Innovations on the horizon for an OA repository indexing system

Petr Knoth
petr.knoth@open.ac.uk
- **Part 1: Intro**
  - Usual COAR slide which talks about the need to pass data to SW etc. (existing slide)
  - One slide to mention that repositories should not block machine agents from indexing their content (Petr can give more info)
  - Introduction to CORE (usual existing slide)
    - Usual slide
    - A slide showing CORE establishing itself in the following key countries (maps of the world with a logo of CORE in the respective places):
      - UK
      - North America (apart from Mexico => US & Canada)
      - South Africa

- **Part 2: Innovations**
  - Deduplication slides
    - Schema of how we do it (perhaps from Matteo's dedup presentation)
    - One slide on the difference between works and outputs (perhaps from Matteo's dedup presentation)
    - Screenshot of deduplication module in Dashboard
  - 1 slide on how our affiliation extraction works (Suchetha to create)
  - Fresh Finds: screenshot with quick description of functionality
  - RRS tool (from Matteo's presentation earlier in the week)
  - SDG prediction
    - One slide on what exists already: e.g. existing solution, problem description, any intermediate results (Suchetha to provide)
    - One slide on the SDG process (existing slide)
    - One slide with the screenshot of the module we are developing (from Figma)
  - TEI dataset: supporting TDM part of funder requirements, such as Nelson Memo. Raw a link between getting indexed in CORE and making content available for TDM
    - Extending this with SoFAIR for Software Citations (add existing SoFAIR slide)
  - AI: leveraging local content for RAG (CORE-GPT but with boosting content from a given repository - @David to supervise)
  - Local vs global context and OA compliance:
    - FAIR certification slide(s)
    - USRN report slide

- **Part 3: Wrap up:**
  - Show CORE subscribing to POSI (existing slide)
  - Show CORE Community governance (existing slide)
  - CORE Membership slide (existing slide)
  - Conclusion slide (leave empty, I will edit)
Each individual repository is of limited value for research: **the real power of Open Access lies in the possibility of connecting and tying together repositories**, which is why we need interoperability. In order to create a seamless layer of content through connected repositories from around the world, Open Access relies on interoperability, the ability for systems to communicate with each other and pass information back and forth in a usable format. **Interoperability allows us to exploit today’s computational power so that we can aggregate, data mine, create new tools and services, and generate new knowledge from repository content.**
Access for machine agents

Repository configuration

A robots.txt file provides a commonly used mechanism with which the webmaster can limit access to a website for machine agents, such as crawlers, and control the load on the server.

The CORE agent obeys information in the robots.txt file. As a result, if the repository has a robots.txt file, then the CORE agent must be allowed access.

IT IS IMPORTANT THAT:

- The “core” agent is allowed to access the repository, including the landing pages and the full text files.
- No crawl-delay should be specified for the “core” agent. As repositories often contain tens of thousands of records, it is important to understand that a crawl-delay of just 10s, could result in the inability to harvest the repository within a reasonable time frame. This is also why some search engines, such as Google, ignore this directive.
What is CORE

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

a comprehensive bibliographic database of the world’s scholarly literature.

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 - Global Repository Coverage

- 298M+ Metadata records
- 226M+ Free to read links to full text papers
- 33M+ Full text hosted directly by CORE
- 150+ Countries
- 12K+ Data providers
- >90 Languages

CORE - Global Repository Coverage
Local context: CORE establishing support for repositories in several English speaking countries
Difference between works and outputs

A **Work** is the single most recent version of a particular paper, that placed on the CORE website. A collection of Works represents all known versions of that paper.

