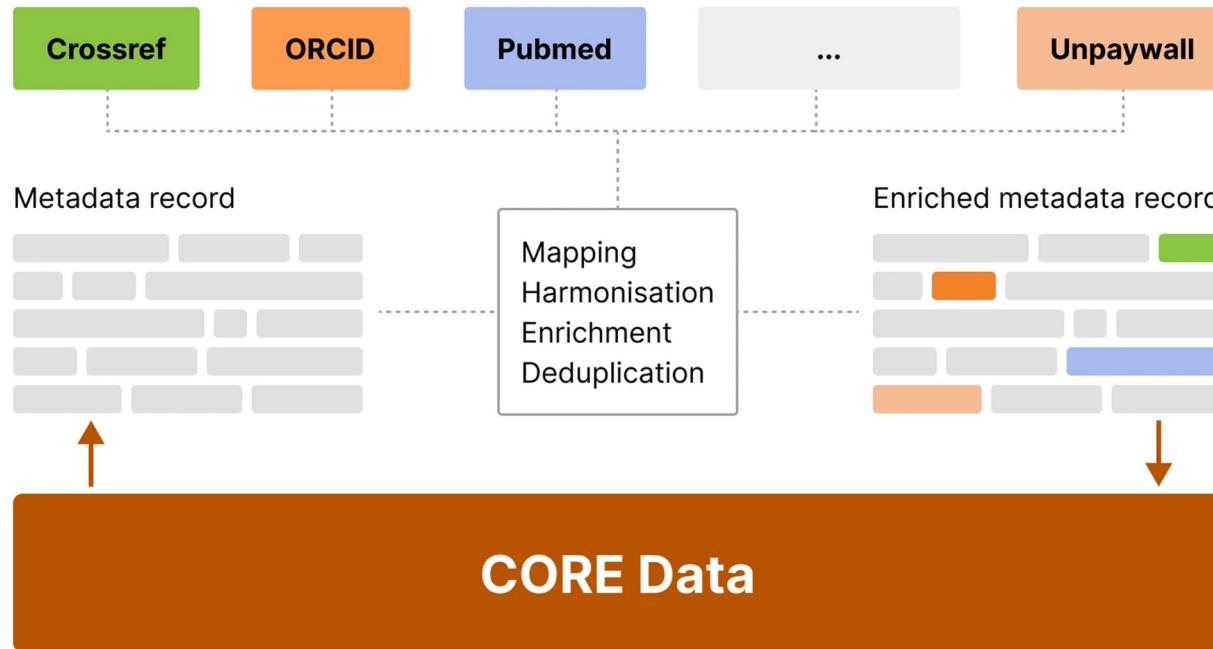
You can input a record XML (what is enclosed in the <rioxx>...</rioxx> tags).

Example 1 (fully compliant) | **Example 2 (partially compliant)**

```
<rioxx
  xmlns="http://www.rioxx.net/schema/v3.0/rioxx/"
  xmlns:ali="http://ali.riox.org/2014/ali/1.0"
  xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:dcterms="http://purl.org/dc/terms/"
  xmlns:rioxxtterms="http://docs.rioxx.net/schema/v3.0/rioxxtterms/" 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:free_to_read start_date="2020-11-17">https://creativecommons.org/licenses/by/4.0</ali:free_to_read>
</rioxx>
```

VALIDATE

Ingestion pipeline: Enrichments



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 |
| 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 | 30.10.2022 | 30.10.2022 |
| 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 |
| | Duplicate | Different version |
| | | Not the same article |

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

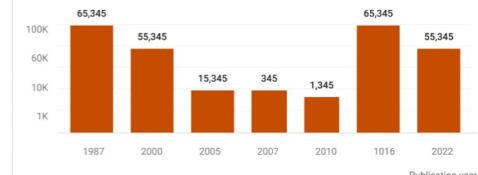
31.05.2021

Number of duplicates

576

Duplicates

Number of duplicates



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

[Get notifications](#)

[BACK](#)

[COMPARE METADATA RECORDS](#)

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

[...](#)

