Institution: The Open University

Unit of Assessment: B11 Computer Science and Informatics

Title of case study: Enabling exploration of hidden, contextual knowledge within large collections of documents

1. Summary of the impact

COnnecting REpositories (CORE) is a system for aggregating, harvesting and semantically enriching documents. As at July 2013, CORE contains 15m+ open access research papers from worldwide repositories and journals, on any topic and in more than 40 languages. In July 2013, CORE recorded 500k+ visits from 90k+ unique visitors. By processing both full-text and metadata, CORE serves four communities: researchers searching research materials; repository managers needing analytical information about their repositories; funders wanting to evaluate the impact of funded projects; and developers of new knowledge-mining technologies. The CORE semantic recommender has been integrated with digital libraries and repositories of cultural institutions, including the European Library and UNESCO. CORE has been selected to be the metadata aggregator of the UK's national open access services.

2. Underpinning research

Research in CORE builds on two earlier projects: Bletchley Park Text and Eurogene. The underlying research topic is the aggregation, semantic enrichment and user exploration of unstructured, mainly textual resources. This problem was tackled in three steps with increasing complexity:

1. In 2002-2007, our research was focused on restricted domains, monolingual text and manual annotation (Bletchley Park Text)
2. Between 2007 and 2011 we extended our investigation to multilingual text and automatic annotation but still within restricted domains (Eurogene)
3. Since 2011, we have been further extending our research to the most general case, i.e. unrestricted domain, multilingual text and automatic annotation (CORE)

In all cases, resources are enriched by calculating cross-resource semantic similarity. In restricted domains we annotated resources using domain ontologies. For large-scale unrestricted domains, we used methods of natural language processing of full-text.

Bletchley Park Text

As a part of the CIPHER project, we used transcripts of Bletchley Park wartime employees' memories, which were manually annotated using domain ontologies. To capture the story structures, we developed an event-based ontology with the museum standard "International Council of Museums Conceptual Reference Model (CIDOC CRM)" ontology as the backbone, which prior to this REF impact census period already contributed to the CIDOC CRM standardisation (ISO 21127:2006). Events are interlinked through domain concepts, such as people, places, objects, temporal data, etc. We developed novel methods for exploring pathways between concepts with mutual information used as a measure of semantic cohesion [3.1].

Eurogene

The Eurogene project processes PowerPoint, Word, pdf and text files in ten languages. In collaboration with geneticists, we developed a multilingual genetic ontology for the automatic annotation of machine-readable content and a theme hierarchy dividing genetics into sub-disciplines. English documents are also annotated using the Universal Medical Language System (the world's largest medical ontology). Our algorithms use these knowledge structures for cross-

language search and the discovery of semantically related documents. We used semantic relatedness to measure the coherence of larger sequences of educational resources that may serve as course material for teaching genetics [3.3].

CORE

CORE harvests and aggregates content from open access repositories across all domains and multiple languages. Bletchley Park Text and Eurogene resources were 'pushed' by the content provider but CORE 'pulls' resources automatically via repository directories. Unlike Bletchley Park Text and Eurogene, CORE's text mining system does not require the construction of any domain ontology and thus can be instantaneously applied to any domain.

The CORE theoretical research questions include:

1. Does the semantic relatedness calculated by natural language processing methods from full-text relate to subjective concept relatedness as perceived by humans? In [3.4] we have shown that the answer is positive.
2. Can relatedness be calculated across different languages? We present a novel approach based on large-scale explicit semantic analysis in multiple languages [3.2].

Application-focused research included the development of scalable algorithms for harvesting large volumes (15m+ records, both full-text and metadata) of resources from heterogeneous distributed repositories. The harvested resources are processed by calculating semantic similarity, de-duplication, language detection, text classification, extraction of citation networks and others. The enriched content aggregated is presented back to the user.

We were the first to introduce the methodological approach of 'three access levels architecture' (preprocessing text in a form suitable for knowledge discovery and text mining, transactional document access and usage analytics), as the necessary condition enabling the mass reuse of open access content [3.6].