An **Outputs** is a particular version of the paper, that has been placed on the repository website and indexed by CORE.
Finding version and near-duplicates

<table>
<thead>
<tr>
<th>Metadata Title</th>
<th>The reference paper</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Title</strong></td>
<td>Transverse Aeolian Ridges (TARs) on Mars</td>
</tr>
<tr>
<td><strong>Author</strong></td>
<td>Balme, M. R., Berman, D. C., Bourke, M. C., Raffkin, J.</td>
</tr>
<tr>
<td><strong>Type</strong></td>
<td>Research</td>
</tr>
<tr>
<td><strong>Field of study</strong></td>
<td>Not available</td>
</tr>
<tr>
<td><strong>DOI</strong></td>
<td>10.1016/j.geomorph.2008.03.011</td>
</tr>
<tr>
<td><strong>Publication date</strong></td>
<td>2008-11-01T00:00:00</td>
</tr>
<tr>
<td><strong>Deposited date</strong></td>
<td>2008-10-20T05:48:00</td>
</tr>
<tr>
<td><strong>Abstract</strong></td>
<td>Not available</td>
</tr>
<tr>
<td><strong>Version</strong></td>
<td>Not available for works</td>
</tr>
<tr>
<td><strong>Action</strong></td>
<td>Mark this paper as</td>
</tr>
</tbody>
</table>

**The record that the reference paper is compared with:**

- **OAI 12674**
- **OAI 12587**

**Transverse Aeolian Ridges (TARs) on Mars**

- Balme, M. R., Berman, D. C., Bourke, M. C., Raffkin, J.
- Not available
- 10.1016/j.geomorph.2008.03.011
- 2008-11-01T00:00:00+00:00
- 2008-10-20T05:48:00
- Abstract not available

**Please indicate the version of articles:**

- AO = Author's Original
- SMUR = Submitted Manuscript Under Review
- AM = Accepted Manuscript
- P = Proof
- VOTR = Version of Record
- CVTR = Corrected Version of Record
- EVTR = Enhanced Version of Record
- NA = Not Applicable (or Unknown)

**Mark this paper as**
**OAI Resolver**

**Repository configuration**

**OAI identifiers** are unique identifiers minted cost-free by repositories. Enables the CORE OAI Resolver to redirect your identifiers to your repository landing pages.

---

**Register OAI mapping**

This allows you to register a mapping between the OAI Prefix and the URL in your repository to which it should resolve. Once resolved and activated your OAI resolver will start resolving your OAI identifiers to your repository metadata page. Where not registered, we will resolve to the metadata page in CORE. To find out more see **OAI identifier documentation**.

The first part of your repository’s OAI identifiers until the colon:

- **OAI Prefix**: oar:oro.open.ac.uk

Example: oar:digital.lib.washington.edu

**Mapping to your repository’s metadata page.**

- **URL Mapping**:
  - http://oro.open.ac.uk
  - Example: digital.lib.washington.edu

Activate resolving to your repository

---

**Settings in the CORE Dashboard**

**CORE website**

**Repository landing page**
Automated detection of near-duplicates

How can you do it at scale?

\[
\begin{array}{cccccccc}
\text{simhash of A} & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 \\
\text{simhash of B} & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 1 \\
A \lor B & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\
\text{Bit population (A \lor B) = 1}
\end{array}
\]

