Systém CORE

Martin Lochman

Ústav informačních studií a knihovnictví FF UK v Praze

E-mail: martin.lochman@gmail.com

Záznam původního příspěvku:

LOCHMAN, Martin. Systém CORE. *Ikaros* [online]. Červenec 2014, roč. 18, č. 7 [cit. 2015-04-01]. ISSN 1212-5075. Dostupný z: http://ikaros.cz/system-core.

English title:

The CORE System

English abstract:

The article focuses on CORE (COnnecting REpositories), a unique system developed by the Knowledge Media Insitute that aggregates both metadata and content from Open Access repositories. Its primary objective is to facilitate free access to this content via various sophisticated tools and services. The first part of the text describes how the system works from a technical standpoint, focusing on the process of ingesting data as well as their exposure to users. The second section of the article is dedicated to thorough examination of search options and services. Finally, a brief evaluation of the entire system is offered.

Klíčová slova:

Informační služby, metadata, digitální archivy, přebírání záznamů, OAI-PMH

Keywords:

Information services, metadata, digital archives, metadata harvesting, OAI-PMH

1. Úvodní informace

Systém CORE (zkratka pro COnnecting REpositories) představuje významný projekt služby nad digitálními archivy celosvětového rozsahu, který sklízí a agreguje data z repozitářů a dalších typů úložišť s otevřeným přístupem. Jeho tvůrcem a provozovatelem je **Institut znalostních médií**, Otevřená univerzita (angl. Knowledge Media Institute, The Open University, http://kmi.open.ac.uk/).

Primárním cílem systému je agregovat obsah s otevřeným přístupem z různých archivů po celém světě, tento obsah obohatit a zajistit k němu přístup prostřednictvím dílčích služeb. Z toho ostatně plyne základní charakteristika, kterou se systém odlišuje od většiny ostatních poskytovatelů služeb – tedy fakt, že od spolupracujících archivů a repozitářů přejímá nejen metadatové záznamy, ale i kopie primárních dokumentů (v rámci procesu zvaného „cachování“), pokud existují, a buduje tak vlastní rozsáhlý repozitář plných textů. Uživatelé tak dostávají možnost tyto texty zpřístupnit přímo, bez nutnosti přesměrování do lokálních úložišť[1].

[1] *CORE* [online]. Milton Keynes (Velká Británie): Knowledge Media Institute, 2011- [cit. 2015-04-01]. Dostupný z: http://core.kmi.open.ac.uk/.

V současné době umožňuje webový portál systému CORE vyhledat přes 24 miliónů metadatových záznamů, z nichž je necelá desetina obohacena plnými texty, pocházejících z více než 650 digitálních repozitářů, což z něj činí nejrozsáhlejší fond plných textů s otevřeným přístupem na světě a mimo jiné jej také řadí mezi deset nejvýznamnějších vyhledávačů pro oblast vědy a výzkumu[2].

2. Vývoj

Prvotní verze systému CORE byla vytvořena na začátku roku 2011 odborníkem Institutu znalostních médií na poli získávání informací a zpracování přirozených jazyků, Petrem Knothem, s cílem usnadnit přístup a dolování z textů velkého množství vědeckých publikací. Již od počátku svého vývoje byl systém finančně podporován mnoha organizacemi a entitami včetně Evropské komise a Komise propojených informačních systémů JISC.

V dubnu 2011 byla spuštěna beta-verze systému, tehdy ještě pod názvem CORE Harvester system. Následovalo testování systému i jeho komponent včetně sklízení metadat a stahování plných textů. Projekt byl mimo jiné představen na Workshopu CERN pro inovace ve vědecké komunikaci OAI7 na Univerzitě v Ženevě (angl. CERN Workshop on Innovations in Scholarly Communication OAI7), který se konal ve dnech 22. – 24. června 2011.

3. Struktura

Srdcem systému CORE je model několikavrstvého agregačního informačního systému. Jednotlivé vrstvy jsou v jeho rámci hierarchicky uspořádány a provázány, přičemž platí, že každá z nich využívá funkce a zdroje nižší vrstvy a zároveň zprostředkuje funkce a zdroje pro vrstvu vyšší. Každá vrstva pracuje s informacemi přijatými od repozitářů na jiné úrovni abstrakce[3].

Nejnižší vrstvou je tzv. **Vrstva pro interoperabilitu přenosu metadat** (angl. The Metadata Transfer Interoperability Layer), která zajišťuje všechny procesy nezbytné pro sklízení metadat z repozitářů. Její klíčovou komponentou je protokol OAI-PMH.

Další vrstvu představuje tzv. **Vrstva metadat a obsahu** (angl. The Metadata and Content Layer), pro kterou jsou klíčové metadatové a obsahové komponenty. Vrstva zprostředkuje procesy potřebné k zajištění jejich ukládání, aktualizace a zpřístupňování. Metadatová komponenta pracuje s objekty ve formátech XML a RDF, popřípadě také ve schématu Dublin Core, obsahová komponenta pak s dokumenty v různých komerčních i otevřených formátech, včetně formátů PDF a DOC.

Další, hierarchicky nadřazená vrstva, je nazývána **Vrstvou obohacování** (angl. The Enrichment Layer). V ní probíhají procesy zpracování a harmonizace metadat a také jejich sémantického obohacování (o další informace a vztahy), které vychází z obsahové analýzy a zejména dolování dat z textů. Na rozdíl od nižších vrstev některé z přítomných procesů mohou být zcela automatizované, poloautomatizované nebo manuální.

