Systémy pro tvorbu digitálních knihoven

Vlastimil Krejčíř
Ústav výpočetní techniky, Masarykova univerzita, Brno
krejcir@ics.muni.cz

INFORUM 2006: 12. konference o profesionálních informačních zdrojích Praha, 23. – 25. 5. 2006

Abstrakt. Článek popisuje a hodnotí základní vlastnosti vybraných softwarových systémů pro tvorbu digitálních knihoven. Těmito vybranými systémy jsou Fedora, DSpace, EPrints, CDSware a Greenstone. Popisovány jsou především zajímavé vlastnosti daných systémů a autorovy osobní praktické zkušenosti s jejich provozem. Zároveň jsou dána doporučení pro nasazování těchto systémů do praktického provozu.

Abstract. The paper describes and evaluates basic characteristics of chosen software systems for digital libraries development. The chosen systems are Fedora, DSpace, EPrints, CDSware and Greenstone. The most interesting features of these systems are described and also the author's own experience of testing and running these systems is provided. Finally recommendations for these systems in real applications are given.

V současnosti je k dispozici celá řada softwarových produktů, které poskytují prostředky pro tvorbu digitálních knihoven. Orientace mezi dostupnými řešeními je proto velmi nesnadná a rozhodnutí o tom, který z nabízených systémů je vhodný právě pro potřeby konkrétní organizace, je složité. Budoucí uživatel si klade mnoho otázek. Podporuje vybraný systém technologie a služby, které vyžaduji? Pokud ano, na jaké úrovni? Jaká je technická podpora daného systému? Nezanikne firma či organizace, která systém vyvíjí a poskytuje? Jistě lze vyjmenovat několik dalších klíčových problémů, které je nutné zkoumat. Do cesty se přitom staví mnoho překážek, které kladou vysoké časové nároky pro získání potřebných informací.

V tomto dokumentu budou představeny některé ze současných volně dostupných open-source systémů pro tvorbu digitálních knihoven. Záměrně bude volen spíše pohled uživatelský než technický. Příslušné technické detaily lze snadno získat z bohaté dokumentace, která je u všech představovaných systémů volně k dispozici. Text se tedy zaměří na popis základních vlastností a principů, na kterých jsou prezentované systémy budovány, a poskytne i autorovo vlastní hodnocení, utvořené na základě zkušeností získaných praktickým testováním. Je nutné předem upozornit, že některé popisované vlastnosti se mohou u různých verzí každého systému lišit.

Výběr systémů DSpace, EPrints, Fedora, CDSWare a Greenstone byl proveden spíše intuitivně, na základě autorových osobních zkušeností. Zohledněna byla především skutečnost, že jsou to systémy, které se nasazují v praxi, mají vybudovanou uživatelskou komunitu, případně jsou zajímavé svým řešením.

1 Fedora

Systém Fedora [1] je svým způsobem mezi jednotlivými systémy výjimečný. Vývojáři si dali za cíl vybudovat systém, který bude realizoval základní teoretické poznatky a modely v oblasti digitálních knihoven. Staví mimo jiné na Kahn/Wilenského architektuře [2]. Systém Fedora poskytuje jen repozitářové služby, jakési jádro ošetřující ukládání, správu a archivaci digitálních objektů. Zároveň obsahuje knihovnu funkcí a volání, která umožňuje programátorovi provádět činnosti nad repozitářem. Systém Fedora zatím není zcela kompletní, okamžitě v praxi nasaditelný a použitelný, protože neposkytuje potřebné uživatelské rozhraní – to si musí instituce, která by chtěla systém používat, vytvořit sama.

Systém Fedora je vyvíjený na Cornell University ve spolupráci s University of Virginia v USA. Název je složen z prvních písmen charakteristiky **Flexible Extensible Digital Object and Repository Architecture**. Cílem celého projektu je poskytnout *univerzální* repozitář pro University of Virginia a další, především akademické instituce. Vývoj je v současnosti financován z grantu Mellonovy nadace.

1.1 Základní vlastnosti

Systém Fedora poskytuje repozitář pro ukládání dat – digitálních objektů. Neklade žádná omezení na typ vkládaných dat (fotografie, dokumenty, video, ...). Pro komunikaci s vnějším světem poskytuje repozitář služby. Ty jsou postaveny na technologii, která má „služby" přímo ve svém názvu, na technologii webových služeb (*Web Services* [3]). Ty jsou však určeny především programátorům a správcům, kteří chtějí budovat nad repozitářem uživatelská rozhraní. Pomocí těchto služeb pak lze systém Fedora spravovat (pomocí Management API) a zároveň z něj získávat data (Access API a Access-lite API).

Samotný koncový uživatel bude pravděpodobně ihned po nainstalování celého systému Fedora zklamán, protože nedostává produkt, s nímž je schopen okamžitě pracovat. K dispozici je pouze (velmi propracovaná) aplikace pro administrátory a programátory a jednoduché www rozhraní (vhodné pouze pro testovací účely).

Výčtem zmiňme některé standardy, které systém Fedora využívá. Pro export a import digitálních objektů lze použít standard **METS** [4], případně vlastní formát systému Fedora – FOXML. Popisná metadata jsou ukládána ve formátu **Dublin Core** [5]. Pro sdílení metadat podporuje Fedora protokol **OAI-PMH** [6]. Systém Fedora je napsán v jazyce Java, může být provozován na počítačích s operačními systémy Windows, Linux, MacOS a další (aktuální kompletní seznam podporovaných platforem lze nalézt na stránkách projektu).