https://moz.com/devblog/near-duplicate-detection
<table>
<thead>
<tr>
<th>Rules</th>
<th>Precision</th>
<th>Recall</th>
<th>Index duration</th>
<th>Search duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hamming distance on simhash 64 bit</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>processed_title</td>
<td>0.83</td>
<td>0.77</td>
<td>48s</td>
<td>38s</td>
</tr>
<tr>
<td>processed_abstract</td>
<td>0.72</td>
<td>0.58</td>
<td>122s</td>
<td>112s</td>
</tr>
<tr>
<td>processed_title + processed_abstract</td>
<td>0.74</td>
<td>0.58</td>
<td>145s</td>
<td>131s</td>
</tr>
<tr>
<td>processed_title + processed_abstract 50</td>
<td>0.70</td>
<td>0.61</td>
<td>61s</td>
<td>53s</td>
</tr>
<tr>
<td>processed_title + processed_abstract 100</td>
<td>0.72</td>
<td>0.60</td>
<td>69s</td>
<td>60s</td>
</tr>
<tr>
<td>processed_title_initials</td>
<td>0.74</td>
<td>0.73</td>
<td>31s</td>
<td>23s</td>
</tr>
<tr>
<td>first_and_last_letters_processed_abstract + processed_title_initials</td>
<td>0.69</td>
<td>0.56</td>
<td>83s</td>
<td>75s</td>
</tr>
<tr>
<td>fullText</td>
<td>0.76</td>
<td>0.63</td>
<td>6201s</td>
<td>6162s</td>
</tr>
<tr>
<td>title+year</td>
<td>0.78</td>
<td>0.69</td>
<td>92s</td>
<td>87s</td>
</tr>
<tr>
<td>title + abstract + year</td>
<td>0.69</td>
<td>0.55</td>
<td>186s</td>
<td>178s</td>
</tr>
<tr>
<td>title + authorstring</td>
<td>0.58</td>
<td>0.55</td>
<td>144s</td>
<td>113s</td>
</tr>
<tr>
<td>title + first author</td>
<td>0.69</td>
<td>0.60</td>
<td>103s</td>
<td>87s</td>
</tr>
<tr>
<td>title + year + first author</td>
<td>0.68</td>
<td>0.57</td>
<td>102s</td>
<td>85s</td>
</tr>
<tr>
<td>Hamming distance on simhash 128 bit</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>processed_title + processed_abstract</td>
<td>0.81</td>
<td>0.60</td>
<td>159s</td>
<td>154s</td>
</tr>
<tr>
<td>processed_title + processed_abstract 50</td>
<td>0.81</td>
<td>0.60</td>
<td>158s</td>
<td>153s</td>
</tr>
<tr>
<td>processed_title + processed_abstract 100</td>
<td>0.81</td>
<td>0.60</td>
<td>161s</td>
<td>147s</td>
</tr>
</tbody>
</table>

near duplicates in practice

Dataset from:
Deduplication of Scholarly Documents using Locality Sensitive Hashing and Word Embeddings
(Gyawali et al., LREC 2020)
Large scale near duplicate checking in CORE

1. CORE papers → Deduplication index

2. New paper → Deduplication index → Near duplicates consolidation

https://github.com/spotify/annoy
How our affiliation extraction works

Papers **without** affiliation string

- Crossref API
- DOI Extraction
- DOI
- OpenAlex API
- affiliation string
- ORCID API
- affiliation string
- ROR API
- ROR ID

Papers **with** affiliation string
How our affiliation extraction works

Less than 30% of records has affiliations associated with them. How can we find affiliations for rest of the publications?

1. Crossref API
   - Papers without affiliation string
   - Feature extraction

2. CORE API
   - Query by author names
   - Records related to author search

3. Filter records
   - (eg. year of publication, keywords, co-authors etc.)

4. Repo ID to affiliation string
   - repo id based sorting

5. ROR API
   - Affiliation string - ROR mapping
   - ROR ID
The module will help an institution to discover research articles authored by its academics across the network of open repositories.

➔ Review papers authored by your institution you might consider adding to your repository.

➔ Browse papers in your repository that we also found in other repositories.

➔ Collaboration network chart
RRS Post deposit

Dashboard

→ Review outputs with Rights Retention records found in your repository.

→ Browse Rights Retention records and include it to your repository

→ Improve your metadata quality

Outputs with Rights Retention records found in your repository

In the list below you can review and approve all RRS.

<table>
<thead>
<tr>
<th>RRS</th>
<th>Status</th>
<th>OAI</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Found</td>
<td>Review</td>
<td>Correct</td>
<td>Lorem ipsum dolor sit amet, consectetur</td>
</tr>
<tr>
<td>Found</td>
<td>Review</td>
<td>Wrong</td>
<td>adipiscing elit, sed do eiusmod tempor</td>
</tr>
<tr>
<td>Found</td>
<td>Review</td>
<td>Correct</td>
<td>incididunt ut labore et dolore magna</td>
</tr>
</tbody>
</table>

Publication date

12.07.2021

Download CSV
SDG prediction

what exists already

→ **Problem Description** - Classification of research into UN SDGs is important for efficiently monitoring and aligning the University’s research efforts with specific SDGs, targeted resource allocation and facilitating impact-driven decision-making.