---

[2] JACOBS, Neil a Rachel BRUCE. Ten search engines for researchers that go beyond Google. *JISC Inform* [online]. July 2013, issue 37, s. 7 [cit. 2015-04-01]. ISSN 1476-7619. Dostupný z: http://www.jisc.ac.uk/inform/inform37/SearchingBeyondGoogle.html.

[3] KNOTH, Petr a Zdenek ZDRAHAL. CORE: Three Access Levels to Underpin Open Access. *D-lib Magazine* [online]. November/December 2012, vol. 18, no. 11/12 [cit. 2014-06-06]. ISSN 1082-9873. Dostupný z: http://www.dlib.org/dlib/november12/knoth/11knoth.html.

Následující vrstvou je tzv. **Vrstva online transakčního a analytického zpracování** (angl. The OLTP and OLAP Layer), která zahrnuje dvě komponenty. První z nich, online transakční zpracování, je souborem funkcí, které zajišťují přístup k informacím a jejich následné zpracování na úrovni jednotlivých článků, popřípadě jejich menšího souboru. Komponenta online analytického zpracování na druhou stranu nabízí procesy pro podporu analýzy metadat a obsahu na úrovni celých sbírek.

Nejvyšší vrstvu představuje tzv. **Vrstva rozhraní** (angl. The Interface Layer), která je zodpovědná za zajištění komunikace systému CORE s koncovými uživateli a nabízí řadu dílčích služeb (o nich dále v podkapitole 4.).

Proces přijetí a uložení digitálního objektu do repozitáře CORE se skládá z několika dílčích a na sebe navazujících fází, které vycházejí z výše popsaného několikavrstvého modelu. V úvodní **agregační fázi** systém provede prostřednictvím příkazů protokolu OAI-PMH sklizeň metadatových záznamů z cílových repozitářů s otevřeným přístupem. Výsledkem úspěšné komunikace je dokument ve formátu XML, který vedle zmíněných metadat obsahuje také odkazy na plné texty dokumentů[4]. Systém tyto odkazy automaticky využije ke stažení plných textů. Následně proběhne extrakce prostého textu a konverze do formátu PDF. Systém CORE umožňuje simultánní sklízení a stahování obsahu z více digitálních úložišť, synchronizaci s registrem repozitářů OpenDOAR a mimo jiné také plánování a monitorování výše uvedených procesů.

Následující fáze **zpracování a sémantického obohacení** zahrnuje několik jednotlivých kroků, v jejichž jádru stojí algoritmy dolování dat z textů. Systém nejprve provede základní úkony tokenizace (rozdělení na jednotlivé slovní tvary[5]), filtrování, stemmingu a indexace přijatých metadatových záznamů a textů. Následují procesy zjišťování sémanticky podobných textů, které slouží nejen jako základ služeb pro uživatele – navigace, doporučení dalších dokumentů, a dalších – ale také k odhalení duplicity a případných plagiátů.

Dalším dílčím procesem této fáze je kategorizace obsahu. Jen velmi malé množství dokumentů pocházejících z archivů s otevřeným přístupem je totiž již před vstupem do systému CORE opatřeno nějakým typem třídníku. Vzhledem k nákladnosti a časové náročnosti manuální klasifikace jeho provozovatelé přistoupili k testování implementace algoritmů zcela automatizované plnotextové klasifikace s cílem množinu dokumentů rozčlenit do 18 hlavních tříd **pořádání DOAJ** (Directory of Open Access Journals, http://doaj.org/)[6].

Posledním z procesů této fáze je extrakce citací z plných textů publikací a jejich propojování s cílovými dokumenty (v případě, že se také nacházejí v repozitáři CORE). K tomuto účelu je využito balíčku programů s otevřeným zdrojovým kódem ParsCit[7]. Proces se podílí také na vytvoření a zpřístupnění propracovaných citačních sítí pro vědecké účely v rámci projektu **DiggiCORE** (Pronikání do propojených repozitářů, angl. Digging into Connected Repositories).

---

[4] Nicméně jen v případě, že je zdrojový archiv uvádí jako součást metadat.

[5] KŘEN, Michal. Srovnávací frekvenční seznamy. In: *Český národní korpus* [online]. Praha: Ústav Českého národního korpusu FF UK, 2000. Dostupné z: http://ucnk.ff.cuni.cz/srovnani10.php.

[6] Internetový adresář DOAJ, založený univerzitní knihovnou Lundské univerzity (http://www.lub.lu.se/) v roce 2003, nabízí hierarchický předmětový pořádací systém. Zahrnuje celkem 18 hlavních tříd, respektive tříd první úrovně, 75 tříd druhé úrovně a 38 tříd třetí úrovně.

[7] KAN, Min-Yen. ParsCit: An open-source CRF Reference String and Logical Document Structure Parsing Package. In: *WING* [online]. Singapur (Singapur): A/P Min-Yen KAN, ca 2000- [cit. 2015-04-01]. Dostupný z: http://aye.comp.nus.edu.sg/parsCit/.

Konečně poslední fází je **vystavení dokumentů a informací** v rámci služeb pro uživatele systému CORE.