Research team with start date at the OU: Zdenek Zdrahal (1991, PI, Senior Research Fellow), Petr Knoth (2006, Research Associate), Paul Mulholland (1995, Research Fellow), Annika Wolff (2000, Research Associate), Trevor Collins (1997, Research Fellow).

3. References to the research (key references in bold)

[3.1] **Zdrahal, Z., Mulholland, P. and Collins, T. (2008) 'Exploring Pathways Across Stories', *IEEE SMC Conference Distributed Human-Machine Systems* (DHMS 2008), March 8-12, 2008, Athens, Greece, pp. 341–6**

[3.2] **Knoth, P., Zilka, L. and *Zdrahal, Z.* (2011) 'Using Explicit Semantic Analysis for Cross-Lingual Link Discovery', *5th International Workshop on Cross Lingual Information Access: Computational Linguistics and the Information Need of Multilingual Societies (CLIA) at The 5th International Joint Conference on Natural Language Processing (IJC-NLP 2011*), 08 - 13 Nov 2011, Chiang Mai, Thailand.**

[3.3] Zdrahal Z., Knoth P., Collins T. and Mulholland P. (2009) 'Reasoning across multilingual learning resources in human genetics', *Interactive Computer Aided Learning*, ICL 2009, Villach, Austria. September 23-25, 2009, ISBN: 978-3-89958- 481-3, Kassel University Press

[3.4] **Knoth, P., Novotny, J. and Zdrahal, Z. (2010) 'Automatic generation of inter-passage links based on semantic similarity', *The 23rd International Conference on Computational Linguistics* (COLING 2010), 23-27 August 2010, Beijing, China. ISBN-9781627481793**

[3.5] Knoth, P. and Herrmannová, D. (2013) 'Simple yet effective methods for cross-lingual link discovery (CLLD) – KMI @ NTCIR-10 CrossLink-2, *NTCIR-10 Evaluation of Information Access Technologies*, Tokyo, Japan, http://people.kmi.open.ac.uk/petr/papers/ntcir-10-crosslink.pdf

[3.6] Knoth, P. and Zdrahal, Z. (2012) 'CORE: three access levels to underpin open access', *D-Lib Magazine*, vol. 18, no. 11/12. http://www.dlib.org/

References [3.1] – [3.5] have been published in peer reviewed, competitive conferences; the key references are in bold. The novel architecture described in [3.6] is the basis of the CORE system selected by Jisc as the UK national Open Access aggregator.

Grant funding

CIPHER (2002-04): Bletchley Park Text was a part of CIPHER project supported by the CEC, funding £1.1M
Eurogene (2007–10): Project supported by the CEC in the eContentPlus programme, funding £1.2m
CORE (2011 – present): Project supported by Jisc, funding £243k

4. Details of the impact

CORE (http://core.kmi.open.ac.uk) is a unique aggregation system, which enables free access and reuse of open access content at a full-text level from a single access point. As of July 2013, CORE harvests 498 open access repositories (including all UK ones) and journals, resulting in 15m+ metadata records and 1.5m full-text documents. CORE detects semantically related or duplicate content from full-text and metadata and, where possible, constructs citation networks. The system provides a range of services for accessing and exposing the aggregated data: CORE Portal, CORE Mobile, CORE Recommendation Plugin, Application Programming Interface (API), Linked Open Data and Repository Analytics. CORE was the first system implementing the 'three access levels from a single source' architecture for open access [3.6]. These levels (transaction, analytical, raw data access) can be mapped to the following REF impact categories.

Impact on public policy and services

Impact on services for the general public. According to Google Analytics, in July 2013 CORE received more than 540,000 visits and served about 90,000 users. Between January and August 2013, this number grew steadily by 10–15% each month. If the requested document is represented only by metadata, CORE offers all available full-text documents that are semantically related [5.2]. For its unique potential of content discovery, CORE has been declared as one of 'The top ten search engines for researchers that go beyond Google' [5.5]. CORE provides free apps for Android and iOS [5.9]. About 1.5% of all requests come from mobile devices.

