Institution: The Open University		
Unit of Assessment: B11 Computer Science and Engineering		
Title of case study: Making scientific knowledge more easily and freely discoverable through CORE.ac.uk		
Period when the underpinning research was undertaken: 2010-2020		
Details of staff conducting the underpinning research from the submitting unit:		
Name(s):	**Role(s) (e.g. job title):**	**Period(s) employed by submitting HEI:**
Dr Petr Knoth	Senior Research Fellow in Text and Data Mining	2008 - present
Period when the claimed impact occurred: 2015 - 2020		
Is this case study continued from a case study submitted in 2014? N		

1. Summary of the impact

Dr Knoth's novel research in aggregating research repositories has created the world's largest Open Access (OA) collection of research literature and pioneered machine access to this content. His not-for-profit CORE.ac.uk is ranked among the top 1,500 websites in the world by website popularity and top 20 websites in Science and Education. It has over 40 million Monthly Active Users (MAU) and is valued at USD94,250,000. CORE.ac.uk has played a pivotal role in the global open science movement by:

- Widening access to and discoverability of scientific knowledge and driving change towards open science;
- Empowering commercial and academic partners to develop novel solutions leveraging scientific literature;
- Ensuring compliance with research funders' open access policies.

2. Underpinning research

Since 2010, Dr **Knoth's** research has tackled the issue of aggregating content from the growing number of online Open Access (OA) research repositories and journals to create CORE.ac.uk (CORE), the world's largest OA collection of scientific outputs. The project has attracted close to GBP4.1 million in research funding and income from licencing, including significant grants for a total of GBP1.6 million from Jisc.

Building CORE's vast, continuously updated dataset required overcoming a number of research challenges, such as those related to the lack of interoperability, scalability, regular content synchronisation and redundancy. Dr **Knoth's** key innovation in this area concerns the development of a unique solution which utiliFPses the OAI-PMH protocol, originally designed for metadata harvesting, for the purpose of efficient content aggregation of scientific literature **[O2, O3]**.

The issue of providing effective content-based recommendation to the aggregated content was tackled in a 2010 paper **[O1]**, which defined a novel approach to the problem, which balances content similarity and diversity. This work laid the foundations of CORE Recommender [O2], one of the key components of CORE.

Later, in a 2018 paper, Dr **Knoth** and his team further enhanced this approach by introducing the use of citation proximity functions in the recommendation engine **[O4]**. Moreover, in an article the following year, the researchers introduced the first 'living lab', to enable researchers in scholarly recommender systems to conduct online evaluations of their algorithms **[O5]**. In particular, this research analysed the performance of CORE Recommender, as measured with respect to a large sample of 1,826,643 recommendations in the 16 months since its launch in 2018, and further contributed to the refinement of CORE Recommender **[O4]**.

In addition to developing novel solutions forming the key components of CORE, Dr **Knoth's** team also investigated the broader issues associated with the application of open access policies. In particular, in a 2019 best paper award winning paper, they showed how one can effectively track the time lag between the date of publication and date of deposit into a research repository across thousands of repositories globally using CORE data **[O6]**. This work was motivated by the need of the research community not just to have research papers available as open access, but to also achieve timely open access. This research work also showed that the introduction of a deposit timeframe significantly reduces the time from a research output being accepted for publication in a repository or journal to it becoming publicly available. This crucial finding supports the argument for setting specific timeframes as a deposit requirement in open access policies and forms the basis of CORE's application in the area of OA policy compliance.

3. References to the research

O1. Automatic generation of inter-passage links based on semantic similarity (2010). **Knoth, Petr**; Novotny, Jakub and Zdrahal, Zdenek. In: Computational Linguistics (COLING 2010) (23-27 Aug 2010, Beijing, China) (pp. 590-598). http://oro.open.ac.uk/22933/. Top conference in natural language processing, 638 submissions, 160 accepted, *A ranked* conference according to CORE Ranking Portal, top 19% of conferences.

O2. CORE: three access levels to underpin open access (2012). **Knoth, Petr**, and Zdrahal, Zdenek. D-Lib 18 (11/12), pp. 1-13. https://doi.org/10.1045/november2012-knoth. Peer-reviewed, highly referenced by others.

O3. Building scalable digital library ingestion pipelines using microservices (2017). Cancellieri, Matteo; Pontika, Nancy; Pearce, Samuel; Anastasiou, Lucas and **Knoth, Petr**. In: MSTR 2017: 11th International Conference on Metadata and Semantics Research (28 Nov - 1 Dec 2017, Tallinn, Estonia). http://oro.open.ac.uk/51070/. Peer-reviewed.