4. Dílčí služby

Specifikem systému CORE je z hlediska výstupu a komunikace informací navenek tzv. **Tříúrovňový přístup** (angl. Three access levels architecture)[8]. Jednotlivé úrovně se navzájem odlišují především druhem dat, která zprostředkují, a z toho plynoucími cílovými skupinami uživatelů. V rámci každé úrovně jsou také nabízeny různé dílčí služby.

První z uvedených úrovní přístupu představuje tzv. **„Úroveň hrubých dat“** (angl. Raw data access). Primárními uživateli tohoto typu přístupu jsou vývojáři, specializovaní badatelé, digitální knihovny, softwarové firmy a další fyzické osoby a korporace, kteří mají zájem o hrubá metadata i vlastní obsah. Ta mohou být dále zpracována, obohacována a normalizována. Metadata a plné texty mohou být získány ve formě souborů ke stažení nebo skrze rozhraní API. Od toho se odvíjejí nabízené služby:

- **Rozhraní CORE pro programování aplikací** (angl. CORE API): Nabízena je dvojice rozhraní, které umožní externím systémům a službám přímou interakci se systémem CORE: **REST API**, které umožňuje vícehlediskové vyhledávání informací, stahování dokumentů ve formátu PDF nebo prostého textu i zjištění tématu odborného textu, a tzv. **SPARQL endpoint**, které poskytuje informace o sklizených datech a podobnostech kódované ve formátu RDF.
- **Soubor dat CORE** (angl. CORE data dump): Služba umožňuje zdarma stáhnout veškerá z repozitářů agregovaná data a metadata ve formě komprimovaných souborů dat (angl. dumps, konkrétním použitým formátem je .gz).

Druhou z úrovní přístupu je tzv. **„Úroveň transakčních informací“** (angl. Transaction information access). Zaměřena je zejména na uživatele z řad studentů, akademiků a obecně širší veřejnosti, kteří usilují o uspokojení svých informačních potřeb. Přístup k obsahu je tak typicky realizován prostřednictvím webového portálu CORE (**CORE portal**) a sady doplňkových služeb, ke kterým patří:

- **CORE pro mobilní platformy** (angl. CORE Mobile): Aplikace pro mobilní zařízení (bez ohledu na to, zda se jedná o chytré telefony nebo tablety) s operačním systémem IOS nebo Google Android nabízí stejné funkce jako plná desktopová verze portálu CORE včetně vyhledávání a navigace v databázi a stahování plných textů dokumentů.
- **Zásuvný modul CORE** (angl. CORE plugin): Nabízen je zásuvný modul, který je nezávislý na konkrétním typu platformy nebo prohlížeče. Plugin na základě sémantické podobnosti metadat nebo plného textu samotného poskytuje informace o spřízněných dokumentech k dokumentu, který si uživatel v rozhraní digitální knihovny nebo institucionálního repozitáře s otevřeným přístupem v daný moment prohlíží.

Poslední (třetí) úroveň přístupu je tzv. **„Úroveň analytických informací“** (angl. Analytical information access). Jejími hlavními uživateli jsou vláda, různorodé nadace, manažeři v podnicích i správci digitálních knihoven a repozitářů, kteří usilují o zjištění statistických

[8] KNOTH, Petr. CORE: Aggregation Use Cases for Open Access. In: *2nd International Workshop on Mining Scientific Publications (WOSP 2013)* [online]. 26 July 2013. Indianapolis (Indiana), The Joint Conference on Digital Libraries, 2013 [cit. 2015-04-01]. Elektronická kopie dostupná z: http://core-project.kmi.open.ac.uk/files/jcdl2013_v7.pdf.

informací na úrovni sbírek a podsbírek často v podobě přehledných grafů a tabulek. K tomu systém CORE nabízí dva sofistikované nástroje:

- **Analytika repozitářů** (angl. Repository analytics): Jedná se o službu, která umožňuje monitorovat proces příjmu metadat a vlastního obsahu ze spolupracujících repozitářů a uživatelům nabízí sadu statistických údajů týkajících se jak kvantitativních charakteristik – množství a velikosti dat – tak i charakteristik kvalitativních – přístupnosti, dostupnosti či platnosti metadat a primárních dokumentů.
- **„CORE Policy compliance analytics“:** Nástroj pro podporu implementace a monitorování britské politiky otevřeného přístupu **Rady pro financování vyššího vzdělání pro Anglii** (angl. Higher Education Funding Council for England, zkratka HEFCE, http://www.hefce.ac.uk/)[9]. Nástroj umožňuje zjistit, do jaké míry[10] se jednotlivé repozitáře ve Velké Británii touto politikou řídí.