Impact on digital libraries and compliance with future REF policies [5.10]. CORE provides analytical information about the content and compliance of repositories with common standards. CORE's repository-related services are critical for the new generation of post-2014 REFs because open access accessibility is a key HEFCE requirement.

Service to institutional repositories. CORE researchers developed a new content recommendation tool, the CORE Widget, that is available in the EPrints Bazaar Store repository (http://bazaar.eprints.org/) [5.7]. This tool helps users to discover open access content effectively.

Impact on institutional repositories. CORE recommender allows the host repository to offer the user semantically similar documents across all harvested repositories. This novel service is used by UNESCO [5.4], The Open University's Open Research Online, University College London Computer Centre, the European Library [5.1], Glasgow University and Goldsmiths College.

Economic impact

A Jisc-commissioned report [5.8] estimates the increase in productivity by text mining as being worth between £123.5m and £156.8m per year. CORE is the only UK service to provide access to raw metadata and the content of research papers as downloadable files or through an API, which

is essential for the development of novel text-mining applications in the digital library and media industry. Between January 2012 and July 2013, thirty-four users requested and received access to CORE preprocessed text to develop their own text-mining applications.

Impact on society, culture and creativity

Since 2012 CORE has been the European Library's default remote search service and is routinely used by their users worldwide. A significant number of access requests to the CORE repository originate from the European Library search engine [5.1].

CORE is listed 2nd in the *TOP 100 Thesis & Dissertation References on the Web* [5.3].

Algorithms for calculating cross-language relatedness developed within CORE scored 'overall-best' in one of the two categories of the cross-language interlinking competitions [3.5] in 2011 and 2013 and were among the top performers in the other.

A member of the CORE team has been invited to present CORE at the World Summit on Information Society, WSIS-10 Review, held in Paris, February 23, 2013 'Using E-Science to Strengthen the Interface between Science, Policy and Society: High Level Roundtable' [5.6].

CORE is the essential component of UNESCO's Repository for Connecting Local and International Content (CLIC). At the World Summit on the Information Society, WSIS-10 Review, Ms Christina von Furstenberg (Senior Programme Specialist, UNESCO) presented the CORE-based CLIC as 'the next generation of tools for the [UNESCO's] Management of Social Transformations (MOST) … that allows comparative access to research, policy recommendations and open access sources, based on semantic analysis'. CORE-based CLIC is named as the key component for UNESCO's MOST programme in the mini-report to be submitted to UNESCO's Executive Board at its 192nd session to be held in Paris from September 24th to October 11th 2013. [5.4].

In recognition of CORE's impact on the open access movement, in June 2013 Jisc invited the CORE team to tender for the national metadata aggregator. This submission was successful and CORE has been selected to be the UK repository aggregator within the UK Repositories Shared Services Infrastructure.

5. Sources to corroborate the impact

[5.1] Technical Manager, The European Library

[5.2] Jisc News, 3 October 2011: 'UK's open access full-text search engine to aid research', http://www.jisc.ac.uk/news/stories/2011/09/openaccess.aspx

[5.3] Online PhD Program (2013) *TOP 100 Thesis & Dissertation References on the Web*, http://onlinephdprogram.org/thesis-dissertation/

[5.4] Chief, Sector for Social and Human Sciences, UNESCO

[5.5] Programme director, Jisc

[5.6] WSIS-10 (February 2013): http://core-project.kmi.open.ac.uk/files/CORE-UNESCO-submit.pptx

[5.7] UKCoRR, Leeds Metropolitan University

[5.8] McDonald, D. and Kelly, U. (2012) *The Value and Benefits of Text Mining*, Jisc Report, http://www.jisc.ac.uk/publications/reports/2012/value-and-benefits-of-text-mining.aspx

[5.9] INFOdocket (2012) *Search/Find Open Access Scholarly Articles, Reports Using New iOS Apps from CORE Project*, http://www.infodocket.com/2012/05/09/searchfind-open-access-scholarly-articles-reports-using-new-ios-apps-from-core-project/

[5.10] Centre for Research Communications, University of Nottingham