Masarykova univerzita, Fakulta informatiky

BASE

Bielefeld Academic Search Engine

Martin Šmíd, 422633
12.12.2015

Úvod

BASE je jedním z nejrozsáhlejších vyhledávačů zaměřených na akademické práce, které jsou volně přístupné na internetu. Projekt je veden v rámci knihovny Bielefeldské univerzity v Německu. Potřeba publikovat dokumenty způsobem, který povede k jednoduchému šíření a následné archivaci, vedla výzkumné instituce k zapojení do Iniciativy otevřených archivů (Open Archives Initiative, OAI). Obsah archivů je nabízen do světa pomocí protokolu pro sběr metadat (OAI-PMH). Projekt BASE se zabývá sběrem, normalizací a indexováním těchto dat, přičemž se zaměřuje na vědecký obsah institucionálních i předmětových repozitářů.

OAI

Iniciativa otevřených archivů (Open Archives Initiative) stojí za vývojem a vydáním standardů pro zlepšení interoperability mezi digitálními archivy. Původní záměr byl zpřístupnit vědecké práce, protože většina institucí přispívajících do vývoje otevřených archivů se zabývala publikováním akademických dokumentů v elektronické podobě (e-printů), nicméně se počítá i s výměnou jiných digitálních dokumentů. Přístup k jednotlivým e-printům se výrazně zjednodušuje, protože lze sklízet metadata z repozitářů a nad sklizenými metadaty lze budovat služby pro uživatele (například vyhledávání).

OAI-PMH

OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) je protokol pro sklízení metadatových záznamů z digitálních repozitářů. Poskytuje technický prostředek pro poskytovatele dat, aby mohli zpřístupnit metadata ze svých archivů poskytovatelům služeb. Poskytované služby jsou založeny na rozšířených standardech internetového protokolu HTTP a značkovacího metajazyka XML.

Účelem OAI-PMH je sběr metadat do jediného místa. Na straně poskytovatele dat musí být podporován minimálně nekvalifikovaný Dublin Core, ačkoli je možné současně s ním podporovat i další metadatová schémata. Služby, které přímo nesouvisejí se sběrem metadat, musejí být poskytnuty dalšími prostředky. Mezi tyto služby, které lze stavět nad získanými metadaty, je například federativní vyhledávání. Jeho prostřednictvím se uživatelé systému snadněji dostanou k požadovaným dokumentům.

BASE

Historie BASE

Práce na projektu, který zatím předcházel BASE, byly silně ovlivněny úspěchem projektu *Digital Library NRW*[1] (DigiBib NRW), na kterém se podílela knihovna Bielefeldské univerzity v letech 1998 až 2000. Digital Library NRW fungoval jako vyhledávač v knihovních katalozích.[2] Snažili se vytvořit systém, jenž by přesahoval knihovní katalogy a dokázal získat informace i z jiných internetových, veřejně dostupných zdrojů.

V červnu roku 2004 proběhlo veřejné spuštění nového projektu, jenž měl demonstrovat sběr metadat kolekcí digitálních dokumentů, nyní již pod názvem Bielefeld Academic Search Engine (BASE) [1]. BASE obsahoval při spuštění přibližně 600 000 dokumentů, jež byly získány z 15 zdrojů. BASE byl vyvíjen za účelem vyhledávání a zpřístupnění kvalitních akademických prací. Dokumenty s otevřeným přístupem se nacházejí na různých místech, aniž by existoval centrální bod, který by udržoval informace o jejich stavech a přístupnosti. Tuto úlohu se snaží BASE vyřešit a přebírá úlohu centrálního bodu, ze kterého je možné snadno vyhledávat dokumenty na základě sklizených metadat s možností přímého přístupu k původním dokumentům.

Zdroje dokumentů

Je nutné zajistit výběr vhodných archivů s velkým množstvím dokumentů, u kterých se lze spolehnout na kvalitu obsahu.

Pro vyhledání vhodných repozitářů probíhá sledování záznamů např. OpenArchives, ROAR, Eprints, DSpace, OpenDOAR (The Directory of Open Access Repositories) [2, 3]. Každý zdroj nabízí různé množství zahrnutých serverů s dokumenty. Liší se však také spolehlivostí a kvalitou nabízených dat.

Současný stav

V současné době je v BASE indexováno přes 81 mil. dokumentů z necelých 3900 serverů (repozitářů). Rozšiřování databáze sesbíraných metadat probíhá téměř denně. Jedná se o sběr dat ze serverů, které byly označeny za vhodné pro zpracování, ale i o zavedení nových repozitářů, ze kterých se budou metadata získávat.

