OTEVŘENÁ BRÁNA K HISTORICKÝM PRAMENŮM

Petr Kocourek, Státní oblastní archiv v Plzni, konference Archivy, knihovny a muzea v digitálním světě 2012

1. Představení webového portálu Porta fontium
2. Úskalí digitalizace a tvorby aplikace založené na open source software v paměťové instituci s regionální působností

Obsah:

- Použité výrazy, (ne)povinně využívaný specializovaný software v českých archivech
- Historie vývoje software ve Státním oblastním archivu (SOA) v Plzni – připomenutí příspěvku z konference Archivy, knihovny a muzea v digitálním světě (AKM) z roku 2010.
- Přeshraniční projekty a nový webový portál Porta fontium
- Zkušenosti s digitalizací
- Úskalí vývoje software

Završením prvního přeshraničního projektu zaměřeného na digitalizaci, kterého se jako hlavní partner přímo účastnil Státní oblastní archiv v Plzni, se otevírá příležitost k představení jeho výsledků. Zároveň právě uplynuly dva roky od chvíle, kdy byla na konferenci Archivy, knihovny a muzea v digitálním světě představena historie digitalizace a vývoje software ve Státním oblastním archivu v Plzni, která mimo jiné vedla k rozhodnutí založit další vývoj na open source software. Nyní se můžeme otevřeně podělit o naše dosavadní zkušenosti.

Krutě zjednodušující seznam použitých výrazů a skutečností

Do příspěvků z konference vystavených na Internetu se občas začtou i lidé z jiných oborů, například studenti informatiky, kteří se chystají na první schůzku s námi v rámci spolupráce s univerzitou. Využívám tedy této příležitosti k velmi stručnému vysvětlení několika pojmů, abychom se případně rychleji domluvili. Ostatní zde mohou alespoň nahlédnout, jak informatik po několika letech práce v archivu pochytil pojmy a skutečnosti z prostředí českého archivnictví:

- Archiválie – dokument v péči archivu (obvykle text nepublikovaný tiskem)
- Fond – logický soubor archiválií (zpravidla vytvořený nebo shromážděný jedním původcem, jinak někdy mluvíme o sbírce)
- Pomůcka – odborný popis fondu a jeho struktury vydaný archivem (formou textu, někdy i databáze) obvykle mluvíme o inventáři. Po vydání pomůcky obvykle prohlásíme fond za zpřístupněný.
- Open source – software s veřejně přístupným zdrojovým kódem. Obvykle jej lze využívat a upravovat bez nutnosti platit někomu za licenci.
- PEvA – jednotný program pro evidenci souhrnných informací o fondech a pomůckách, povinně používaný všemi archivy v ČR.
  Pro podrobnější popis a zpřístupňování archiválií u nás povinný program neexistuje, každý archiv si to řeší po svém.

Vývoj software pro archivy ve Státním oblastním archivu v Plzni

Pro lepší pochopení pozadí našeho rozhodnutí vyvíjet si software pro zveřejňování archiválií sami na bázi open source software stručně připomenu některé informace, které zazněly na konferenci Archivy, knihovny a muzea v digitálním světě v roce 2010.

2002-2003

Po zrušení okresních úřadů přechází státní okresní archivy pod státní oblastní archivy, které působí na území bývalých krajů. Některé okresní archivy nemají ani počítačovou síť, jiné mají vnitřní informační systém (Tachov) nebo i zveřejňují archiválie na Internetu (Cheb). Začíná snaha o sjednocování informačních a komunikačních technologií.

2004-2008

Vývoj nového informačního systému ve spolupráci s komerční firmou, která plánuje vznikající systém prodávat i do dalších archivů. Inspirace úspěšným modelem vývoje spolupráce mezi společností Bach a Zemským archivem v Opavě.

2009

Ukončení obchodního partnerství. Kvůli nevýhodné smlouvě není software naším majetkem a informační systém dokonce ani nesmíme dále používat, zůstává nám jen obsah databází a hardware. Rozhodnutí nekupovat jediný zbývající zavedený informační systém pro archivy v ČR (od firmy Bach), ale raději vyvíjet vlastní software (podobně jako oblastní archiv v Třeboni), navíc založený na open source projektech. Část informačního systému nezbytnou pro provoz archivu nahrazujeme vlastními silami (intranet, ekonomické evidence, webové stránky) nebo ve spolupráci se studenty informatiky Západočeské univerzity (jednoduchý podací deník).