① 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 | ... |

[DOWNLOAD CSV](#)

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

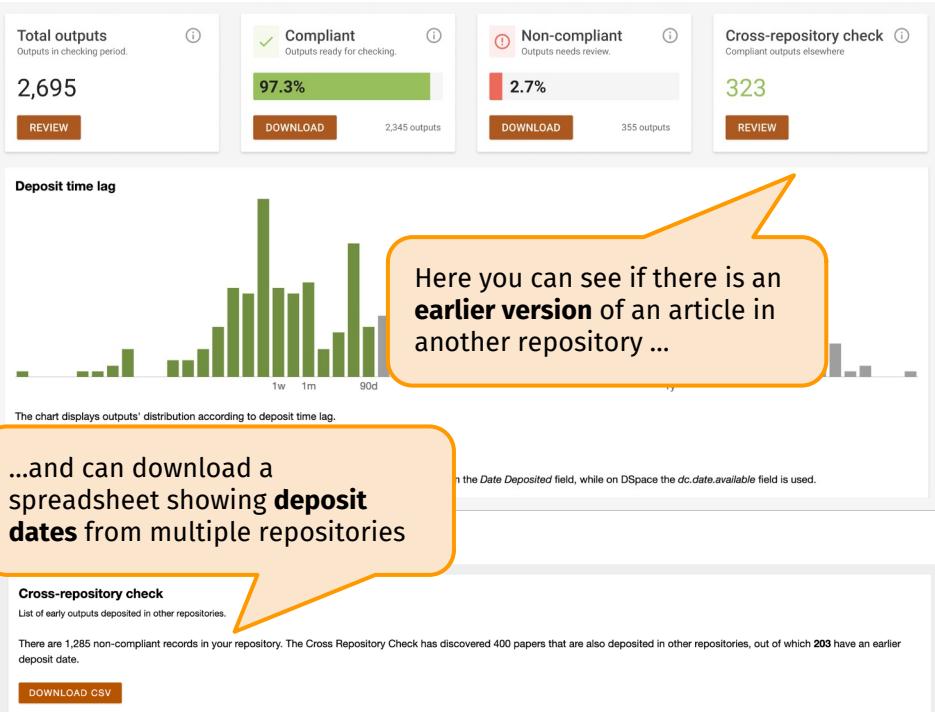
CORE has discovered more DOIs which are not listed in the repository.
14
more available

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

Browse DOI records

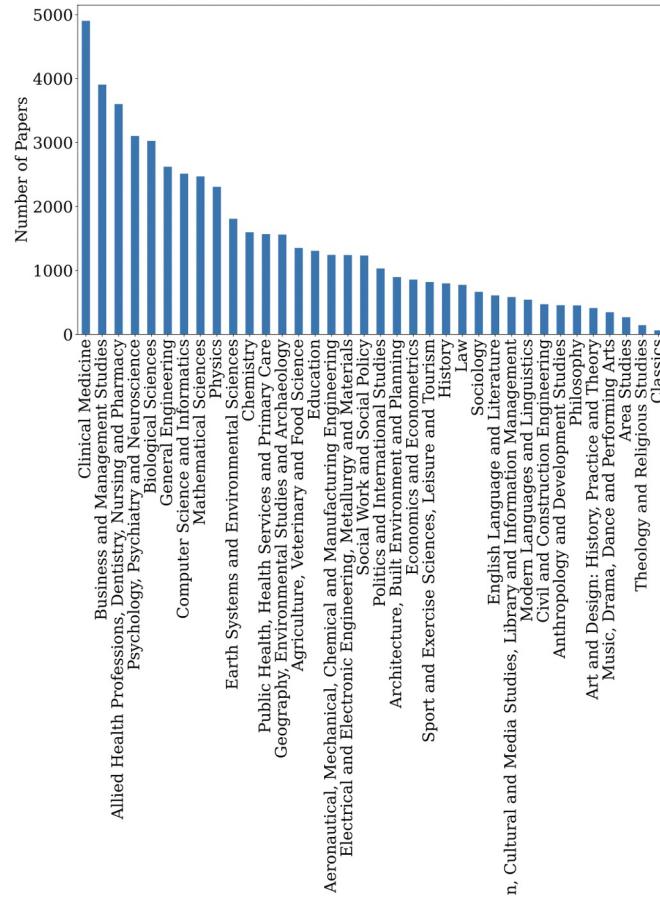
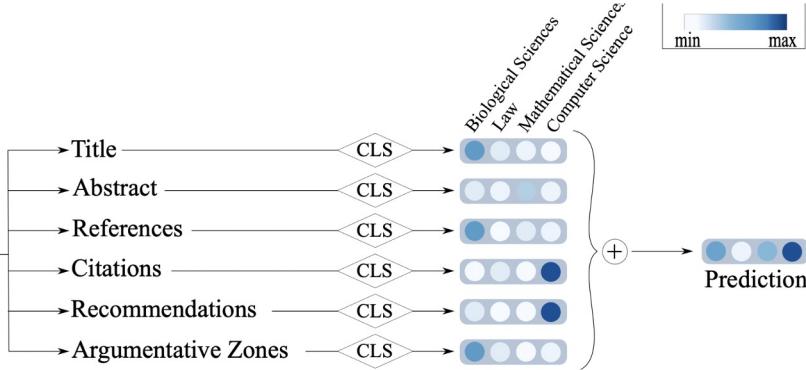
Search