5. Vyhledávání

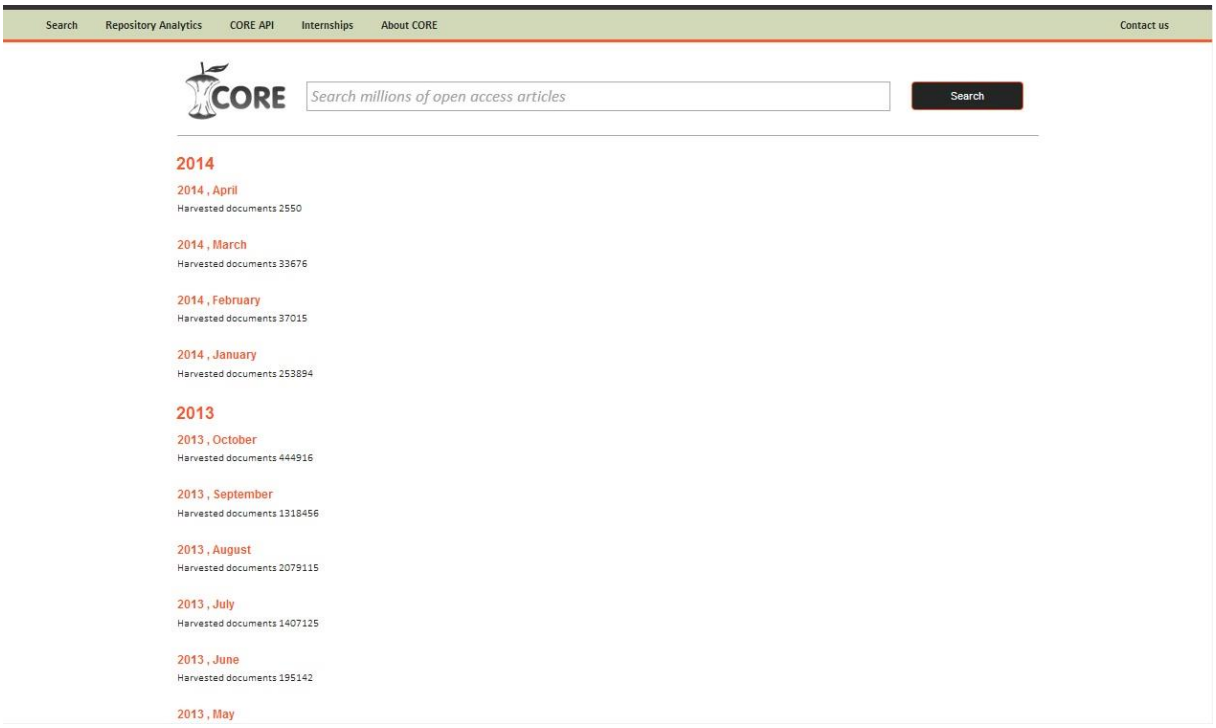
Vyhledávání metadatových záznamů a plných textů dokumentů, jsou-li k dispozici, je v rámci systému CORE možné realizovat klasickým **dotazovým vyhledáváním** pomocí vyhledávacího okna. Provozovatelé na hlavní stránce portálu také nabízejí alternativní možnost **prohlížení** podle data přidání, popřípadě prohlížení menší množiny nejnovějších přírůstků.

Režim **prohlížení** se vyznačuje velmi omezenou funkčností (viz obr. 1). Umožňuje sice zobrazit množiny záznamů na úrovni jednotlivých dnů, kdy byly do systému CORE přidány, vztahuje se však jen na období od února 2012 do současnosti, navíc je toto členění značně nekompletní. Součástí tohoto režimu je i výše uvedená funkce zobrazení nejnovějších přírůstků (angl. Latest additions), jejímž hlavním úskalím je z pohledu koncového uživatele nemožnost nabízenou množinu záznamů jakkoliv utřídit či uspořádat.

Dotazové vyhledávání je k dispozici v **jednoduché** i **pokročilé** podobě. Základ první z uvedených představuje jediné vyhledávací okno, do kterého lze zadat jednoslovný nebo víceslovný informační dotaz. Jednoduchý režim také dovede pracovat se syntaktickými prvky typu pravostranného/levostranného omezení, nahrazení jednotlivých znaků zástupným znakem či specifikací dotazu za použití znaku („-“). Možné je samozřejmě užití jednoho ze sady základních booleovských operátorů AND, OR nebo NOT.

[9] Policy Guide: Open access research. In: *Higher Education Funding Council for England (HECFE)* [online]. Bristol (Velká Británie): Higher Education Funding Council for England (HECFE), 1992- [cit. 2015-04-01]. Dostupný z: http://www.hefce.ac.uk/whatwedo/rsrch/rinfrastruct/oa/.

[10] Na základě analýzy dostupnosti všech publikací v repozitáři vypočítá tzv. míru dodržování (angl. Compliance Rate) v procentech. Sledována je také její proměnlivost v čase.