2010

Počátek vývoje náhrady software pro zpřístupňování archiválií. Rozhodnutí zveřejňovat digitalizované archiválie i na Internetu.

Přeshraniční projekty

Od roku 2010 již zveřejňujeme listiny v rámci mezinárodního projektu Monasterium. V současnosti zde máme zpřístupněny listiny z 33 fondů západočeských archivů. Snímky jsou uloženy na našich serverech, popisná metadata odesíláme provozovatelům portálu ve formátu excelovské tabulky.

Velká priorita je on-line zpřístupnění badatelsky nejžádanějšího materiálu – sbírky matrik, kterou digitalizujeme již od roku 2005.

Moravský zemský archiv (MZA) v rámci projektu přeshraniční spolupráce České republiky a Rakouska zprovoznil novou webovou aplikaci pro zpřístupnění matrik – Acta Publica. Protože SOA v Plzni, SOA v Praze a MZA používaly původně pro evidenci matrik stejný „modul Matriky“, předpokládáme kompatibilní databáze (dokumentace neexistuje, zdrojový kód původní aplikace nemáme k dispozici, pouze obsah databáze). Díky vstřícnosti MZA začínáme na Acta Publica již od roku 2010 zveřejňovat naše matriky, později se připojuje také SOA v Praze. Zadávání metadat bylo dokončeno v červnu roku 2012, postupné skenování snímků bude

završeno na jaře roku 2013. Na rozdíl od Monasteria jsou snímky předzpracovány a uloženy na serverech v Brně, protože jejich stahování z našich serverů, zpracování a posílání do klientského prohlížeče v reálném čase na základě aktuálního požadavku na prohlížení se záhy ukázalo příliš zdlouhavé, uživatel musel na zobrazení obrázku čekat desítky sekund. Hlavní problém aplikace Acta Publica pro plzeňský oblastní archiv je, že kvůli ne zcela kompatibilní datové struktuře zde asi nikdy nebude zprovozněna zásadní funkce – vyhledávání podle obcí. Uživatel si musí hledanou obec nejprve najít v inventáři zveřejněném ve formátu PDF, aby zjistil název původce (obvykle farnost) a signaturu příslušné knihy.

Inspirován Moravským zemským archivem (MZA) připravuje i Státní oblastní archiv v Plzni několik projektů přeshraniční spolupráce.

2010-2012: V rámci prvního projektu „Bavorsko-česká síť digitálních historických pramenů“ vznikl webový portál portafontium.eu, na kterém jsme ihned začali zveřejňovat velké množství digitalizovaných archiválií včetně jejich základních popisů.

2013-2015: Navazující projekt „Česko-bavorský archivní průvodce“ se bude zabývat vyhledáním a popisem dalších materiálů a fondů, a to nejen v západočeských archivech, ale na území celé České republiky a Bavorska.

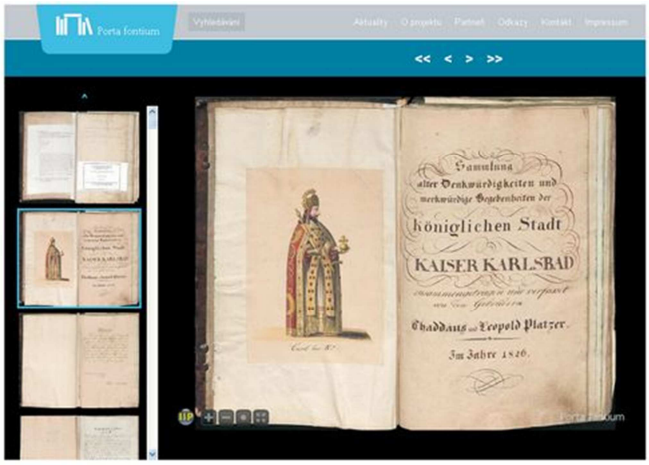
Cílem našich projektů je především zmapovat a zpřístupnit archivní fondy a dokumenty uložené v bavorských archivech týkající se české historie a naopak české fondy a dokumenty vztahující se k osobám, místům a jevům týkajícím se Bavorska.