1.2 Digitální objekty

Základním prvkem, se kterým systém Fedora pracuje, je digitální objekt (DO). Ten obsahuje především perzistentní jednoznačný identifikátor a systémová metadata. Další obsah objektu se již řídí tím, jakého je daný objekt typu. Digitální objekty v systému Fedora mohou být tří typů.

Datový objekt (Data Object) obsahuje zejména metadata a data – tedy obsah, který chceme uchovávat v repozitáři. Kromě samotných dat může obsahovat jen odkazy na data – ta mohou být uložena v samotném repozitáři i mimo něj (na vzdáleném serveru na Internetu apod.).

Specialitou systému Fedora jsou další dva typy digitálních objektů. Především je to *objekt popisu chování* (Behaviour Definition Object). V něm jsou popsány služby, které se váží na datové objekty. Službou může být například „náhled fotografie". Tuto službu mohou využívat všechny datové objekty, které obsahují fotografie. Objekt popisu chování však pouze službu popisuje, neříká jakým způsobem má být služba provedena.

Konkrétní realizaci služby zajišťuje poslední typ objektu – *objekt implementace chování* (Behavior Mechanism Objects). Je to vpodstatě samotný program, který provedení služby zajistí. V případě naší služby „náhled fotografie“ vrátí tento program zmenšenou fotografii jejíž originál je uložen v datovém objektu.

Služby jsou v *objektu popisu chování* definovány pomocí jazyka WSDL [7]. Tím je dosaženo platformové nezávislosti (implementace služby v *objektu implementace chování* může být teoreticky napsána v jakémkoli programovacím jazyce). Výchozí službou pro jakýkoli objekt je služba, která vrací data objektu přesně tak jak jsou uložena v repozitáři (neprovádí s nimi žádné transformace). V základní instalaci jsou k dispozici i služby pro manipulaci s obrázky a fotografiemi (zmenšení, převod do stupňů šedi apod.), služby pro provádění transformací v jazyce XSLT a další.

Typický scénář vytvoření digitálního objektu může vypadat následovně. Uživatel vytvoří datový objekt s fotografií. Tento objekt naváže na objekt popisu chování s názvem „Služby pro zpracování fotografií“ (jedna z těchto služeb je „náhled fotografie“). Objekt popisu chování pak sváže s konkrétní implementací – programem na zpracování fotografií. Tímto je tento program zároveň navázán na datový objekt. Při každém požadavku uživatele na náhled fotografie je originál fotografie předán tomuto programu na zpracování fotografií. Program vyrobí z fotografie náhled a ten předá uživateli, který jej požadoval.

1.3 Správa verzí (Versioning)

Z dalších zajímavých vlastností systému Fedora lze zmínit podporu ukládání různých verzí téhož obsahu objektu – uživatel například uloží nějaký textový dokument a později chce uložit jeho opravenou verzi. Může ji uložit přímo pod stejným jménem do datového objektu s původní verzí dokumentu. Systém automaticky uchová obě verze dokumentů. Samozřejmě je možné kdykoli získat jakoukoli z uložených verzí.

1.4 Závěr

Systém Fedora není v současné době kompletní hotový knihovní systém, který je možné okamžitě nasadit do běžného provozu. Jedná se spíše o jakési jádro, základ pro budování digitální knihovny. Lze jej připodobnit ke skládačce závodního vozu Formule 1 – uživatel dostává do rukou motor a další díly. Záleží pak jen na něm, jaký bude výsledný monopost. Musí však počítat i s tím, že není vůbec snadné celý vůz poskládat, tak aby byl rychlý a výkonný. Zvolené technologie systému Fedora jsou velmi robustní a dostatečně platformově nezávislé, ale tyto vlastnosti zároveň zvyšují režii práce s repozitářem. Přesto šikovný „konstruktér“ dokáže vytvořit funkční a rychlý systém. Příkladem je Encyclopedia of Chicago [8], který je odkazován přímo ze stránek projektu Fedora. A protože se celý systém Fedora drží standardů a má i slušnou technickou podporu a finanční zázemí, lze jej doporučit jako robustní řešení datového skladu zejména s výhledem do budoucnosti.

2 DSpace

DSpace (Digital Archive Project) [9] je systém poskytující zázemí pro provozování digitálních knihoven (akademické) instituce a s tím spojenou základní funkcionalitu. Původním cílem projektu DSpace bylo vyvinout spolehlivý otevřený univerzální digitální knihovní systém, který bude nasazen na MIT (Massachusetts Institute of Technology) a zároveň dán k volně dispozici jiným potenciálním uživatelům. Postupem doby se vytvořila velmi živá komunita kolem celého systému a na jeho vývoji se podílí programátoři z celého světa. V současnosti se jedná pravděpodobně o jeden z nejživějších, funkčních a nejrychleji se rozvíjejících projektů v oblasti volně dostupných systémů digitálních knihoven.