---

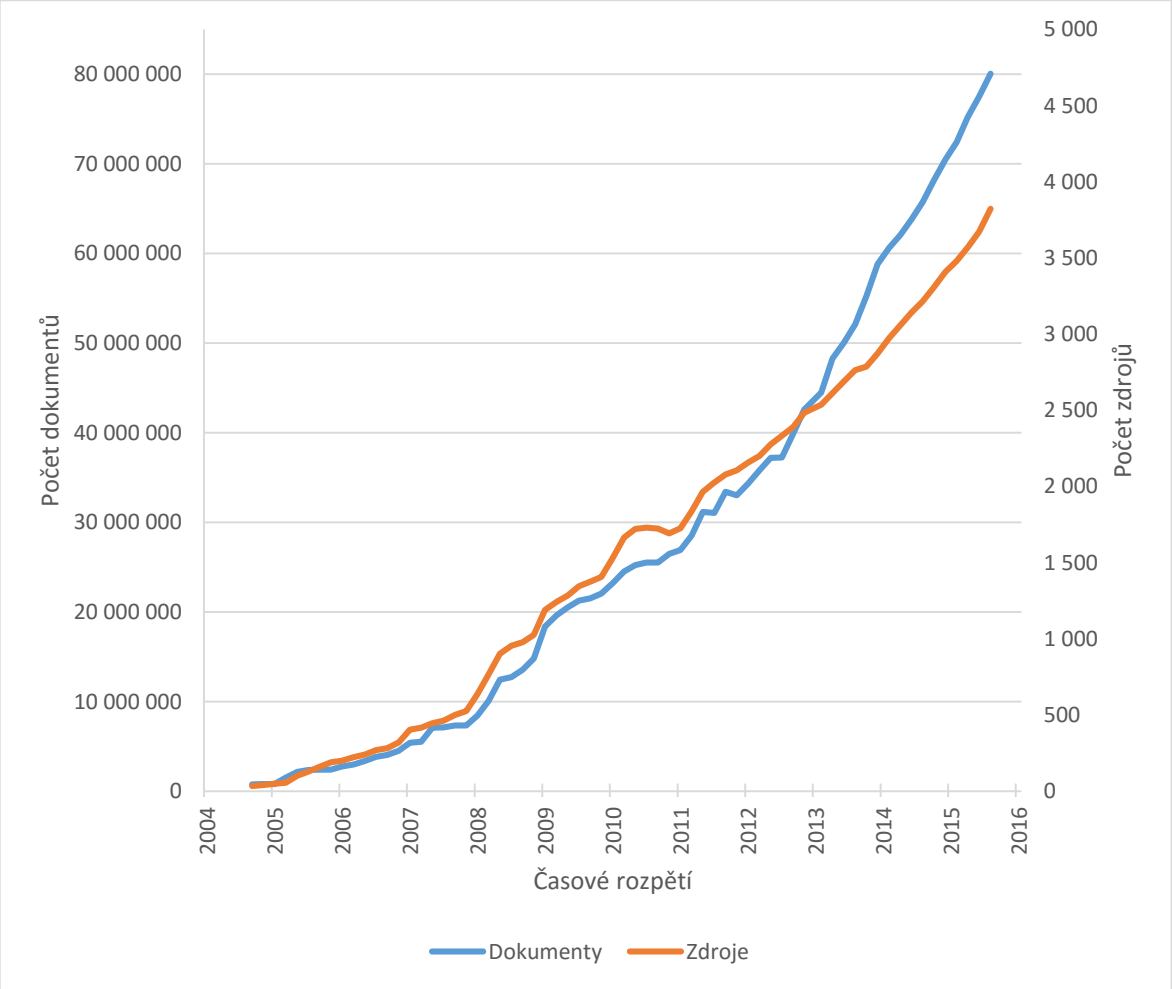
[1] Digitální knihovna Severního Porýní-Vestfálska

[2] V roce 2000 byl název změněn na DigiBib. Oficiální stránky projektu: https://www.digibib.net