→ However, manual SDG classification is tedious and therefore costly. There is currently no mechanism for classifying all OU research outputs uploaded onto the University’s repository, ORO, into SDGs. This implies that the University is not fully aware of its research contributions to SDGs.

→ Important impactful contributions might be missed and any current analysis of the proportion of OU outputs that fall under specific SDGs will be by definition incomplete.

---

Baseline - Similarity-based Model

<table>
<thead>
<tr>
<th>Research O/P</th>
<th>SDG Class Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title + Abstract</td>
<td>SDG Goals/ Targets/ Indicators</td>
</tr>
</tbody>
</table>

Encoder (Sentence BERT)

Paper representation → SDG class representation

Similarity Check

UN SDG label(s)

Results

Number of Publications and Error Rates per SDG Category

- # Publications
- Error Rates

SDGs: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17
AI tools for automating the classification of research into UN SDGs

1. Indexing repository research outputs
2. Systematic processing of all research outputs
3. Metadata/full text passed to a ML model
4. AI SDG classification
5. Integration of SDG-labelled research output into CORE
6. API response with SDGs

CORE
A global comprehensive bibliographic database of open research

UN SDGs

SDG label confidence score

UN SDG label(s)

API

Dashboard

The Open University

OSC Platform

OU

Systematic processing of all research outputs

Metadata/full text passed to a ML model
Sustainable Development Goals (SDGs)

The Sustainable Development Goals (SDGs) are a collection of 17 interlinked global goals designed to be a "blueprint to achieve a better and more sustainable future for all". The SDGs were set up in 2015 by the United Nations General Assembly and are intended to be achieved by 2030.

Number of articles with SDG

- Metadata records: 52,286K
- 76% Full texts
TEI dataset: supporting TDM part of funder requirements (such as Nelson Memo)

Plans should describe:

“How to maximize equitable reach of public access to peer-reviewed scholarly publications, including by providing free online access to peer-reviewed scholarly publications in formats that allow for machine-readability and enabling broad accessibility through assistive devices”

*From the Nelson Memo on Ensuring Free, Immediate, and Equitable Access to Federally Funded Research*
Integrate Software citation in the CHARS pipeline
AI: leveraging local content for Retrieval Augmented Generation (RAG)

- Question / Answer platform
- Retrieval Augmented Generation (RAG)
- Leverages LLM + full text content from CORE to generate authentic, trustworthy answers

**CORE-GPT**

- For domain or subject-specific repositories
- Generate answers based only on local content

**CORE-GPT + Domain Specific Knowledge**
Generate a comprehensive answer to the following question (but not more than 160 words) solely based on the provided content. Format the links to the papers as follows: `{url:$url, abstract:$abstract},$question`
Local vs global context and OA compliance:

**UK**
- Research Councils UK (RCUK) Policy on Open Access
- Wellcome Trust

**EU**
- Horizon2020

**USA**
- OSTP (Nelson Memo)
- National Institutes of Health (NIH) Public Access Policy

**Global**
- PlanS
- Gates Foundation
CORE FAIR certification

The Rights Retention Strategy (RRS) enables authors to exercise the rights they have on their manuscripts to deposit a copy of the Author Accepted Manuscript (AAM) in a repository on publication and provide open access to it.

Findable

1. Your repository supports OAI-PMH?
   The Open Archives Initiative Protocol for Metadata Harvesting (referred to as the OAI-PMH in the remainder of this document) provides an application-independent interoperability framework based on metadata harvesting. Read more.
   Yes

2. Your repository metadata supports a widely used Application Profile
   OpenAIRE Guidelines 2.0+ or MDRIF.
   Yes

Accessible

Once the user finds the required data, they need to know how they can be accessed, possibly including authentication and authorization.

Findability

- Your repository supports OAI-PMH?
  Yes

Wideness

- Your repository metadata supports a widely used Application Profile
  Yes

Accessibility

- Once the user finds the required data, they need to know how they can be accessed, possibly including authentication and authorization.

