Masarykova univerzita

Fakulta informatiky

eSciDoc - middleware pre eScience aplikácie

http://www.escidoc-project.de/JSPWiki/en/Startpage

Martin Jantošovič

N-IN PSK [sem 3, roč. 2]

22. novembra 2009

1 Úvod

Digitálne repozitáre začali pred niekoľkými rokmi, s cieľom nahradiť klasické knižnice. Boli zamerané hlavne na publikácie. Dnes sa stávajú bežným nástrojom výskumníkov, akademikov a vedcov. Repozitáre častokrát iba poskytujú priestor na ukladanie, perzistentnú identifikáciu, archiváciu a služby vyhľadávania. Tieto funkcie pred koncovými užívateľmi skrývajú nadstavbové a špecializované aplikácie.

Podobnú štruktúru dodržiava aj projekt eSciDoc. Už od začiatku projektu boli separátne vyvíjané backendové služby (eSciDoc Infrastructure) a frontendové aplikácie (eSciDoc Solutions)[3].

2 Čo je to eSciDoc?

eSciDoc je spoločný projekt spoločností Max Planck Society a FIZ Karlshruhe, ktorý do polovice roku 2009 podporovalo aj Federal Ministry of Education and Research (BMBF). Vývoj začal v roku 2004 a tento rok (t.j. 2009) bol ukočený.

Cieľom bolo vyvinúť a poskytnúť výskumným organizáciam webovú platformu pre elektronický výskum a využiť tak informačné technológie na urýchlenie procesov. Zároveň tvorcovia zamýšľali vytvoriť systém, ktorý by zaručoval pernamentný prístup k výsledkom a materiálom výskumov hlavne spoločnosti Max Planck Society, ale aj globálnému vedeckému svetu.

Mnohé eScience aplikácie sa zameriavajú hlavne na obrovský objem dát, ako ich uložiť, ako s nimi zaobchádzať a analyzovať ich. eSciDoc sa však zameriava viac na riadenie a spravovanie poznatkov počas celého procesu výskumu. Do repozitára je možné zaznamenávať nielen koncové výsledky, ale aj iné kroky, od základnej myšlienky, cez procesy experimentov až po analýzu a vyhodnotenie výsledkov. Digitálna knižnica sa tak stáva infraštruktúrou celého výskumu a dovoľuje tak spoluprácu a komunikáciu vedcov z rôznych kútov sveta.

3 SOA

Architektúra eSciDoc systému je známa ako SOA (Service-Oriented Architecture)[1]. Jednotlivé časti systému sú navrhnuté a implementované tak, aby boli čo najviac škálovatelné a znovuvyužiteľné. SOA rozdeľuje funkcie do samostatných jednotiek alebo služieb, ktoré sú dostupné cez sieť. To dovoľuje ich kombinovanie a znovuvyužívanie pri vývoji nových aplikácií. Tieto služby komunikujú medzi sebou posielaním dát z jednej služby do inej.

Systém má implementované dva prístupy k samostatným službám. Jedným je Simple Object Access Protocol (SOAP) a druhým je Representational State Transfer (REST). Tieto prístupy dovoľujú vývojárom nezávislosť pri výbere programovacieho jazyka a tak urýchľujú ich vývoj. Dokonca je podporovaný aj Web 2.0.

```
SOAP: ItemHandler.retrieve("escidoc:123")
REST: GET /ir/item/escidoc:123
```

Príklad SOAP a REST získania objektu

Všetky služby sa delia do troch vrstiev. Vrstva služieb jadra poskytuje 4 základné operácie: vytvorenie, získanie, modifikácia a zmazanie dátových objektov, ktorými sú napríklad kontext, kontajner alebo položky v repozitári. Tieto služby sú bezstavové a využívané vyššími vrstvami.

Druhou vrstvou sú „prostredné“ služby. Pridávajú ďalšie funkcie k službám jadra. Sú tiež bezstavové a môžu manipulovať s vlastnými dátami. Medzi také služby patrí napríklad validácia, notifikácia alebo handler obrázkov.

Poslednou vrstvou sú aplikačné služby. Sú to služby, ktoré implementujú biznis logiku špecifických riešení. Medzi ne patria služby vyhľadávania alebo služby transformácií medzi dátovými formátmi.

4 Architektúra

Celý systém sa skladá z troch hlavných častí, ktoré sú na sebe nezávislé. „eSciDoc Infrastucture“ (infraštruktúra) poskytuje základné služby a bežne používané funkcie. Časť „eSciDoc Services“ (služby) je zameraná na aplikačné služby, ktoré vytvárajú rozhranie pre tretiu časť s názvom „eSciDoc Solutions“ (riešenia).

4.1 eSciDoc Infrastructure

eSciDoc Infrastructure je v podstate middleware, ktorý zaobaľuje repozitár a implemetuje služby prvej vrstvy SOA, popísanej v predchádzajúcej kapitole.