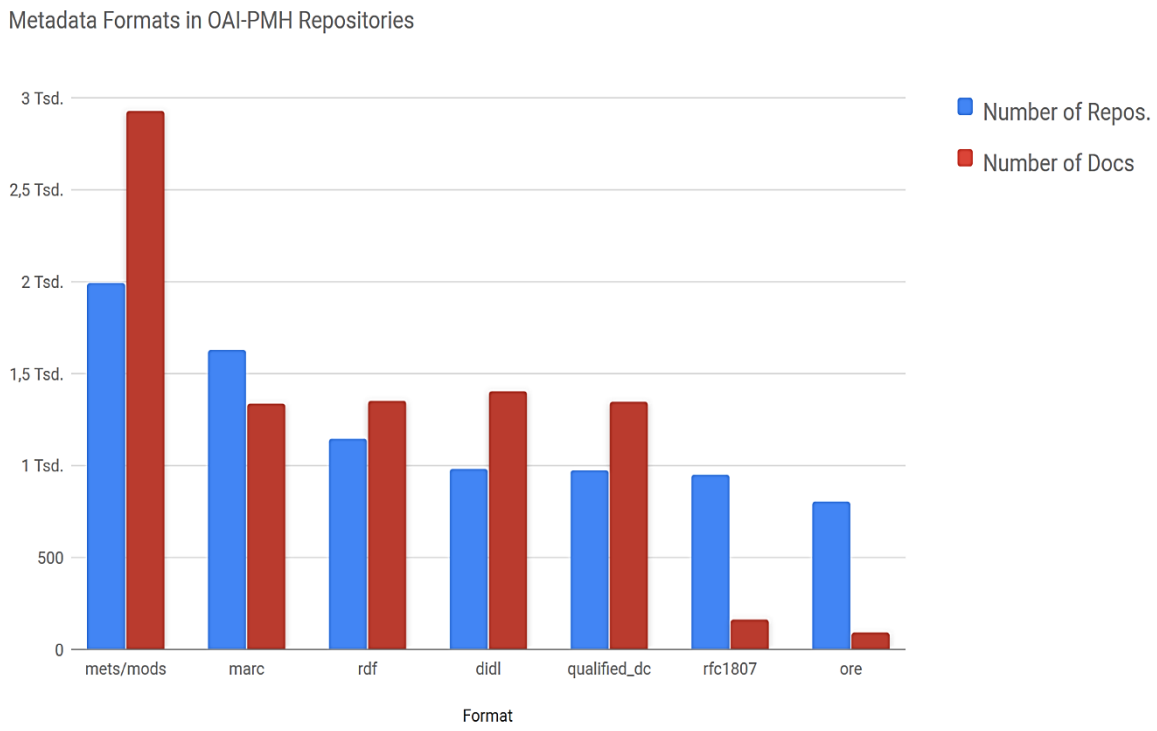
Na stránkách projektu BASE je přehledně vystavený seznam všech indexovaných repozitářů. Uvedená statistika dovoluje uživateli prohlédnout si, ze kterých zemí repozitáře pocházejí a v jakém jazyce jsou dokumenty napsány. Momentálně se lze prostřednictvím služeb BASE setkat s dokumenty ve 108 různých jazycích [4].

Z českých repozitářů to jsou např. Národní technická knihovna, digitální knihovna ČVUT v Praze, DML-CZ nebo DSpace Masarykovy univerzity v Brně a veřejné služby jejího Informačního systému.

Protokol OAI-PMH vyžaduje podporu nekvalifikovaného DC. Při sběru metadat je možné se setkat i s dalšími metadatovými formáty, jako jsou kvalifikovaný DC, formáty z rodiny standardů MARC nebo RDF. Přehled dalších formátů ukazuje obrázek 2, kde je znázorněn počet zaindexovaných repozitářů v tisících a druhy metadatových formátů.



Obrázek 1: Počet indexovaných dokumentů a repozitářů [5]

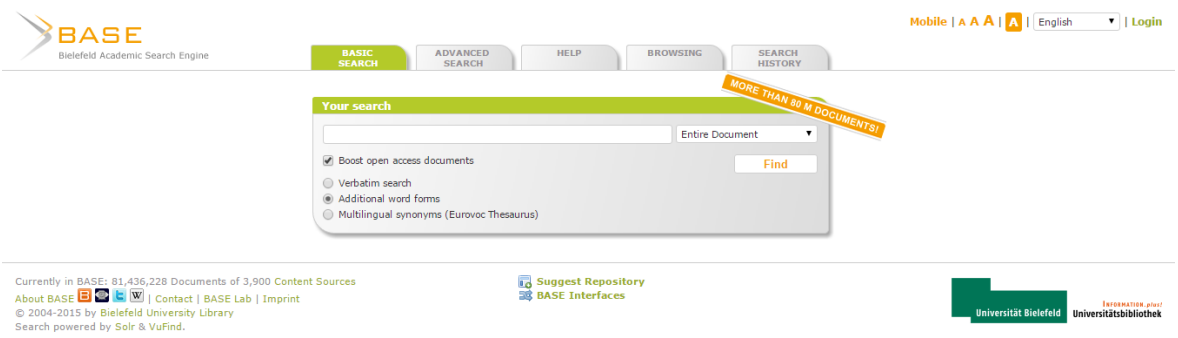


Obrázek 2: Metadatová schémata dokumentů [6]