2.1 Základní vlastnosti

Na rozdíl od systému Fedora je systém DSpace okamžitě po instalaci připraven k provozu v reálném prostředí. Poskytuje jak samotný repozitář pro ukládání dat, tak kompletní www rozhraní pro přístup – pro koncové uživatele i pro správce. Systém DSpace je poskytován volně včetně zdrojových kódů. Komunita uživatelů a vývojářů, která se kolem systému DSpace postupně vytvořila, má poměrně propracovaný systém vývoje. Při dodržení pravidel se může na rozvoji systému podílet kdokoli. Zmíněná pravidla definují zejména okruh vylepšení, která mohou být přijmuta do oficiální verze, a také popisují způsob, kterým lze tato vylepšení dělat. O zařazení konkrétního vylepšení do oficiální verze systému DSpace pak rozhoduje komise složená z hlavních vývojářů. Omezení nejsou kladena ani na modifikace pro lokální potřebu – tyto modifikace se však do oficiálních verzí nezařazují.

Pro popisná metadata používá systém DSpace standard **Dublin Core** (navíc DSpace poskytuje velice pěkné rozhraní pro práci s tímto standardem), pro sdílení metadat standard **OAI-PMH**. Perzistence objektů a jejich jednoznačné identifikátory jsou řešeny pomocí **CNRI Handles** [10]. Identifikátory „handles“ (ukazatele) na objekty jsou vyjádřeny pomocí URN. DSpace podporuje i standard **OpenURL** [11].

Systém DSpace může být provozován na operačních systémech typu UN*X. Verze pro operační systém Windows není v současnosti k dispozici.

2.2 Digitální objekty a struktura

Digitální objekt systému DSpace je struktura s jednoznačným identifikátorem, která zapouzdřuje metadata i samotná data. Na rozdíl od systému Fedora však systém DSpace poskytuje kromě digitálních objektů i struktury pro jejich logické rozčlenění. Již při vytváření nového digitálního objektu uživatel určuje, do které *kolekce* (Collection) bude daný objekt patřit. Kolekci mohou například tvořit příspěvky z konference Inforum 2006. Digitální objekt může být zařazen i v několika kolekcích (nejméně však v jedné). Kolekce jsou zařazeny pod další logické celky, které se nazývají *komunity* (Communities). Každá komunita může být zařazena pod jinou komunitu. Komunitou může být například Masarykova univerzita, jinou komunitou zařazenou pod komunitu Masarykova univerzita může být například Ústav výpočetní techniky a podobně.

Z výše uvedeného textu lze usuzovat, že komunity jsou vhodné právě a pouze pro reprezentaci struktury instituce, která systém DSpace využívá. Je pravdou, že toto je základní cíl, který mají komunity plnit. V praxi je však lze s úspěchem použít i k vytvoření zcela odlišných struktur, které jsou komunitami pouze technicky. Při ukládání článků z časopisů lze pomocí zanořování komunit dosáhnout struktury *Časopis* → *Ročník* → *Číslo*. Komunitami jsou v tomto případě *Časopis* a *Ročník*. *Číslo* je zde kolekcí, která obsahuje digitální objekty, kterými jsou jednotlivé články (které jsou součástí daného čísla).

Zavedené struktury slouží k snadné navigaci pro uživatele přistupujícího k systému DSpace přes webové rozhraní.

2.3 Webové rozhraní a jeho funkce

Webové rozhraní je velmi dobře propracováno. Nabízí širokou paletu funkcí, které může uživatel při práci s repozitářem potřebovat – procházení obsahem repozitáře (logicky přes komunity a kolekce k samotným digitálním objektům), vyhledávání v repozitáři, propracované vkládání nových dat, vytváření kolekcí, systém správy uživatelů a metadat, a další funkce, které umožňují pohodlnou práci s celým systémem DSpace. Rozhraní působí velmi příjemně a pohodlně se s ním pracuje. Vlastní vzhled (barevné schéma, velikosti fontů, ...) je vytvořen pomocí technologie CSS a správce jej může snadno změnit.

2.4 Další funkce

Z dalších funkcí, které systém DSpace poskytuje zmiňme podporu uživatelský účtů včetně základního systému autentizace (který lze opět změnit například na autentizaci pomocí protokolu LDAP aj.). Dále je k dispozici možnost vytváření skupin uživatelů, kterým je možné přidělovat přístupová práva ke kolekcím i digitálním objektům. Uživatele lze například ustanovit správcem některé z kolekcí apod.

Vyhledávání v systému DSpace je umožněno jak v metadatech, tak samotných datech (DSpace umí kromě obyčejných textových souborů indexovat i soubory typu PDF a Microsoft DOC).

Systém DSpace podporuje tzv. worklflow proces vkládání nových objektů. V praxi to znamená, že uživatelem vložený objekt nemusí být okamžitě zařazen do repozitáře, ale může být předán do schvalovacího řízení. To se skládá z několika kroků v nichž pověření uživatelé mají možnost nový objekt zkontrolovat (případně opravit metadata apod.) a postoupit ho do další fáze schvalování (popř. je-li to fáze poslední schválit uložení do repozitáře). Systém DSpace napevno definuje několik fází (např. schvalování metadat objektu, samotné potvrzení vložení do repozitáře apod.). Jednotlivé kroky schvalování se definují pro kolekce. Každá kolekce tedy může mít libovolnou kombinaci schvalovacích fází. Objekt, který je do této kolekce vkládán, pak musí projít všemi definovanými fázemi schvalování a až poté je vložen do repozitáře.