Tvorcovia sa rozhodli pre kombináciu technológií Java a XML. Miesto budovania úplne novej infraštruktúry, si radšej vybrali cestu integrovania už existujúcich open-source komponentov. Z voľne dostupných balíčkov boli zvolené PostgreSQL ako databázový server, JBoss Application Server a Tomcat Servlet Container ako webový server. Repozitárový systém je vybudovaný na FEDORA (Flexible Extensible Digital Object Repository Architecture), konkrétne Fedora Commons.

4.2 eSciDoc Services

Ako názov predpovedá, táto časť systému je zameraná na služby, ktoré pridávajú ďalšie funkcie k službám jadra. Medzi aktuálne dostupné služby patria:

- **Citation Style Manager** - služby, ktoré pridávajú možnosť pracovať s citačnými formátmi APA, AJP.
- **Data Acquisition Handler** - služby na prácu s internými a externými dátovými zdrojmi.

- **Depositing** - pridáva podporu SWORD protokolu. Bližšie informácie o tomto protokole nájdete na jeho stránkach [2].
- **Digilib** - prostredie pre prácu s obrázkami.
- **Duplicate Detection** - nástroj implementujúci aplikáciu, ktorá prechádza metadáta v repozitári Fedora a hľadá prípadné duplicity.
- **Named Entities** - práca s metadátami.
- **OAI-PMH Provider** - rozhranie pre OAI-PMH.
- **PID Manager** - služba na vytváranie a generovanie perzistentných identifikátorov.
- **Search&Export** - možnosť nastaviť, ktoré zdroje môžu byť prehľadávané a aký bude výstup pri exporte.
- **Technical Metadata Extraction**
- **Transformation**
- **Validation** - je možné validovať položky alebo kontajnery s ohľadom na určené pravidlá.

4.3 eSciDoc Solutions

eSciDoc Solutions predstavuje už hotové riešenia založené na eSciDoc Infrastructure. Je možné si napísať vlastné riešenie alebo použiť nejaké, ktoré samotný projekt eSciDoc ponúka. Momentálne sú dostupné riešenia:

- **Publication Management** - správa publikácií
- **ViRR** - digitálna kolekcia Rímskej Ríše
- **Faces** - kolekcia fotografií tvárí 171 žien a mužov vyjadrujúcich 6 emócií
- **Scholarly Workbench** - riešenie vytvorené pre odbor umenia a humanitných vied

5 Vlastné zhodnotenie projektu

Počas písania tejto eseje som sa pokúsil nainštalovať eSciDoc Infrastructure. Bohužiaľ, nepodarilo sa mi to nainštalovať na mojej distrubúcii linuxu Linux Mint. Z tohto dôvodu nemôžem popísať vlastné skúsenosti.

Výhodou projektu je, že je voľne dostupný pod licenciou Common Development and Distribution Licence (CDDL), takže si ho môže každý nainštalovať a vytvoriť si vlastné riešenie pre svoje potreby.

Autori určite dosiahli toho, čo si určili ako cieľ projektu. Vytvorili robusný a škálovateľný middleware na tvorbu aplikácií určených pre eScience, aby tak urýchlili vedcom prácu. Celý systém je navrhnutý ako „enable technology“, t.j. to čo potrebujete si povolíte alebo nainštalujete.

Referencie

[1] SOA na wikipédii.
http://en.wikipedia.org/wiki/Service-oriented_architecture.

[2] SWORD web stránka.
http://www.swordapp.org/.

[3] Webová stránka eSciDoc projektu.
https://www.escidoc.org/JSPWiki/en/Startpage.

[4] M. Agosti. *eSciDoc Infrastructure: A Fedora-Based e-Research Framework z knihy: Research and Advanced Technology for Digital Libraries*, volume 5714/2009, pages 227–238. 2009.

[5] Malte Dreyer, Ulla Tschida, Natasa Bulatovic, and Matthias Razum. eSciDoc – a Scholarly Information and Communication Platform for the Max Planck Society. *German e-Science*, 2007.

Metadata v Dublin Core

```
<link rel="schema.DC" href="http://purl.org/dc/elements/1.1/" />
<meta name="DC.Title" content="eSciDoc -
 middleware pre eScience aplikácie" />
<meta name="DC.Creator" content="Martin Jantošovič" />
<meta name="DC.Subject" content="eSciDoc -
 middleware pre eScience aplikácie" />
<meta name="DC.Description" content="eSciDoc je open-source
 middleware urceny hlavne pre eScience aplikacie" />
<meta name="DC.Date" content="22.11.2009" />
<meta name="DC.Type" content="Text" />
<meta name="DC.Type" content="Esej" />
<meta name="DC.Format" content="application/pdf" />
<meta name="DC.Format" content="computerFile" />
<meta name="DC.Identifier"
 content="http://www.fi.muni.cz/~xjant/PV070/eSciDoc.pdf" />
<meta name="DC.Source"
 content="https://www.escidoc.org/" />
<meta name="DC.Source"
 content="http://colab.mpdl.mpg.de/mediawiki/ESciDoc_Overview" />
<meta name="DC.Language" content="sk" />
```