CORE: Infrastructure for text mining of open access content at scale
STM Week 2019 - Innovations

Dr Petr Knoth

Big Scientific Data and Text Analytics Group
Knowledge Media Institute, The Open University
CORE’s mission

Aggregate all open access research articles worldwide ...

... enrich this content and provide seamless access to it through a set of data services ...
CORE harvests from repositories
Harvesting is challenging

Many OAI-PMH implementations challenges ...

- No content harvesting support
- Restrictions on full text downloading
- Failing resumption tokens
- Incremental updates
- Reliability
- Scalability
- Sequential nature of OAI-PMH
- Locating full text URLs in metadata
- Metadata interoperability
Harvesting data is challenging
Harvesting data is challenging
CORE Processing pipeline

• Metadata download, extraction and harmonisation
• Full text download
• Text extractions, sections extraction
• Metadata validation and enrichment (DOI, ORCID, etc.)
• Thumbnails generation
• References and citation contexts extraction
• API enrichment (e.g. finding DOIs, linking to other systems)
• Document type classification
• Deduplication
• Indexing
• Exposing (data dumps, API, FastSync)
How often is the CORE content updated

- Data providers harvested as frequent as hardware allows
- Harvesting time is specified by the CORE scheduler
  - Last time the repository was harvested
  - Repository size
  - Repository location
  - Repository harvesting performance
  - Previous information about harvesting errors
- Schedule functionality reviewed on a regular basis
World’s largest dataset of Open Access full texts

- 14,389,274 Hosted full texts
- 24,936,921 Access to free to read full texts
- 135,539,113 Metadata records
- 9,645 Data providers

Majority of records in CORE do not have an equivalent in Crossref.
January 2019 – CORE reached over 10M monthly active users for the first time

571% increase from January 2018
CORE Usage

- December 2019 – CORE within top 5k websites globally by user engagement.
- A combination of daily visitors and page views on a website over a 3 month period.
- core.ac.uk by usage in the top 0.0009% of global websites.
CORE wins an award

- Outstanding Impact of Research on Society and Prosperity Award 2019
- Greatly motivated to serve the community even more!
CORE services

• Access to raw data
  - API
  - Dataset
  - FastSync

• Content discovery
  - Recommender
  - Discovery

• Managing content
  - Repository Dashboard
CORE’s raw data services
Raw data services – CORE API

• Enables the development of new applications
• Real-time machine access to the world's largest collection of open access papers
• Harmonised access to data from across the network of CORE providers
• Direct machine access to full texts of research papers
Raw data services – CORE Dataset

• Download millions of research papers for text and data analysis
• Prototype, analyse and mine your data in your infrastructure
Raw data services – CORE FastSync

- Keeps your data in sync with research content from around the world
- Fast and incremental updates as soon as they become available. No usage restrictions
- Based on ResourceSync
Types of collaborations

ontochem IT SOLUTIONS
Microsoft
INNOVATION ENGINEERING

IRIS.AI
turnitin

NAVER
LEAN Library
CACTUS
Artificial Researcher
Use cases powered by CORE

- It is beyond human capacities to read all scientific literature
- Example use cases in which CORE is applied:
  - Improving discovery
  - Plagiarism detection
  - Question answering in science
  - Literature based discovery
  - Fact checking and detection of misinformation
  - Analysing research trends
  - Finding experts in a particular domain
  - Research evaluation and scientometrics
  - Exploratory and visual search
  - Classifying citations based on context
  - …
CORE Opportunities

• Growing demand for raw data access services
  • Help companies to develop innovative services analysing and mining research papers

• Monitor compliance with Plan S
  • Help institutions to comply as well as monitor their compliance

• Development of products to serve the needs of HEIs
  • Help institutions to increase the discoverability of their research outputs via CORE services (Recommender, Discovery, Search, integrations with other systems, etc.)
  • Make repositories more engaging
Take home

- Data providers (repositories, preprint servers, journals, etc.) and aggregators need to work together to allow text and data analysis, processing and reuse of large volumes of research papers.
- CORE provides the tools for programmatically processing open access data fast, reliably and from across the global network of repositories.
- If you are a developer or analyst:
  - Build your own stuff using CORE’s data services on top of the global full text open access corpus
- If you are a company wanting to text mine research papers globally:
  - Talk to us and register for our raw data services.
Thank you!

https://core.ac.uk