Vyhledávání v BASE

Obrázek 3 ukazuje stránku se základním vyhledáváním v sytému BASE. Je možné pozorovat, že se tvůrci snaží o jednoduché prostředí, které není přeplněné nepotřebnými prvky. BASE nabízí základní a rozšířené vyhledávání pro specifičtější potřeby uživatele.

Vyhledává se pouze v metadatech indexovaných souborů, kvůli dodání výsledku v rozumném čase a kvůli náročnosti na výkon. Uživateli je umožněno vybrat, ve kterých polích chce vyhledávat. Svůj dotaz může hledat ve všech dostupných metadatech dokumentu, v názvu dokumentu, podle jména autora nebo předmětu zařazení dokumentu.

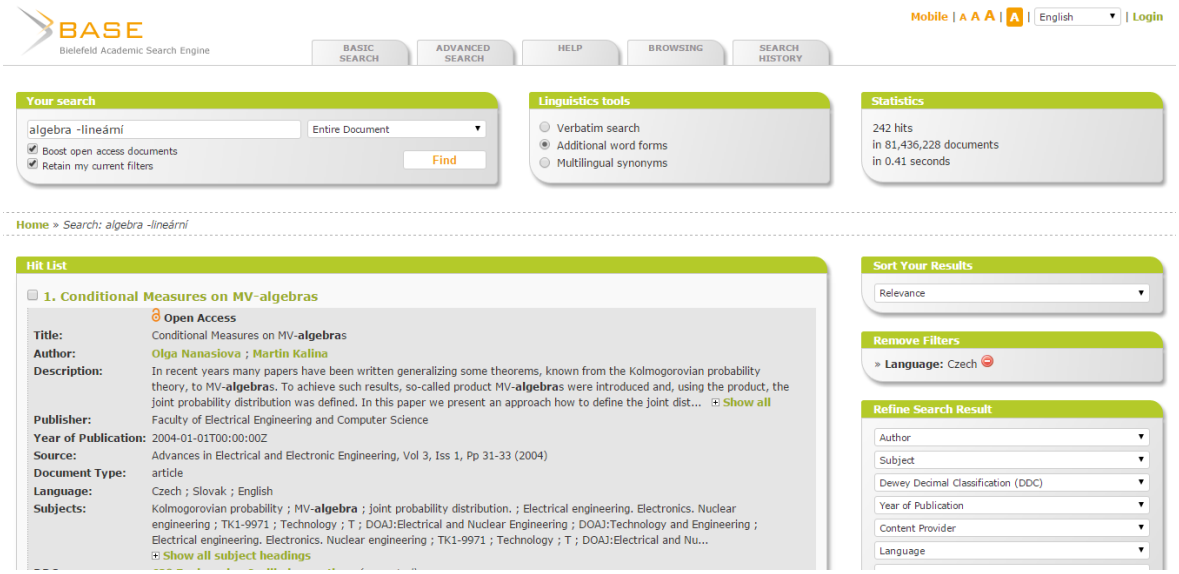


Obrázek 3: Úvodní stránka vyhledávání v BASE

Zadáním více hledaných výrazů lze kombinovat způsoby vyhledávání. Základní interpretace termů je jejich konjunkce. Po nalezení výsledků je uživateli umožněno upravovat zobrazení výsledných dokumentů. Nabízí se možnosti, jako jsou řazení dokumentů nebo filtrování podle různých kritérií. Filtrovat je možné podle autora, předmětu, Deweyova desetinného systému klasifikace, roku publikování, jazyku publikace aj.

