Invenio

Jaroslav Čechák
učo 410322
třetí ročník bakalářského studia
Fakulta informatiky Masarykovy univerzity

6. prosince 2014

1 Co je Invenio

Invenio je svobodný systém s veřejně dostupným zdrojovým kódem, který poskytuje nástroje pro provoz digitální knihovny a dokumentového archivu. Systém je modulární a skládá se z mnoha menší balíků, kde každý z nich řeší pouze část funkcionality celého systému. Dohromady tvoří Invenio prostředek, který umožňuje různorodé elektronické dokumenty přidávat, zařazovat a indexovat, spravovat a nabízet koncovým uživatelům.

2 Stav projektu

Projekt Invenio byl vyvinut výzkumným centrem *Conseil Européen pour la Recherche Nucléaire* (CERN) v rámci snahy o jednoduchý, efektivní a veřejný přístup k výsledkům výzkumu a publikacím tohoto centra. Výsledkem této snahy byla v roce 2006 nový vzhled *CERN Document Server* (CDS), což byla produkční podoba systému CDS Invenio ve verzi 0.90.0 [1]. Od počátku byl projekt veden jako otevřený a veškerý kód je dostupný pod licencí *GNU*

General Public License v2.0 (GNU GPLv2). Tato licence umožňuje provádět libovolné úpravy a/nebo kód dále předávat, a to vždy pod svobodnou licencí GNU GPLv2, nebo novější a se zachováním informací o původních autorech [2].

V současné době je na stránkách projektu (`http://invenio-software.org/`) k dispozici stabilní verze Invenio 1.1.4 z 31. srpna 2014 [3]. Tato verze se skládá z 39 modulárních částí psaných z velké části v programovacím jazyce Python. Na jejím vývoji se kromě CERNu podílí také instituce Deutsches Elektronen-Synchrotron (DESY), École polytechnique fédérale de Lausanne, (EPFL), Fermilab (FNAL) a Stanford Linear Accelerator Center (SLAC) National Acceleration Laboratory. Pro svoji digitální knihovnu systém Invenio využívá 31 institucí. Jako zajímavé příklady projektů bych zde rád uvedl kromě výše uvedeného CDS také INSPIRE – informační systém o vysokoenergetické fyzice, Aristotle University of Thessaloniki – knihovna řecké univerzity a Zenodo – webový archiv dat pořízených během nebo za účely výzkumu [4]. Právě CDS, který spravuje více než milion bibliografických záznamů zahrnujících mimo jiné články, knihy, časopisy, fotografie a videa [5].

3 Okénko dovnitř programu Invenio

3.1 Vyhledávání

Mezi hlavní vlastnosti jistě patří vyladěný vyhledávač, který spolu se speciálně navrženými indexy, umožňuje velmi rychle (podobně jako Google) vyhledávat až nad pěti miliony záznamů. Dokáže kombinovat vyhledávání na základě metadat, seznamů zdrojů i fulltextového dotazu. Využilo se zde především předpokladu, že databáze digitální knihovny potřebuje zvládat velké počty hledání, ale změny provádí pouze ojedinělé. Vše, co je možné uchovávat v mezipaměti, se v ní uloží. Dále byly vybudovány rozsáhlé indexy a to jak dopředné, kde jsou pro každé slovo poznačeny dokumenty,

v nichž se vyskytuje, tak zpětné, kde se naopak uchovávají slova, která se v daném záznamu vyskytují. Systém umožňuje vyhledávat jednotlivá slova, stejně jako celé fráze a dokonce i regulární výrazy.

Zde je ukázka průběhu vyhledávání pro dotaz „Ellis AND moun IN Theses“. Parser dotazů rozdělí dotaz na čtyři části, a to: slova „Ellis“ a „moun“, logickou spojku „AND“ a označení kolekce „Thesis“. Ke slovům „Ellis“ a „moun“ se z indexů naleznou příslušně seznamy výskytů, které před dalším zpracováním musejí projít dekompresí. Plně rozvinuté seznamy výskytů putují do vyhodnocovače booleovských dotazů, kde se provede jejich průnik, protože v dotazu byla spojka „AND“. Shlukovač mezitím paralelně vyhledá univerzum kolekce „Thesis“ a dojde k poslednímu průniku univerza a společného seznamu výskytů. Takto vzniklé finální výsledky hledání se setřídí a na základě šablony naformátují do výsledné webové stránky, která se prezentuje uživateli.

Původní implementace z roku 2002 [6] využívala datový typ `Numeric` v jazyce Python pro reprezentaci binárních vektorů, nad kterými se následně dělaly pomocí bitových operací průniky a sjednocení. V roce 2007 [6] došlo k optimalizace paměťové náročnosti napsáním rozšíření v programovacím jazyce C, které dovolovalo adresovat na úrovni bitů oproti bajtům v případě typu `Numeric`. V roce 2011 došlo k propojení systému s externími nástroji na získávání informací Solr[1] a Xapian[2] [6]. Ty slouží zejména k indexování záznamů v databázi a k řazení nalezených výsledků [7].