| 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 |



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.

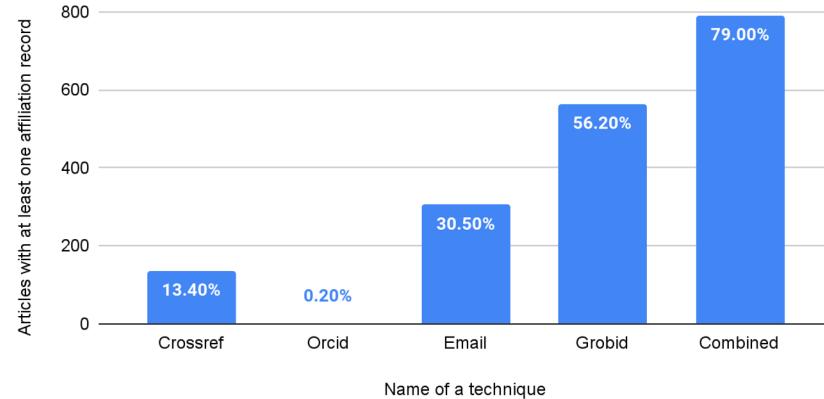


Affiliation extraction

- 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.
- Harmonise and map affiliation to a ROR ID
- Will propagate to the CORE API and Dashboard.

Techniques comparison 1

Testing was performed on a sample of 1000 research papers in CORE



Using AI with the corpus - and to curate the corpus

Question answering - CORE-GPT

Citation classification

Systematic Reviews

Supporting Peer Review

+ many more

Metadata Enrichment

Deduplication

Affiliation Extraction

Rights Retention Statements

Fresh Finds



AI Needs Open Data and Open Infrastructures

The power of data vs the power of models

- To increase performance, shall I spend more time on tuning models or on improving my data?
- In most situations, spending time on data leads to higher performance improvements.
- But, it takes considerable time, effort and resources needed to collect such data.