2.5 Závěr

Systém DSpace se ukazuje být nejživějším z prezentovaných systémů. V současné verzi již jeho vývoj značně pokročil a nadále probíhá. Je to kompletní systém, což je jeho výhoda i nevýhoda. Výhodné je, že ho lze prakticky okamžitě nasadit do reálného provozu. A to může být zároveň nevýhodou – systém DSpace je pevně naprogramován a některé jeho funkce jsou dány a nelze je zcela snadno přizpůsobit vlastním potřebám (například prezentace digitálního objektu, jeho metadat a obsahu je jednotná, ať je již obsahem cokoli). Toto může být značným omezením pro instituci, která má konkrétní požadavky jež se neshodují s tím, co DSpace poskytuje. Uživatel, který se rozhodne systém DSpace používat, se buď musí smířit s daným stavem, nebo začít systém DSpace upravovat dle svého. Úpravy však znamenají obvykle větší zásahy do samotného kódu a při přechodu na novou verzi DSpace je nutné celé přeprogramování provést znovu (s větším či menším úsilím v závislosti na rozsahu úprav). Tyto problémy by ovšem měly být vyřešeny v budoucnu, kdy se plánuje podpora tzv. *Addons*, což jsou jakési pluginy, kterými bude možné systém DSpace vylepšovat.

Příkladem možné úpravy systému DSpace je „plugin“ Tapir [12] (nejedná se o plugin v pravém slova smyslu, autoři tak Tapir jen označují – ve skutečnosti je to přepis zdrojových kódů systému DSpace). Plugin Tapir vyvíjí programátoři na University of Edinburgh ve Velké Británii. Je specializací systému DSpace na ukládání diplomových a disertačních prací. Obvykle je tento plugin pro novou verzi DSpace vydáván s dvou až tříměsíčním zpožděním. Existuje mnoho dalších konkrétních úprav DSpace, které jsou dělány zejména lokálně – instituce, které úpravy udělaly pro vlastní potřeby, jsou většinou ochotny je případným zájemcům poskytnout zdarma. Tímto postupem lze řešit i některé speciálnější požadavky na systém DSpace.

Systém DSpace je nasazován v praxi, zajímavé instalace jsou například [13], [14], [15] nebo [16] (zájemce také může pro vyhledání aktuálních instalací použít vyhledávač Google[1] [17]). Systém DSpace lze doporučit každé instituci, která hledá rychlé, snadné a levné řešení pro strukturované ukládání dat, které je zároveň i dostatečně živé a robustní. A v případě, že daná instituce nemá extrémně zvláštní požadavky, je systém DSpace jedním z nejvhodnějších řešení.

3 EPrints

Systém EPrints [18] je určen zejména pro správu vědeckých a technických informací a pro sdílení výsledků a novinek vědeckého výzkumu. Příkladem takových informací jsou diplomové práce, vědecké články, technické zprávy, příspěvky z konferencí apod. EPrints je kompletní systém připravený k okamžitému nasazení. Je vyvíjen na University of Southampton ve Velké Británii. Šířen je pod hlavičkou organizace GNU [19].

3.1 Základní vlastnosti

Výchozímu zaměření systému EPrints odpovídají i veškeré procesy, uživatelské rozhraní, metadata digitálních objektů a celý systém správy. Vše je samozřejmě konfigurovatelné a nic nebrání s jistým úsilím vytvořit obecný repozitář – výchozí orientace systému je však lehce omezující.

Z obvyklých standardů podporuje systém EPrints především protokol **OAI-PMH**. Pro metadata používá vlastní vnitřní formát. Celý systém je kompletně naprogramován ve skriptovacím jazyce Perl.

Systém EPrints nepoužívá termín digitální objekt, ale užívá termín položka (*Item*). Položka je v principu digitálním objektem – zapouzdřuje metadata i data. Dále v textu bude pojem položka v tomto smyslu užíván.

3.2 Archivy

Specialitou systému EPrints je podpora archivů – což jsou v podstatě zcela samostatné digitální knihovny s vlastní konfigurací, které běží nad jednou instalací systému EPrints (a samozřejmě tak společně využívají velkou část kódu). Je tedy možné, aby jeden server obsluhoval samostatné a navenek zcela odlišné archivy dvou i více institucí. Archivy jsou obvykle řešeny pomocí virtuální serverů webového serveru Apache [20] (který systém EPrints užívá).

3.3 Rozhraní a jeho funkce

Systém EPrints nabízí kompletní uživatelské rozhraní, podobně jako systém DSpace. Vzhled a částečně i samotné funkce rozhraní jsou dobře konfigurovatelné. K dispozici jsou standardní

[1] Doporučuji zadat vyhledávači řetězec „DSpace at“

služby – vyhledávání, odkazy na jednotlivé části systému, procházení obsahem repozitáře, výpis nejnovějších přidaných položek, dále odkazy na vstup do uživatelské oblasti, registrace nového uživatele a další. Správci mohou využívat další funkce, které umožňují celý systém konfigurovat a obsluhovat.

Z výše zmíněného výčtu funkcí je patrné, že systém EPrints má kompletní podporu uživatelských účtů. Přidělování práv není tak propracované jako v systému DSpace. Uživatelé jsou pouze tří typů (což většinou zcela postačuje) – *administrátor*, *editor* a *obyčejný uživatel*. Funkce *administrátora* je zřejmá. *Obyčejný uživatel* může přidávat vkládat nové položky. Systém EPrints podporuje workflow proces přidávání nových položek, podobně jako je tomu v systému DSpace. Proto jsou zavedeni uživatelé – *editoři*. Ti mají právo schvalovat přidání nové položky do repozitáře (a tím schválit její zveřejnění). Obyčejný uživatel smí svoji vlastní položku, která již byla finálně schválena pro vložení do repozitáře editovat. Takto editovaná položka však musí opět projít celým schvalovacím procesem.