O4. Towards effective research recommender systems for repositories (2017). **Knoth, P**.; Anastasiou, L., Charalampous; A., Cancellieri, M.; Pearce, S.; Pontika, N.; & Bayer, V. In: Open Repositories 2017, Brisbane, Australia. http://oro.open.ac.uk/49366/. Peer reviewed.

O5. Online Evaluations for Everyone: Mr. DLib's Living Lab for Scholarly Recommendations (2019). Beel, Joeran; Collins, Andrew; Kopp, Oliver; Dietz, Linus W. and **Knoth, Petr**. In: *Advances in Information Retrieval (Part 2)*, Lecture Notes in Computer Science, Springer, Cham, pp. 213–219. https://doi.org/10.1007/978-3-030-15719-7_27. Peer-reviewed, presented at ECIR 2019, *A ranked* conference according to CORE Ranking Portal, top 19% of conferences.

O6. Do Authors Deposit on Time? Tracking Open Access Policy Compliance (2019). Herrmannova, Drahomira; Pontika, Nancy and **Knoth, Petr**. In: 2019 ACM/IEEE Joint Conference on Digital Libraries (2-6 Jun 2019, Urbana-Champaign, IL). https://doi.org/10.1109/JCDL.2019.00037. *A** conference according to CORE Ranking Portal, top 4% of conferences, 155 submissions, highest ranked conference in Digital Libraries, **Vannevar Bush Best Paper Award.**

4. Details of the impact

Impacts on participation, learning and understanding: Widening access to and discoverability of scientific knowledge and driving change towards open science

Dr Knoth's research **[O1-O6]** has created the world's largest **[C1]** and continuously growing aggregator of full-text open access research content from repositories and journals. Since 2015, CORE has been delivered by The Open University as a national aggregation service and financially supported by a partnership with Jisc, the UK digital solutions provider for education and research. The not-for-profit CORE hosts more than 25 million free-to-read full texts (9x larger than the world's second largest OA collection provided by PubMed Central) and provides access to many more papers via its 224 million metadata records collection, acquired from over 10k data providers. As of 22nd December 2020, the site welcomes more than 40 million Monthly Active Users (MAU) (making it by far the most used OA aggregator globally**) [C3**, p.1**]**. It ranks 1,445 in Internet traffic according to Amazon's Alexa Global Rank (compared to rank 256,508 of its nearest competitor BASE-search.net) **[C3**, p.2**]**. CORE is also in the top 20 websites globally in the category of Science and Education according to SimilarWeb **[C3**, p.3**]** and valued at USD94,250,000 by WorthOfWeb **[C3**, p.4**]**.

Jisc's Director of Open Access Services says: "*CORE has been one of Jisc's key infrastructure open access services and a strategic service facilitating transition to open access and science in scholarly communication. CORE has also been acting as an enabling infrastructure for other Jisc* *digital services powered through CORE's technology. Since 2011, CORE has greatly advanced openness by providing digital services and support to UK and global HEIs in the delivery of an interoperable network of open access repositories, a cornerstone of Jisc's fight for affordable and equitable research communication*” **[C2**, p.1**]**.

Executive Director of the Confederation of Open Access Repositories (COAR), with a membership of over 150 institutions worldwide from over 50 countries in 5 continents, stated: “*CORE has significantly assisted the academic institutions participating in our global network with their key mission, which is their scientific content exposure. In addition, CORE has helped our content administrators to showcase the real benefits of repositories via its added value services*” **[C2**, p.4**]**.

CORE provides an aggregator capable of quickly drawing together full-text scientific content from distributed research repositories and journal websites at scale **[O2-O3]**. As a result, CORE has become an essential open science infrastructure, which is relied on by millions of people around the world and substantially contributes to the global OA movement. In particular, CORE provides an important service for disadvantaged groups and in developing countries. More than 60% of its users come from less economically developed countries, where access to education and research is more challenging **[C3**, p.5**]**.

Open Access Programme Manager for Eifl, quotes: “*Most of the world’s peer-reviewed articles are still locked in subscription journals, inaccessible to researchers in the Global South. The COVID-19 pandemic has demonstrated the need for the rapid sharing of research findings to inform urgent public health responses. CORE has been enabling free and unrestricted availability of peer-reviewed research literature, removing the barriers that researchers, policy makers, professionals and the general public in the Global South face trying to access critical research and information. We have seen again and again how access to such vital information has been improving people’s lives*” **[C2**, p.7**]**.