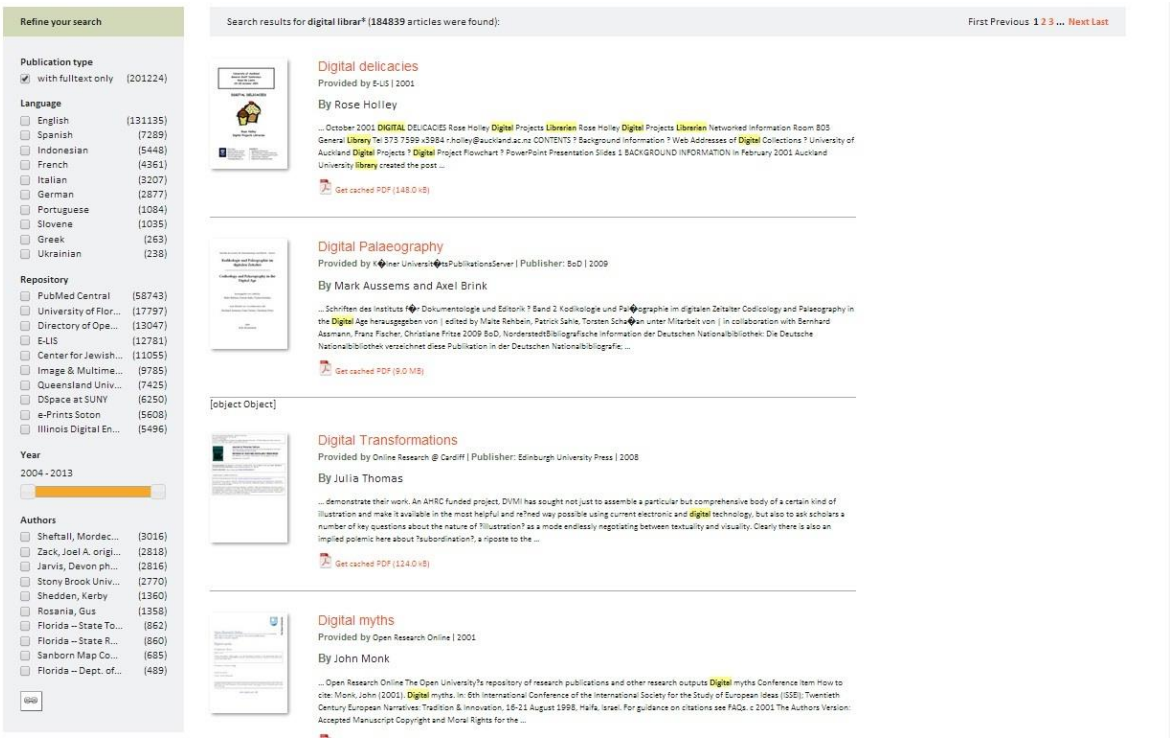
Obrázek 1 - Náhled režimu prohlížení

Pokročilá varianta (viz obr. 2) vyhledávání umožňuje využít až osm pevně definovaných selekčních polí, z nichž čtyři simulují použití výše uvedených booleovských operátorů a syntaktických prvků – pole „všechna slova“ (angl. all of the words), „přesná fráze“ (angl. exact phrase), „alespoň jeden z výrazů“ (angl. at least one of the words) a „bez těchto výrazů“ (angl. without the words). Jejich doplňkem je pak rozbalovací nabídka, která dává možnost použité informační dotazy aplikovat na vyhledávání v názvu, v názvu a abstraktu nebo přímo v plném textu samotného dokumentu. Prostřednictvím zbývajících čtyř polí lze upřesnit autora, vydavatele, repozitář, ze kterého byl záznam sklizen, a rok (popřípadě rozmezí let). Zajímavým specifikem – a zároveň nedostatkem – pokročilé varianty je fakt, že není přímo dostupná z hlavní stránky portálu, nýbrž se zobrazí jako možnost až poté, co je provedeno vyhledávání v jednoduchém režimu.

Obrázek 2 - Náhled rozhraní pokročilého vyhledávání

Poté, co systém vykoná základní vyhledávací proces nad indexem metadatové databáze CORE, nabídne množinu záznamů uspořádaných sestupně podle klesající relevance. Záznamy jsou zde zobrazeny ve zkrácené podobě: obsahují pouze následující prvky: název (který je zároveň vybaven hypertextovým odkazem vedoucím na úplný záznam), autor, typ dokumentu, rok publikace a zdrojový archiv. V případě, že je k dispozici také kopie plného textu, je zkrácený záznam opatřen stručným úryvkem, odkazem k jejímu stažení ve formátu PDF a náhledem titulní stránky.

Vedle výše popsaného pokročilého vyhledávání je s výslednou množinou záznamů možné dále pracovat prostřednictvím **faset** údajů, které jsou zobrazeny na levé straně obrazovky (viz obr. 3). Nabízené údaje jsou však ryze formální povahy, věcné hledisko zde zcela chybí. Pravděpodobně nejdůležitějším z nich je **typ publikace**, který umožňuje rešerši omezit pouze na záznamy opatřené plnými texty, nebo naopak rozšířit o ty, které jimi nedisponují. K dalším údajům patří **jazyk** dokumentu, **zdrojový repozitář** a **autor**. Možnost vyhledávání omezit z hlediska roku publikace reprezentuje interaktivní **časová osa**.