Velmi propracované (snad nejlépe ze všech prezentovaných systémů) je rozhraní pro vložení nové položky. To samozřejmě souvisí se zaměřením systému EPrints na vědecké dokumenty. Veškerá metadata, která musí uživatel vyplnit, jsou tomu přizpůsobena – uživatel zadává kdy byl dokument publikován, v jakém časopise, jakým nakladatelem, ISSN, ...). Vkládání probíhá v několika krocích a množství informací, které je potřeba vyplnit, je značné a celý proces zabere netriviální množství času. Ovšem pro publikování článků a vědeckých prací je celý systém naprosto vynikající (nepředpokládá se, že uživatel bude vkládat více než jednotky položek v krátkém čase).

Vyhledávání je umožněno v metadatech i datech. Nainstaluje-li správce systému potřebné aplikace, je možné indexovat (a následně prohledávat) i obsah dokumentů typu PDF, Microsoft DOC, HTML a další.

3.4 Struktury

Systém EPrints má i vlastní řazení dokumentů do stromové struktury, která pak umožňuje uživatelům snadnější orientaci a procházení obsahem repozitáře. Protože je systém zaměřen na dokumenty, je výchozí struktura předem dána a uživatel do ní dokument pouze vkládá. Členění struktury je stejné, jaké používá Library of Congress [21] v USA (tvořit však lze i členění vlastní). Kořen tohoto členění je označován jako oblast (*Area*), zde *Library of Congress Subject Areas*. Uživatel tak může postupně určit, že vkládaný dokument je například práce z oblasti medicíny, dále že se jedná o oblast výzkumu nervové soustavy atp. Uživatelé hledající dokumenty z oblasti výzkumu nervové soustavy se snadno díky tomuto členění dostanou ke všem pracím, které jsou na toto téma v repozitáři uloženy.

3.5 Statické generování stránek

V systému EPrints se předpokládá, že vkládání dokumentů nebude příliš časté (myšleno ve stovkách denně). Proto byla z důvodů rychlosti přístupu k datům zvolena metoda generování statických stránek. Očekává se, že většina operací na serveru bude spojena se čtením dokumentů, hledáním, listováním atp. a zápis a editace budou tvořit jen malou část procesorového času. Neustálé dynamické generování by pak znamenalo zbytečnou zátěž. Proto je část rozhraní pro přístup k systému tvořena statickými HTML stránkami, které se jednou za čas (obvykle v době nejmenší zátěže, například v noci) generují. Pokud je nějaká položka vložena do repozitáře, pak se ve veřejné části webu objeví až po té, co je spuštěno přegenerování HTML stránek. Načítání statických předgenerovaných stránek je mnohem rychlejší než jejich dynamické generování za běhu a nezatěžuje celý systém. Tento přístup je volen zejména v případech, kdy se předpokládá skladování obrovského množství dat. Představme si situaci, kdy uživatel chce vypsat názvy všech dokumentů v repozitáři – v dynamickém systému se musí všechny názvy přečíst z databáze (čtení z databáze je časově náročná operace) a zobrazit uživateli. Pokud tuto operaci bude požadovat postupně několik desítek uživatelů, časová náročnost (i přes optimalizace opakovaného čtení z databáze) vzroste. Naopak pokud již máme tento seznam vygenerován, pouze jej uživateli zobrazíme (stejně jako dalším desítkám uživatelů, kteří jej budou požadovat) a celá zátěž systému výrazně poklesne.

3.6 Závěr

Systém EPrints byl hlavně vytvořen pro jeden účel – uchovávání a sdílení dokumentů v akademické komunitě. Pro tento účel je naprosto vynikající a splňuje asi většinu požadavků, které na něj lze klást. Což ho samozřejmě znevýhodňuje v soutěži s jinými, více univerzálními systémy. Systém EPrints lze samozřejmě s jistým úsilím přizpůsobit a přepracovat, zůstává však otázkou, jestli se taková práce vyplatí (a jestli není spíše vhodné sáhnout po jiném systému). Prvotní otázku při rozhodování o nasazení systému EPrints, kterou si musí budoucí uživatel položit, je, jaký typ dat chce v systému uchovávat. Například pro uchovávání fotografií není systém EPrints zcela vhodný (uživatelé vkládají i desítky fotografií denně, mnohem dynamičtěji je v systému editují, mohou spíše požadovat, aby se vložené fotografie okamžitě objevily ve veřejné části webového rozhraní apod.).

V praxi je systém EPrints využíván poměrně hojně. Na domovské www stránce systému je statistika, která uvádí aktuální množství položek ve veřejně běžících systémech EPrints. Toto číslo se v současnosti pohybuje v desítkách tisíc. Většina těchto instalací je využita právě k účelu, pro který byl systém EPrints vytvořen. Příkladem může být například E-LIS [22] (digitální knihovna, která uchovává a zpřístupňuje dokumenty z oblasti knihovní a informační vědy). Systém EPrints lze proto doporučit zejména pro tuto oblast nasazení.