The aforementioned COAR Executive Director, adds: “*CORE has significantly contributed to the global scholarly communications landscape by facilitating a transition from unaffordable and spiralling costs of publisher subscription models to new affordable and equitable systems where no-one, wherever they are, is denied access to knowledge resulting from publicly-funded research*" **[C2**, p.4**]**.

Impacts on commerce and the economy: Empowering commercial and academic partners to develop novel solutions leveraging scientific literature

Underpinned by Dr **Knoth’s** novel research to overcome the lack of interoperability between different research repositories **[O2, O3]**, CORE has become a treasure trove for:

- developers, analysts, text and data miners, and
- scientific repositories and Higher Education Institutions (HEIs),

who have adopted, used and benefited from the following CORE services:

1. **CORE Raw Data Services** (**CORE Dataset, CORE API, CORE FastSync)** provide machine access to CORE data, enabling developers to build and run innovative applications. These services have been provided either as **strategic partnerships** or supplied under a **commercial licence**.

a. Strategically, CORE has partnered with Microsoft, who link CORE Data to enrich Microsoft Academic Graph (MAG), a heterogeneous graph containing scientific entities which Microsoft uses to power experiences in Bing, Cortana, Word, and Microsoft Academic. Writing in November 2020, Microsoft Outreach Academic Services Managing Director, states: “*In 2018, Microsoft partnered with CORE to link documents in the MAG with open access articles in CORE, as the datasets complement each other. The outcome is the world’s largest scholarly graph with both bibliographic information, including citations, and full texts of academic papers supplied by CORE for machine processing. In addition to enriching MAG, these links to open access research papers in CORE have also been fed directly to the Microsoft Academic Search service, enabling millions of Microsoft Academic Search users to access full text research papers more easily*” **[C2,** p.10**]**.

b. As of 22nd December 2020, a total of 3,891 developers have registered for access to CORE's raw data services. Motivated by the ability to reduce costs and bring their new products and services faster to the market by adopting CORE's new technology, eleven companies in 9 countries have so far purchased a licence allowing them to commercialise CORE's raw data services in their businesses. These include:

   i. Turnitin: Senior Vice President Business & Corporate Development, market leading plagiarism detection company Turnitin, acquired in 2019 by Advance Publications for USD1.75 billion, wrote how CORE FastSync ensures Turnitin's "*database remains at the forefront of publishing trends and can continue to best serve the needs of our customers and partners*" **[C2**, p.26**]**. Content Manager at Turnitin, adds that now, thanks to the integration with CORE, "*Turnitin will be able to ensure that we will always be crawling the latest and most valuable open access content into our database*" **[C2**, p.13**]**.

   ii. **Artificial Researcher:** CEO of Artificial Researcher, testifies that CORE powers all of her company' services, enabling the company to offer "*trustworthy, informed and rich results to clients.*" She adds that the partnership with CORE "*represents a substantial cost benefit that has enhanced the growth of Artificial Researcher as a whole" and that "the inclusion of CORE content greatly enhances the end-user experience for researchers, academic libraries and commercial clients*" **[C2**, p.16**]**.

   iii. **IRIS.ai:** CTO of Artificial Intelligence Science Assistant platform Iris.ai, states in his 2019 testimonial how CORE is helping the company to speed up its ambition to "*make open access articles more easily accessible*" **[C2**, p.28**]**.

   iv. **Ontochem:** CEO Ontochem GmbH, states: "*Our partnership with CORE has significantly increased the depth and breadth of full text content available to our customers, greatly enhancing the value of our offering overall while delivering savings in terms of time, cost and human resource to our company in the area of data aggregation. The full text data provided by CORE is fundamental to the operation of our SciWalker platform, a leading search and analysis tool for the life sciences."* **[C2**, p.19**]**.

2. **CORE Recommender.** Capitalising on Dr **Knoth's** research into recommender systems **[O3, O4]**, the CORE Recommender suggests free relevant articles to repository users based on what they read. More than 335 scientific repository systems have been actively using the CORE Recommender as of December 31st 2020, including Cambridge University and the most used repository in the world, arXiv.org. In December 2020, arXiv IT Lead states: "*Fast and free access to scientific literature is a cornerstone of the arXiv ethos. Our partnership with CORE via the Recommender allows our users to easily and quickly discover relevant research literature, benefits groups with restricted access to scientific literature and contributes to the global open access mission that both arXiv and CORE support"* **[C2**, p.22**]**. Writing on CORE.ac.uk, in a March 2019 testimonial, Deputy Manager of Scholarly Communication (Open Access) at The University of Cambridge stated that the service "*greatly enhances the functionality of our repository and provides our users with topical open resources that are only a click away*" **[C2**, p.31**]**.