Gold

Silver

Bronze

Certified 16.08.2023

Dashboard

Repository name

Overview

Harvesting status

Content

OA Compliance

UKRI compliance

Plan S compliance

DDI

ROXX Validator

Deduplication

Footprints

SGO

IRIS Policy

Plugins

FAIR certification
USRN Desirable Characteristics of Digital Publication Repositories Report

In defining the desirable characteristics outlined below, the U.S. Repository Network (USRN) Desirable Characteristics Working Group strove to balance feasibility and necessity.

26.05.2024

Free and Easy Discoverability & Access

The repository provides broad, equitable, and maximally open access to resources and their metadata in both machine- and human-readable formats at no cost to users. This should be done in a timely manner after submission and with limited unscheduled downtime, consistent with legal and policy requirements. The repository’s content should be discoverable both within the repository (e.g., via a search interface) and externally (i.e., via discovery services and aggregators). The repository supports access to its content for persons with disabilities and adheres to current web accessibility standards. Provide an accessibility statement describing efforts to increase accessibility of the platform/content and how to seek help if an accessible version of a digital publication is not available.

1 Your repository supports OAI-PMH?

The Open Archives Initiative Protocol for Metadata Harvesting (referred to as the OAI-PMH in the remainder of this document) provides an application-independent interoperability framework based on metadata harvesting. Read more.

2 Your repository metadata supports a widely used Application Profile

OpenAIRE Guidelines 3.0+ or RIOXX

Your metadata compliant with RIOXX: 97.3%

3 Indexed content

Number of metadata records: 18.69K
CORE - An adopter of POSI

The Principles of Open Scholarly Infrastructure

Core is a mission-driven not-for-profit endeavour and a signatory of the Principles of Open Scholarly Infrastructure.
Community governance

Participation of members:

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
CORE Members
Supporting and Sustaining
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:
  - Participate in governance and help shape our development roadmap
  - Support and sustain CORE
Three levels of CORE Membership

**Starting** (Free)
Organisations who want to get their content indexed in CORE and benefit from increased discoverability (in search, Recommender, Discovery and PMC) and persistent identification (via OAI Resolver), while staying in control of how their content is exposed using the CORE Dashboard.

**Supporting**
Organisations who want to be widely recognised for their open research. In addition to all the STARTING benefits, this membership also comes with extra visibility for your organisation in the OA network, additional technical support, CORE API access for your researchers, and a seat on our Board of Supporters.

**Sustaining**
Organisations who are truly open research champions. In addition to all the SUPPORTING benefits, this membership comes with additional tools to help stay compliant with OA policies and FAIR, unlimited technical support, CORE Dataset access for your researchers for text and data mining, and two seats on our Board of Supporters.
<table>
<thead>
<tr>
<th><strong>Starting</strong></th>
<th><strong>Supporting</strong></th>
<th><strong>Sustaining</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Free</td>
<td>All Supporting benefits PLUS</td>
<td>All Supporting benefits PLUS</td>
</tr>
</tbody>
</table>

- CORE Repository Dashboard
- Get your OAI identifiers resolved to your repository
- CORE Discovery
- CORE Search
- CORE Recommender
- General support
- Download statistics
- CORE Discovery Boost:
  - PubMed Central (PMC)
  - Recommended across the repositories network
- RIOXX metadata validator

- Versions and duplicates detection
- Rights Retention Strategy (RRS) Identification
- API use for your university
- Technical support (£5)
- Logos for your papers
- Logo banner
- Monitor RIOXX compliance
- Board of Supporters (1 vote)

- CORE OA Compliance dashboard
- Dataset use for your university
- Repository health check
- Unlimited and prioritised support
- Personalised banner
- Promoted / spotlight / featured articles
- Hosted or interview style blog post on mission-aligned activities
- Board of Supporters (2 votes)
Conclusion

- Balancing global and local contexts.
- Continuing to address global challenges.
- Working closely with individual repositories in the UK, North America (USRN) and South Africa to address their specific local needs.
- New functionalities to provide added value back to repositories.
- CORE Dashboard an important gateway for repositories to access services from, but repository integrations in the future also to play a role.
Thank you!

Petr Knoth
petr.knoth@open.ac.uk