4 CDSware

Systém CDSware [23] je aplikací vyvíjenou ve švýcarském CERNu. Primárně je určen právě pro potřeby této uznávané vědecké instituce, ale zároveň je dán zdarma (včetně zdrojových kódů) k dispozici dalším zájemcům. Je to rozsáhlý kompletní systém poskytující datové úložiště i uživatelské (a administrátorské) webové rozhraní. Je poskytován pod hlavičkou GNU a licencí GNU Public Licence. Systém CDSware je také částečně lokalizován do českého jazyka. Jako jediný z prezentovaných systémů nabízí i možnost placené podpory a údržby systému.

Koncepcí přístupu k uživatelskému rozhraní se řadí na rozhraní mezi systémy Fedora a DSpace. Filozofií prezentování informací se spíše blíží systému EPrints.

Interoperabilitu řeší systém CDSware pomocí protokolu **OAI-PMH** (k dispozici je harvester i provider OAI), metadata ukládá v knihovnickém formátu **MARC 21** (resp. na tento formát metadata mapuje a toto mapování je možné měnit).

4.1 Uživatelské rozhraní

Blízko k systému Fedora má systém CDSware především díky řešení vlastního uživatelského rozhraní. To je v základní instalaci systému CDSware značně nekompletní a není okamžitě provozuschopné. Uživatel má ihned po instalaci dvě možnosti – využít tzv. demo instalace, která vytvoří komponenty, které udělají systém CDSware okamžitě provozuschopným, nebo začít vytvářet potřebné komponenty systému sám.

Výše zmíněnými komponentami systému jsou například systém pro vkládání digitálních objektů (je třeba nadefinovat metadata a formuláře pro každý konkrétní typ dat, která bude chtít uživatel vkládat), přiřadit akce, které lze na konkrétních typech dat provádět (například editace, mazání, . . .) a další. Pro tvorbu těchto komponent je k dispozici velmi propracovaný systém, který je také velmi rozsáhlý a komplikovaný a orientace v něm není příliš intuitivní. Je nutné sledovat podrobně dokumentaci a vytvoření právě takové komponenty, kterou uživatel chce, není jednoduchou záležitostí. Nelze se přitom vyhnout ani programování (v jazyce Python). Komplikovanost celého systému tvorby komponent plně vyvažuje skutečnost, že pokud uživatel celý proces tvorby komponenty zvládne, je schopen vyrobit v podstatě „cokoli“.

Pokud je použita demo instalace, pak jsou základní komponenty již připraveny – je však nutné se smířit s tím, jak vypadají (formuláře pro vkládání typů objektů, vzhled atd.). Demo instalace přirozeně nepokrývá zdaleka všechny typy dat.

Webové rozhraní je poměrně jednoduché a přehledné, spočívá především v možnosti procházet obsah serveru a propracovaného vyhledávání dokumentů. Jeho vzhled je samozřejmě možné měnit dle potřeby.

Systém CDSware poskytuje rovněž kompletní systém přihlašování uživatelů, přidělování práv a rolí. Lze jej hodnotit jako jeden z nejlepších mezi prezentovanými systémy.

Podporováno je bohaté strukturování objektů do složek a kolekcí, do kterých se dají vkládat libovolné typy dat (u každé kolekce/složky je možné nastavit, které typy dokumentů do ní lze vkládat). U každé kolekce lze nastavit způsob výpisu objektů v ní obsažených, způsob řazení, dodatečné informace (odkazy na jiné logicky související kolekce) atp.

Podobně jako systém EPrints používá i systém CDSware předgenerovávání statických www stránek.

4.2 Rozšíření a další vlastnosti

Filozofií systému CDSware je, že cokoli, co umí udělat někdo jiný, nemusí dělat přímo sám systém. Výsledkem je, že nenáročný uživatel si nainstaluje jen základní systém a spokojí se s jeho funkcemi. Pokud však žádá navíc další služby (indexování obsahu dokumentů, slovníky autorit aj.), musí si doinstalovat aplikace, se kterými je pak systém CDSware schopný pracovat. Celkem je možné nainstalovat asi 15 aplikací (většina je GNU), které výrazně zvyšují funkcionalitu celého systému. Je přitom důležité dodržet kompatibilitu jednotlivých verzí aplikací s nainstalovanou verzí systému CDSware. Navíc mnohé z aplikací jsou sice uvedeny jako nepovinné pro instalaci základního systému, ale zároveň jsou uvedeny jako doporučené – v praxi to většinou znamená raději je nainstalovat a vyhnout se tak případným problémům s následným provozem celého systému.

Z dalších zajímavých vlastností zmiňme možnost měnit vlastnosti vyhledávacího algoritmu, nastavovat pravidla pro řazení výsledků vyhledávání, nastavit která metadata mají při vyhledávání větší váhu atd.

4.3 Závěr

Systém CDSware je obrovský, komplexní a propracovaný produkt, který má nesmírně širokou funkcionalitu (a stále je vyvíjen). Domnívám se však, že právě to veliké množství funkcí, různých nastavení a kombinací s dalšími aplikace z něj dělají i systém velmi těžkopádný. Jen samotná instalace je malým dobrodružstvím – zkompilovat všechny potřebné aplikace ve správných verzích (navíc každá z těchto aplikací je závislá na dalších aplikacích a knihovnách operačního systému, které opět musí být ve správných verzích). Správně a bez problémů vše zkombinovat je i dílem štěstí. Podpora uživatelům je velmi dobrá – samotní vývojáři a správci CDSware jsou velmi vstřícní a snaží se pomáhat při řešení všem problémů.