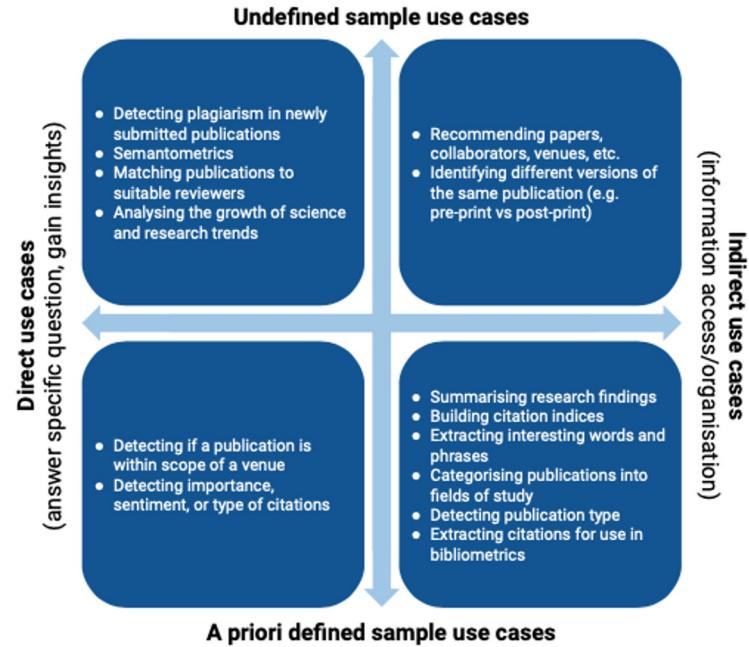


Wide variety of use cases over research literature

- A limited number of use cases can be satisfied with a sample of scholarly content (slice and dice the data as needed)
- Many use cases require machine access to all existing research from everywhere and always up-to-date.

=>

AI needs access to an open, comprehensive, always up-to-date corpus of research content at a full text level.



CORE provides access to machine-readable research



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
 - A wide range of application areas: plagiarism detection, fact checking and misinformation detection, research graphs, rapid systematic reviews, detection of research trends, research assessment, identifying suitable peer-reviewers, recommender systems, innovation engineering, compliance monitoring, drug discovery, ...



Dataset



FastSync

Opportunities of AI for academic research

Facilitating and systematising literature review

Hypothesis generation (literature-based discovery)

Generating experimental code

Automating experimentation process

Data interpretation?

Drafting and/or improving the quality of manuscripts

Supporting peer-review

Improving research assessment

Dissemination of research

...

Challenges of AI for academic research

Algorithmic Bias

Model interpretability

Validation

Computational resources

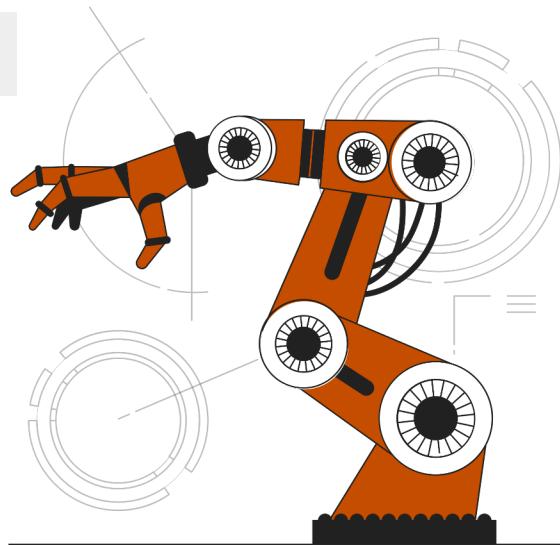
Reproducibility

Legality

Ethics

Credit assignment

Integration with existing systems



Many difficult questions to answer

- Is it ethical to use AI to evaluate research?
- How shall we assign credit for research that was created with AI assistance?
- Is it acceptable to allow academics to use generative AI in drafting manuscripts and research proposals?



Take home ...

- **Scholarly data** in a machine readable form are highly valuable in a wide range of application both within and outside of academia.
- **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.
- CORE provides access to a **comprehensive always up-to-date corpus of research papers**. Come and work with us to tackle any AI challenges and opportunities that need this data.

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.

More reading: references

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

Thelwall, M., Kousha, K., Wilson, P., Makita, M., Abdoli, M., Stuart, E., Levitt, J., Knoth, P. and Cancellieri, M. (2023) **Predicting article quality scores with machine learning: The UK Research Excellence Framework**, *Quantitative Science Studies*, pp. (early access), MIT Press

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

More reading: references

Suchetha N. Kunnath, David Pride, and Petr Knoth. 2023. **Prompting Strategies for Citation Classification.** In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23)*, October 21–25, 2023, Birmingham, United Kingdom. ACM, New York, NY, USA, 11 pages

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

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*