Hledáme	**Příklad**	**Počet výskytů**
A **a** B	lineární algebra	67
A a B jako **fráze**	"lineární algebra"	30
A **nebo** B	(lineární algebra)	355 783
A a B **nebo** A a C **nebo** A, B a C	algebra (lineární numerická)	74
A **bez** B	algebra -lineární	354 699

Tabulka 1: Příklad složeného hledaného výrazu



Obrázek 4: Výsledky hledání pro výraz "algebra -lineární"

Pro zkušeného uživatele je připraveno i rozšířené hledání, ve kterém mu je umožněno upřesnit svůj dotaz ještě před položením dotazu. Kromě jednoho dotazovacího pole se mu tak naskytne příležitost vyplnit více polí, ve kterých si vybere možnosti, jakým způsobem chce ovlivnit svoje hledání.

Obrázek 5: Rozšířené hledání v BASE

Závěr

BASE umožňuje pohodlné vyhledávání nad sklizenými metadaty. Do repozitářů s akademickými pracemi lze proto přistupovat z jednoho místa, aniž by byl uživatel nucen prohledávat každý repozitář zvlášť. Projekt se neustále vyvíjí a rozšiřuje. Oproti Google Scholar, který taktéž nabízí přístup k obrovskému počtu dostupných dokumentů, se BASE snaží cílit na repozitáře s otevřeným přístupem. Může tedy motivovat k tomu, aby se lidé nebáli publikovat do repozitářů. Pokud si vyberou jeden z větších OA repozitářů, je poté téměř jisté, že se k jejich práci snadno dostanou i ostatní právě skrze BASE.

Použité zdroje

[1] SUMMANN, Friedrich a Norbert LOSSAU. Search Engine Technology and Digital Libraries: Moving from Theory to Practice. *D-Lib Magazine* [online]. 2004, roč. 10, č. 9 [vid. 1. prosinec 2015]. ISSN 1082-9873. Dostupné z: doi:10.1045/september2004-lossau

[2] PIEPER, Dirk a Friedrich SUMMANN. Bielefeld Academic Search Engine (BASE): An end-user oriented institutional repository search service. *Library Hi Tech* [online]. 2006, roč. 24, č. 4, s. 614–619 [vid. 2. prosinec 2015]. ISSN 0737-8831. Dostupné z: doi:10.1108/07378830610715473

[3] Bielefeld University Library. *BASE - Bielefeld Academic Search Engine | FAQ* [online]. c2015 [vid. 12. prosinec 2015]. Dostupné z: http://www.base-search.net/about/en/faq.php

[4] Bielefeld University Library. *BASE - Bielefeld Academic Search Engine | Content Sources* [online]. c2015 [vid. 2. prosinec 2015]. Dostupné z: http://www.base-search.net/about/en/about_sources_date_dn.php?menu=2

[5] Bielefeld University Library. *BASE - Bielefeld Academic Search Engine | Statistics* [online]. c2015 [vid. 2. prosinec 2015]. Dostupné z: http://www.base-search.net/about/en/about_statistics.php?menu=2

[6] PIEPER, Dirk a Friedrich SUMMANN. 10 years of „Bielefeld Academic Search Engine“ (BASE): Looking at the past and future of the world wide repository landscape from a service providers perspective. In: [online]. 2015 [vid. 1. prosinec 2015]. Dostupné z: http://pub.uni-bielefeld.de/publication/2766308

[7] Bielefeld University Library. *BASE - Bielefeld Academic Search Engine* [online]. c2015 [vid. 1. prosinec 2015]. Dostupné z: http://www.base-search.net/about/en/index.php

[8] HANOUSEK, Tomáš. *OAI-PMH pro začátečníky* [online]. B.m.: Národní archiv. 2009 [vid. 1. prosinec 2015]. Dostupné z: http://www.nacr.cz/Z-files/moznosti_06.pdf

Metadata DC

<?xml version="1.0" encoding="UTF-8"?>

<metadata
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:dc="http://purl.org/dc/elements/1.1/">
<dc:title>BASE: Bielefeld Academic Search Engine</dc:title>
<dc:creator>Martin Šmíd</dc:creator>
<dc:subject>OAI</dc:subject>
<dc:subject>OAI-PMH</dc:subject>
<dc:subject>BASE</dc:subject>
<dc:subject>Bielefeld Academic Search Engine</dc:subject>
<dc:description>BASE se zabývá sběrem, normalizací a indexováním metadat. Zaměřuje se především na OA repozitáře s akademickými pracemi.</dc:description>
<dc:date>2015-12-12</dc:date>
<dc:type>Text</dc:type>
<dc:format>text/docx</dc:format>
<dc:language>cs</dc:language>

</metadata>