Na systému CDSware je patrné, že je vyvíjen především pro velikou instituci, jakou je CERN. Musí pokrýt veškeré požadavky, který na něj pracovníci z celého CERNu kladou, což z něj dělá obrovský a komplikovaný systém. Zároveň jej však lze (s netriviálním úsilím) plně nakonfigurovat a přizpůsobit vlastním potřebám. Lze jej proto doporučit pro velké a velmi různorodé instituce – pro takové instituce pak bude výhodné zaplatit si instalaci, přizpůsobení a údržbu, než řešit nasazení systému CDSware vlastními silami.

5 Greenstone

Systém Greenstone (Greenstone Digital Library Software) [24] je vyvíjen na University of Waikato na Novém Zélandě. Má podporu organizace UNESCO. Je to volně dostupný open-source systém šířený pod licencí GNU General Public License. Systém je částečně lokalizován do češtiny. Autor tohoto příspěvku nemá osobní zkušenosti s provozem systému Greenstone a převážnou část informací čerpal z obsáhlé diplomové práce Jakuba Řehana [25].

Systém Greenstone se svojí filozofií nejvíce blíží systému CDSware – zejména v oblasti přístupu ke konfiguraci systému a tvorbě uživatelského rozhraní.

5.1 Základní vlastnosti

Ze standardů využívá systém Greenstone především protokol **Z39.50**, který zajišťuje interoperabilitu. Po instalaci speciálního pluginu je k dispozici i protokol **OAI-PMH**. Pro uchovávání popisných metadat užívá systém Greenstone standard **Dublin Core**.

Základní jednotkou pro uchovávání dat v systému Greenstone je *dokument*. Dokument je nosičem informací, kterými mohou být texty, obrázky, audio nebo video. Dokumenty jsou sdružovány do *sbírek* a sbírky jsou sdružovány do *knihoven*. *Sbírka* je základním kamenem, ze kterého systém Greenstone vychází. Je to samostatná jednotka, která uchovává logicky související dokumenty.

5.2 Tvorba sbírek a ukládání dokumentů

Se *sbírkami* úzce souvisí systém ukládání dokumentů do repozitáře. Řešení systému Greenstone je v přístupu k tomuto problému ojedinělé. Uživatel nejdříve vytvoří sbírku s určitými vlastnostmi, které přesně vymezí, jaké dokumenty se do sbírky budou moci vkládat, jaký bude celý proces jejich vložení a jak bude vypadat uživatelské rozhraní pro danou sbírku (jak budou prezentovány dokumenty ve sbírce uložené). Vzhled www rozhraní, které bude poskytovat přístup k obsahu sbírky, lze nadefinovat pomocí systému šablon a maker.

Proces vkládání dokumentů je složen ze dvou hlavních kroků – importu dat a metadat do interního formátu (v jazyce XML) systému Greenstone a následného zařazení již takto zpracovaných dat do repozitáře. Samotný import dat do interního formátu zahrnuje několik kroků. Především se jedná o samotné zpracování dat, která mají být uložena v repozitáři.

Pro zpracování každého datového typu (PDF, DOC, JPG, AVI, ...), musí existovat plugin. Pro textové dokumenty je tímto zpracováním například převod do znakové sady UTF-8, indexování obsahu dokumentu, získání části metadat z dokumentu apod. Při importu je vygenerován i jednoznačný identifikátor (generuje se pomocí hashovací funkce[2], kterou vytvořil jeden z autorů systému Greenstone).

5.3 Uživatelská rozhraní

Pro správu sbírek a ukládání dokumentů je spolu se systémem Greenstone dodávána aplikace Librarian (napsaná v programovacím jazyku Java). Díky dobře navržené architektuře systému je možné vytvářet vlastní aplikace, kterými lze celý systém obsluhovat – uživatelské rozhraní poskytuje část systému zvaná *recepční*, která komunikuje s jinou nezávislou částí zvanou *server sbírky* (ta se stará o poskytování obsahu). Obě tyto části spolu komunikují prostřednictvím interního protokolu systému Greenstone, který však může být nahrazen protokolem systému CORBA [27].

Uživatelské rozhraní pro přístup k obsahu repozitáře je webové a umožňuje základní procházení sbírek a vyhledávání v nich. Procházení obsahem sbírky je umožněno díky klasifikátorům dokumentů – což jsou v podstatě jakési složky, do nichž lze logicky související dokumenty zatřiďovat. Jednotlivé klasifikátory lze vnořovat do sebe a tím tvořit tématickou strukturu, ve které jsou dokumenty uloženy, a jíž může uživatel snadno procházet.

5.4 Závěr

Systém Greenstone se ukázal být v praxi použitelný. Dosáhnutí určitých úprav však nemusí být zcela snadné (viz [25]). Slabinou může být také současný stav v oblasti uživatelských účtů a práv – v podstatě se omezuje pouze na definování administrátorů a správců sbírek. Jemnější přidělování práv uživatelům zatím není zcela podporováno. I proto je systém Greenstone vhodný především pro instituce, které nepředpokládají, že uživatelé budou sami vytvářet obsah repozitáře.