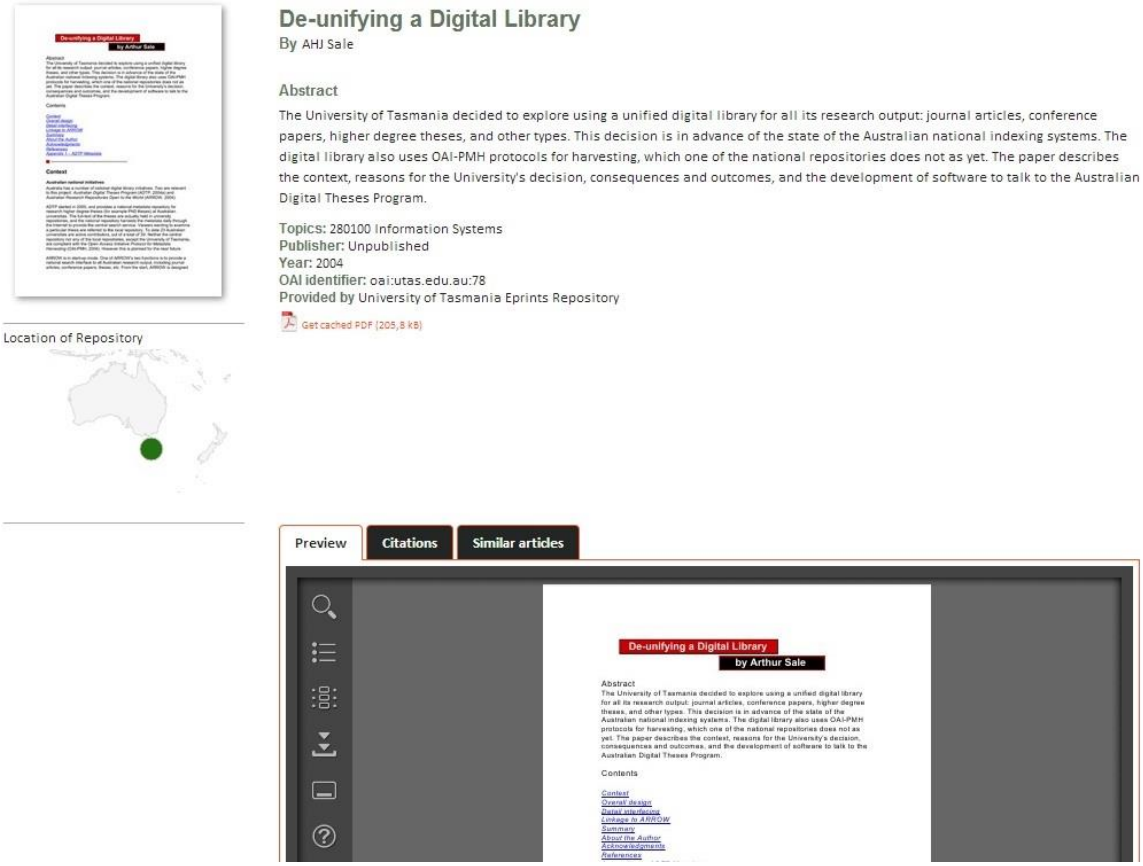


Obrázek 3 - Náhled množiny výsledků a faset údajů

Úplný záznam dokumentu nabízí kromě základní sady metadat (popsány v podkapitole 6.) náhled titulní stránky ve větším rozlišení než v případě zkráceného záznamu, výsek slepé mapy znázorňující lokaci zdrojového repozitáře a znovu také odkaz ke stažení plného textu ve formátu PDF (viz obr. 4). Ten si uživatelé mohou alternativně také v plném rozsahu prohlížet přímo v režimu online prostřednictvím vestavěné aplikace Multivio[11]. Opomenout nelze funkce „Podobné články“ (angl. Similar articles)[12], která nabízí stručné seznamy tematicky podobných textů a dokonce umožňuje graficky znázornit vztahy mezi nimi, a „Citace“ (angl. Citations), která zobrazí bibliografické záznamy použitých zdrojů převzaté z plného textu.

[11] *Multivio* [online]. Martigny (Švýcarsko): RERO: Réseau des Bibliothéques de Suisse occidentale, 2007- [cit. 2015-04-01]. Dostupný z: https://www.multivio.org/main/.

[12] Funkce vychází z výše popsané sémantické analýzy, která probíhá v rámci vstupních procesů dokumentů do systému.



Obrázek 4 - Ukázka metadatového záznamu australské provenience

6. Metadatový záznam

Systém CORE používá blíže nespecifikovaný metadatový formát. Vzhledem k rozmanitosti archivů a repozitářů, ze kterých sklízí metadata a případně i plné texty, délka jednotlivých záznamů kolísá. Přesto je možné identifikovat prvky, které jsou všem záznamům společné:

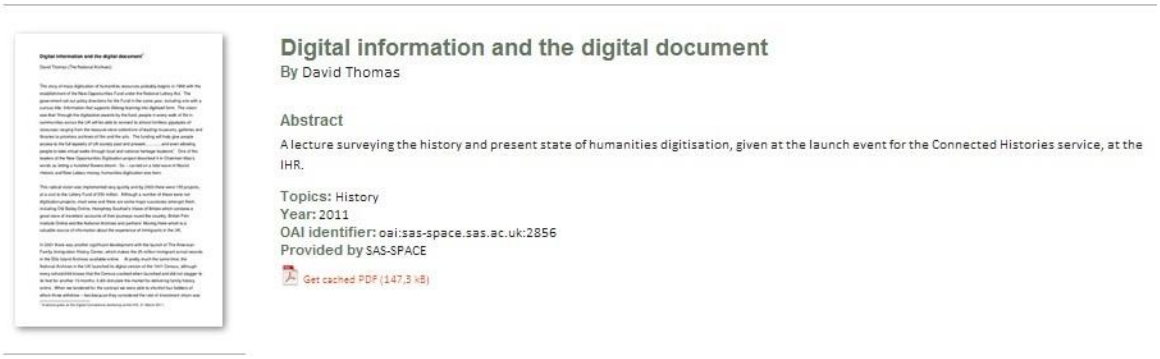
- **Název**: název, pod kterým je záznam v systému vyhledatelný
- **Autor:** fyzická osoba nesoucí primární odpovědnost za dílo
- **Rok:** rok, kdy byl primární dokument publikován
- **Identifikátor OAI**: jednoznačný a jedinečný identifikátor metadatového objektu ve zdrojovém repozitáři
- **Poskytovatel** (angl. „Provided by“): digitální archiv nebo repozitář, od kterého systém CORE získal metadata, popřípadě kopii plného textu dokumentu