3.2 Standardy

Systém Invenio používá pro svoji činnost dva základní standardy z oblasti digitálních knihoven. Těmi jsou *MAchine-Readable Cataloging* (MARC)[3] a *Open Archives Initiative Protocol for Metadata Harvesting* (OAI-PMH)[4].

[1] `http://lucene.apache.org/solr/`

[2] `http://xapian.org/`

[3] `http://www.loc.gov/marc/bibliographic/`

[4] `http://www.openarchives.org/pmh/`

3.2.1 MARC

Díky standardu MARC 21 je možné popsat podrobnými metadaty i velice různorodé dokumenty, od článků až po videa. MARC je již dlouhou dobu považován za zavedený a i přes stáří svého vzniku stále dokáže poskytnout možnost, jak digitální objekt obstojně popsat. Pokud je potřeba s metadaty pracovat mimo systém skrze soubor, využívá se MARC XML[5] [8].

3.2.2 OAI-PMH

Standard OAI-PMH zajišťuje výměnu bibliografických metadat s ostatními archivy jako je například arXiv.org. Invenio metadata nejenom nabízí, ale i přijímá a indexuje, aby v nich bylo možné vyhledávat.

4 Zhodnocení

Projekt Invenio je podle mého názoru velkým přínosem do světa digitálních knihoven zejména díky své otevřenosti, modularitě a použitelnosti. Zároveň se jedná o projekt se stále aktivním vývojem [**blog**], což dovoluje reagovat na nové výzvy a problémy v oblasti digitálních knihoven. Svůj podíl na aktivním vývoji mají i studenti, kteří zpracovávají některé nové funkcionality jako své závěrečné práce, což hodnotím kladně. Jednak se tímto způsobem zaručuje vývoj systému, ale také se zvedá povědomí o problematice digitálních knihoven.

Instalace a nasazení systému v základním nastavení je velice jednoduchá a rychlá. Sám jsem zprovoznil vlastní server během necelé půl hodiny. Invenio při instalaci vyžaduje jen minimální množství závislostí. S ohledem na rozsah systému mě taková míra samostatnosti zaskočila.

Dostupnost zdrojového kódu a zvolená licence umožňuje libovolný zásah a úpravu podle individuálních požadavků. Navíc programovací jazyk Python byl vybrán i z důvodu snadnější čitelnosti kódu oproti jiným

[5] `http://www.loc.gov/standards/marcxml/`

jazykům. Díky modularitě je možné v budoucnu vyměnit některé prvky za novější a lepší, aniž by tím být výrazně narušen chod celého systému.

Odkazy

1 ŠIMKO, Tibor. *CDS Invenio v0.90.0 is released* [online]. 2006 [cit. 2014-12-05]. Dostupný z WWW: ⟨https://raw.githubusercontent.com/inveniosoftware/invenio/v0.90.0/RELEASE-NOTES⟩.

2 FREE SOFTWARE FOUNDATION, Inc. *GNU General Public License* [online]. 2007 [cit. 2014-12-05]. Dostupný z WWW: ⟨http://www.gnu.org/licenses/gpl-2.0.html⟩.

3 ŠIMKO, Tibor. *GNU General Public License* [online]. 2014 [cit. 2014-12-05]. Dostupný z WWW: ⟨http://invenio-software.org/blog/invenio-1.1.4⟩.

4 CERN. *General/Demo – Invenio* [online]. 2014 [cit. 2014-12-05]. Dostupný z WWW: ⟨http://invenio-software.org/wiki/General/Demo⟩.

5 CERN. *Invenio* [online]. 2014 [cit. 2014-12-05]. Dostupný z WWW: ⟨http://invenio-software.org⟩.

6 ŠIMKO, Tibor. *Invenio Technology: Selected Practical Software Development Lessons From A Large Digital Library System* [online]. 2013 [cit. 2014-12-05]. Dostupný z WWW: ⟨http://simko.home.cern.ch/simko/talks/invenio-technology-openlab-aug-2011.pdf⟩.

7 GLAUNER, Patrick Oliver. *Enhancing Invenio Digital Library With An External Relevance Ranking Engine* [online]. 2012 [cit. 2014-12-05]. Dostupný z WWW: ⟨http://cds.cern.ch/record/1456329/files/CERN-THESIS-2012-074.pdf⟩.

8 CHHIBBER, Nalin. *Enriching The Metadata On CERN Document Server* [online]. 2014 [cit. 2014-12-05]. Dostupný z WWW: ⟨http://cds.cern.ch/record/1750268/files/Report_Nalinc.pdf⟩.