6 Lokalizace prezentovaných systémů

Všechny z prezentovaných systémů s výjimkou systémů CDSware a Greenstone jsou k dispozici pouze v anglické verzi. Obvykle jsou však s nimi dodávány nástroje, které podporují tvorbu lokalizací pro různé jazyky.

Všechny systémy si v současnosti dokáží přinejmenším částečně poradit se všemi českými znaky s diakritikou. Bez problémů funguje ukládání metadat a vyhledávání nad nimi. Problémy dělá všem systémům indexace (pokud ji podporují) ukládaných dokumentů – většinou si systém dokáže poradit jen s dokumenty, které jsou uloženy v kódování, které sám používá jako výchozí (problémy jsou například s Win-1250, kódovou stránkou systému Windows, která je v současnosti u nás pravděpodobně nejpoužívanější).

Druhým závažnějším problémem je „ceske vyhledavani“. U plně lokalizovaných vyhledávání jsou čeští uživatelé uvyklí na to, že po zadání dotazu bez diakritiky jsou nalezena i

[2]Zjednodušeně lze říci, že hashovací funkce je funkce, která spočítá z obsahu dokumentu (nebo množiny dat) jeho jednoznačný identifikátor (označovaný jako otisk). Nejznámějšími příklady hashovacích funkcí jsou například MD5 nebo SHA-1.

odpovídající slova s háčky a čárkami. Například na dotaz „skola“ odpoví systém nalezením všech záznamů se slovem „skola“ i „škola“. A tato funkcionalita není dosud podporována.

7 Závěr

Z popisu uvedených systémů lze vyvodit, že žádný z nich není ideálním univerzálním softwarovým systémem, který by řešil zcela všechny problémy a požadavky. Každý z nich má kladné i záporné vlastnosti. Instituce, která uvažuje o nasazení některého z prezentovaných systémů, by se měla rozhodovat na základě cílů, které by měl daný systém splnit. Důležité je zaměření na otázky jaká data bude systém uchovávat, v jakém množství, v jakých strukturách, kdo budou koncoví uživatelé a jak bude vypadat jejich práce se systémem, kdo a jakým způsobem bude data do systému vkládat, jak sofistikovaný má být systém přístupových práv, a další. Tento příspěvek je stručnou odpovědí na některé z výše uvedených otázek a podává alespoň obecnější přehled oblasti volně dostupných systémů pro tvorbu digitálních knihoven. Některé podané informace o daných systémech jsou částečně závislé na konkrétní verzi – při případném rozhodování o nasazení vybraného systému je velmi vhodné studovat dostupnou dokumentaci dané dostupné verze vybraného systému. Další informace lze nalézt také v [28].

Reference

[1] Fedora, Flexible Extensible Digital Object and Repository Architecture. `http://www.fedora.info/`

[2] Robert Kahn, Robert Wilensky. *A Framework for Distributed Digital Objects Services*. May 13, 1995. cnri.dlib/tn95-01. `http://www.cnri.reston.va.us/k-w.html`. Dostupný na www dne 13. 4. 2006

[3] Web Services. `http://www.w3.org/2002/ws/`

[4] METS, Metadata Encoding & Transmission Standard. `http://www.loc.gov/standards/mets/`.

[5] Dublin Core. `http://dublincore.org/`

[6] OAI-PMH, Open Archives Initiative – Protocol for Metadata Harvesting. `http://www.openarchives.org/OAI/openarchivesprotocol.html`

[7] Web Services Description Language. `http://www.w3.org/TR/wsdl`

[8] Encyclopedia of Chicago. `http://www.encyclopedia.chicagohistory.org/`

[9] DSpace, Digital Archive Project. `http://dspace.org`

[10] CNRI, Corporation for National Research Initiatives. Handle System. `http://www.handle.net/`

[11] OpenURL. `http://library.caltech.edu/openurl/`

[12] The Tapir for DSpace. `http://www.thesesalive.ac.uk/dsp_home.shtml`

[13] DSpace@Cambridge. http://www.dspace.cam.ac.uk/

[14] DSpace at MIT. https://dspace.mit.edu/

[15] Sissa Digital Library. https://digitallibrary.sissa.it/

[16] RepositóriUM. https://repositorium.sdum.uminho.pt/

[17] Google. http://www.google.com/

[18] EPrints. http://software.eprints.org/

[19] The GNU. http://www.gnu.org/

[20] Apache HTTP server. http://httpd.apache.org/

[21] Library of Congress. http://www.loc.gov/

[22] ELIS – The open archive for Library and Information Science. http://eprints.rclis.org/

[23] CDSware – CERN Document Server Software. http://cdsware.cern.ch/

[24] Greenstone Digital Library Software. http://www.greenstone.org/

[25] Jakub Řehan. *Systémy na podporu digitálních knihoven (Greenstone)*. Diplomová práce. Fakulta informatiky, Masarykova univerzita, 2004.

[26] Z39.50. http://www.loc.gov/z3950/agency/

[27] Common Object Request Broker Architecture (CORBA). http://www.corba.org/

[28] Vlastimil Krejčíř. *Univerzální digitální repozitář*. Diplomová práce. Fakulta informatiky, Masarykova univerzita, 2005. http://eprints.rclis.org/archive/00005076/. Dostupný na www dne 18. 4. 2006