Vedle uvedených pěti prvků je zapotřebí uvést několik dalších, jejichž výskyt závisí na kvalitě metadatového popisu jednotlivých zdrojových archivů:

- **Typ dokumentu**: údaj o typu primárního dokumentu, v některých případech doplněný o informaci, zda je recenzovaný či nikoliv
- **Abstrakt:** redukovaný text vztahující se k tematice a obsahu dokumentu
- **Předmět:** v rámci systému CORE prvek uveden jako „Témata“ – angl. „Topics“; obsahuje jednoslovná, popřípadě víceslovná označení vystihující obsah dokumentu.

Na základě průzkumu bylo možné určit, že v některých případech se jedná o autory přímo zprostředkovaná klíčová slova (viz obr. 5), v dalších případech je pak aplikováno Třídění Kongresové knihovny (viz obr. 6). U příspěvků pocházejících z repozitářů australské provenience se dokonce podařilo identifikovat užití Australské standardní vědecké klasifikace (angl. Australian Standard Research Classification).

- **Identifikátor DOI**: trvalý identifikátor informačního zdroje přidělený Mezinárodní nadací DOI
- **Nakladatel:** organizace nebo instituce, která dokument komerčně publikovala
- **Staženo z:** hypertextový odkaz vedoucí na elektronickou kopii dokumentu uloženou ve zdrojovém archivu



Obrázek 5 - Ukázka metadatového záznamu s autorským klíčovým slovem

Obrázek 6 - Ukázka metadatového záznamu s Tříděním Kongresové knihovny

7. Uživatelský přístup

Systém CORE svým uživatelům prozatím nenabízí možnost registrace a vytvoření osobního uživatelského účtu. Všechny nabízené služby spojené s vyhledáváním a získáváním obsahu i doplňkové nástroje jsou k dispozici bez jakýchkoliv omezení zdarma. V tomto ohledu se tedy systém výrazně odlišuje od většiny služeb nad digitálními archivy.

Systém mimo jiné také poskytuje službu signálního informování v podobě RSS kanálu. Pro účely komunikace s provozovatelem je uvedena e-mailová adresa.

8. Webové rozhraní

Webové rozhraní systému CORE je koncipováno přehledně a jednoduše (viz obr. 7). V jeho ústřední části se nachází vyhledávací okno jednoduchého vyhledávání, které tak umožňuje

přímý přístup do metadatové databáze CORE a zároveň podává aktuální informaci o celkovém počtu záznamů. V horní části hlavní stránky nad vyhledávacím rozhraním jsou k dispozici základní odkazy na **vyhledávání**, **nástroj pro analytiku repozitářů**, informace o **rozhraní CORE pro programování aplikací** a **stručnou charakteristiku** celého systému. Nechybí výše zmíněný odkaz na kontaktní informace provozovatele.

Prostor pod vyhledávacím oknem je vyplněn dalšími, graficky lépe zpracovanými odkazy na informace o **procesech**, které systém CORE používá k agregaci obsahů z jednotlivých repozitářů, **aplikaci CORE pro mobilní platformy**, **nástroji pro analytiku repozitářů** a **CORE rozhraní pro programování aplikací** (obsahují stejné hyperlinky jako výše uvedené odkazy).

Spodní část stránky je věnována stručným popisům systému samotného, jeho komponent a nabízených služeb, novinek a možností, jak se mohou další repozitáře a archivy s otevřeným přístupem zapojit do projektu, a neustále aktualizovanému seznamu nejnovějších článků z blogu CORE (http://core-project.kmi.open.ac.uk/blog).

Webové rozhraní v plném rozsahu včetně funkčních součástí je lokalizováno výhradně do anglického jazyka.

Obrázek 7 - Hlavní stránka webového rozhraní systému CORE

9. Statistiky

Tabulka č. 1 zprostředkuje dílčí statistické údaje o systému CORE. K jejich zjištění bylo využito rozhraní jednoduchého a pokročilého vyhledávání a přímé komunikace s provozovateli systému. Uvedená data jsou aktuální k 1. dubnu 2015.

K 1. dubnu 2015 umožňuje systém zpřístupnit až 24 271 248 metadatových záznamů, z nichž až 1,8 miliónů je opatřeno kopiemi plných textů dokumentů ve formátu PDF. Tyto záznamy pocházejí z 659 digitálních repozitářů a úložišť, z nichž největší zastoupení má CiteSeer X.

Nezanedbatelné množství záznamů pochází také z databáze DOAJ: Directory of Open Access Journals a známého preprintového archivu arXiv.org.

Vzhledem k faktu, že v množině zdrojových repozitářů převládají úložiště britské provenience, převládá z hlediska jazyka primárních dokumentů angličtina. Početně hojně zastoupené jsou také indonéština[13], španělština, francouzština a němčina.

Systém CORE agreguje data i z repozitářů a archivů české provenience. Největší počet záznamů pochází z digitálního repozitáře Vysoké školy báňské – Technické univerzity Ostrava.