3. **The CORE Repository Dashboard** is an online interface that offers academic partners valuable technical information and statistics. Between its launch in 2015 and December 2020, repository managers and research communications professionals from more than 75 per cent of UK universities have used the service to analyse their research outputs and manage their repository collections **[C3**, p.6**]**. Users from 27 countries, such as the United States, Japan, Russia, have created 330 Dashboard accounts since 2015. University of Strathclyde, Institutional Repository Manager explains that the service "*provides an excellent level of feedback and control over what gets harvested, aggregated and exposed*". He also noted that "*it provides intelligence on harvesting errors which may need troubleshooting, thereby ensuring optimum repository health for text mining as well as discovery of repository content*" **[C2**, p.34**]**. Executive Director of COAR, states: "*These services have introduced content administrators to a wealth of new information relating to their repository systems and research outputs that they were not aware of, and have made it easier for them to meet their daily tasks and speed up their efforts*" **[C2**, p.4**]**.

Impacts on public policy, process and professional practice: Ensuring compliance with research funders' open access policies

In 2019, based on award winning research **[O6]**, a cooperation of CORE with Research England was initiated. Research England subsequently announced that CORE Data will be used in the REF2021 OA Policy Audit **[C4**, pp.8 and 10; para. 40 and 49**]**, to enable the REF 2021 audit committee to identify non-compliant research outputs. In doing so, the method of measuring *deposit time lag* proposed in **[O6]** was implemented in CORE and a new process leveraging CORE data to support REF2021 Audit has been agreed between CORE and Research England, establishing a formalised collaboration.

Additionally, in 2020, CORE released a REF2021 OA policy compliance tool that has been embedded into the CORE Repository Dashboard. As of 23rd December 2020, our access logs show that 127 UK HEIs, i.e. 77% of the total, have used free compliance checking features provided by this tool **[C3**, p.13**]**. Moreover, at the time of writing, London City University, St George's University London, Nottingham Trent University and Anglia Ruskin University have already purchased a premium licence to this tool and the number of premium subscribers is increasing.

5. Sources to corroborate the impact

C1. Evidence of CORE being the world's largest aggregator of full-text open access scientific content from repositories and journals:

- Jisc scholarly communications blog post (pp.1-3).
- Continuously updated statistics at CORE website (https://core.ac.uk/data/) (p.4).

C2. Evidence of CORE.ac.uk testimonial letters and web testimonials

- Director of Open Access Services, Jisc (pp.1-3).
- Executive Director, Confederation of Open Access Repositories (pp.4-6).
- Open Access Programme Manager, Eifl (pp.7-9).
- Microsoft Research Academic Services (pp.10-12).
- Content Manager, Turnitin (pp.13-15).
- CTO - CEO, Artificial Researcher IT GmbH (pp.16-18).
- CEO, Ontochem (pp.19-21).
- IT Lead, arXiv (pp.22-24).
- Senior Vice President Business & Corporate Development, Turnitin, UK (pp.25-27).
- CTO, IRIS AI AS (pp.28-30).
- Deputy Manager of Scholarly Communication (Open Access), Cambridge University, (p.31-33).
- Institutional Repository Manager, University of Strathclyde, UK (pp.34- 36).

C3. CORE.ac.uk traffic and services statistics:

- CORE's 41 million monthly active users (p.1).
- CORE ranking in global Internet user engagement according to Alexa Global Rank (p.2).
- CORE ranking as top 4 website in the category of Science and Education globally by SimilarWeb (p.3).
- CORE is estimated to be worth USD94,250,000 by WorthOfWeb (p.4).
- Screenshot on CORE's users per country - more than 60% of these users come from less economically developed countries, where access to education is more challenging (p.5).
- Dataset of more than 50% repository managers and research communications professionals from UK universities have been using the CORE Repository Dashboard (p.6).
- Dataset with UK HEIs access logs of the CORE REF2021 Open Access policy compliance tool (p.13).

C4. REF2021 Audit Guidance where CORE has been selected to provide data for the audit. REF 2021 OA Policy Audit will use CORE data to measure and assess policy compliance.