Celkový počet záznamů v indexu	24 271 248
Celkový počet záznamů obohacených plnými texty	1,8 miliónů
Celkový počet zdrojových archivů	659
Počet záznamů pocházejících z České republiky	129 771
Převládající druh dokumentu	Článek (výhradně)
Převládající rok publikování dokumentu	2013 (1 510 314)
Převládající jazyk dokumentu	Angličtina
Zdrojový archiv s nejvyšším počtem poskytnutých záznamů	CiteSeer X
Zdrojový archiv s nejvyšším počtem poskytnutých záznamů české provenience	DSpace VŠB-TUO
Země s nejvyšším počtem poskytnutých záznamů	Velká Británie

Tabulka 1 - Statistické údaje o systému CORE k 1. dubnu 2015

10.Vyhodnocení

Systém CORE představuje významný počin v oblasti vývoje vyhledávačů a služeb působících nad digitálními archivy a repozitáři. Vedle archivů britské provenience agreguje data také z úložišť z dalších zemí, jeho význam pro vědeckou komunitu je tedy celosvětový. Systém je navíc založen na principech otevřeného přístupu, obsahy vysoké informační hodnoty nabízí bez omezení uživatelům ze všech zemí.

Nezpochybnitelnou předností systému CORE je fakt, že neagreguje jen metadatové záznamy ale i kopie příslušných plných textů. Dále lze vyzdvihnout také kvantitativní charakteristiky: celkovou velikost indexu čítající přes 24 miliónů záznamů, vysoké množství plných textů dokumentů, počet a rozmanitost zdrojových archivů. Význam nesou i rysy kvalitativní, zejména pestrý výběr nabízených služeb a doplňkových nástrojů pro individuální i institucionální uživatele, kvalitně pojaté procesy vstupu dat do systému a v neposlední řadě také přehledně navržené uživatelské rozhraní.

Vzhledem k tomu, že systém dosud prochází neustálým vývojem, je nutné zmínit také několik podstatných nedostatků. Tím nejvážnějším se jeví vyhledávání informací. Režim prohlížení nabízí velmi omezené možnosti a zvolenou množinu záznamů neumožňuje nijak utřídit.

[13] Jde o nahodilý úkaz, jehož důvodem je sklízení velkého množství záznamů a plných textů z repozitáře indonéské univerzity v Diponegoro (angl. Diponegoro University Institutional Repository).

V případě dotazového vyhledávání sice uživatel má možnost s nalezenými záznamy dále pracovat prostřednictvím faset údajů, tyto jsou však výhradně formálního charakteru. Věcné vyhledávání také do značné míry komplikuje různorodost věcného pořádání metadatových záznamů. Problematické je mimo jiné i rozhraní pokročilého vyhledávání, které je možné zpřístupnit až po provedení vyhledávacího procesu pomocí jednoduchého vyhledávání – uživatel je tedy nucen provést zbytečný krok navíc.

Literatura

CORE [online]. Milton Keynes (Velká Británie): Knowledge Media Institute, 2011- [cit. 2015-04-01]. Dostupný z: http://core.kmi.open.ac.uk/.

JACOBS, Neil a Rachel BRUCE. Ten search engines for researchers that go beyond Google. *JISC Inform* [online]. July 2013, issue 37, s. 7 [cit. 2015-04-01]. ISSN 1476-7619. Dostupný z: http://www.jisc.ac.uk/inform/inform37/SearchingBeyondGoogle.html.

KAN, Min-Yen. ParsCit: An open-source CRF Reference String and Logical Document Structure Parsing Package. In: *WING* [online]. Singapur (Singapur): A/P Min-Yen KAN, ca 2000- [cit. 2015-04-01]. Dostupný z: http://aye.comp.nus.edu.sg/parsCit/.

KNOTH, Petr a Zdenek ZDRAHAL. CORE: Three Access Levels to Underpin Open Access. *D-lib Magazine* [online]. November/December 2012, vol. 18, no. 11/12 [cit. 2014-06-06]. ISSN 1082-9873. Dostupný z: http://www.dlib.org/dlib/november12/knoth/11knoth.html.

KNOTH, Petr. CORE: Aggregation Use Cases for Open Access. In: *2nd International Workshop on Mining Scientific Publications (WOSP 2013)* [online]. 26 July 2013. Indianapolis (Indiana), The Joint Conference on Digital Libraries, 2013 [cit. 2015-04-01]. Elektronická kopie dostupná z: http://core-project.kmi.open.ac.uk/files/jcdl2013_v7.pdf.

KŘEN, Michal. Srovnávací frekvenční seznamy. In: *Český národní korpus* [online]. Praha: Ústav Českého národního korpusu FF UK, 2000. Dostupné z: http://ucnk.ff.cuni.cz/srovnani10.php.

Multivio [online]. Martigny (Švýcarsko): RERO: Réseau des Bibliothéques de Suisse occidentale, 2007- [cit. 2015-04-01]. Dostupný z: https://www.multivio.org/main/.

Policy Guide: Open access research. In: *Higher Education Funding Council for England (HECFE)* [online]. Bristol (Velká Británie): Higher Education Funding Council for England (HECFE), 1992- [cit. 2015-04-01]. Dostupný z: http://www.hefce.ac.uk/whatwedo/rsrch/rinfrastruct/oa/.