Dojde také k virtuálnímu propojení archivních fondů, které byly v minulosti rozděleny a nyní leží v různých archivech na obou stranách hranice. Příkladem jsou zveřejněné listiny z kláštera Waldsassen, kde 1606 listin je ze státního archivu Amberg a 57 listin ze Státního okresního archivu Cheb. Obdobně 9556 fotografií z fondu Sudetoněmeckého archivu v Mnichově doplňuje zatím 5416 fotografií z našich okresních archivů (hlavně Cheb a Karlovy Vary).

Na webu Porta fontium jsme také zveřejnili již přibližně 5100 kronik (600 tisíc snímků) ze všech západočeských okresních archivů. Z fondu Sudetoněmeckého archivu v Mnichově je přístupných 22 kronik (některé jsou vícesvazkové).

Kromě práce v navazujícím přeshraničním projektu chceme v příštím roce web Porta fontium i software, na kterém je postaven, dále vyvíjet a digitální archiv doplňovat i o materiál, který není přímo předmětem přeshraničních projektů.

Kromě vylepšování uživatelského komfortu aplikace zveřejníme společně s matrikami rozsáhlý rejstřík míst, který bude umožňovat i sledování správního a názvoslovného vývoje obcí, podobný seznam původců a napojení na databázi našich archivních fondů a pomůcek, kterou budeme importovat z databáze PEvA a doplňovat o schválené archivní pomůcky ve formátu PDF.



Do konce roku 2013 bude na portál přidáno všech přibližně 11 tisíc svazků a 1,25 milionu snímků matrik. Dokud s tím bude souhlasit MZA a nezmění se podmínky zveřejnění, plánujeme všechny matriky ponechat také na portále Acta Publica, kde navíc úspěšně funguje např. diskusní fórum. Oproti tomu nový portál Porta fontium nabízí kromě lepší technologie zobrazování snímků i možnost vyhledávání podle obcí, plánujeme také filtrování obsahu podle typu záznamu, vyhledávání podle časového rozsahu a další vylepšení.

V rámci nového projektu Česko-bavorský archivní průvodce připravujeme možnost podrobnějšího popisu struktury fondů v souladu s mezinárodními standardy ISAD(G) a EAD. V budoucnu lze též očekávat využití možnosti exportů pro zveřejnění dat prostřednictvím národních i evropských integrátorů (např. budoucí Národní digitální archiv, Europeana), záleží mimo jiné na tom, jak bude v tomto směru koordinován postup mezi archivy v České republice.

Zkušenosti s digitalizací

Následujících několik postřehů nemá zprostředkovat souvislý návod na digitalizaci, ale snad by někomu mohlo udělat radost, když se zde otevřeně podělíme o několik našich bolestně získaných zkušeností.

Snímání knih

Nejprve jsme knihy, podobně jako pro listiny, fotografovali. Zatímco u listin je náročnější příprava a zpracování snímku profesionálním fotografem ku prospěchu věci, u masové digitalizace knih je takovýto postup neefektivní. Snímky tedy zhotovují méně kvalifikovaní pracovníci a brigádníci. Aby výstupy byly ostré a rozměrově nezkreslené, bylo třeba vytvořit pracoviště se speciální podložkou, stativem, světly a vzdáleným ovládáním spouště. Podobné vybavení nadále využíváme pro obsluhu badatelských požadavků na kopie stránek z knih, a to i v okresních archivech mimo hlavní digitalizační pracoviště.

Při každodenním nepřetržitém několikahodinovém provozu se však fotoaparáty brzy opotřebovaly. Na konci roku 2010 jsme tedy využili příležitosti, kdy byl mimořádně k dispozici větší objem

finančních prostředků, a pořídili jsme si velký knižní skener formátu A1. Knižní skener přinesl výrazně lepší poměr kvality a rychlosti, takže počítáme-li s náklady na obsluhu, kontrolu a následnou úpravu snímků, je přes vysokou pořizovací cenu hospodárnější. Později si archiv, ať už z vlastního rozpočtu, nebo v rámci přeshraničního projektu, pořídil další dva menší knižní skenery formátu A2. Množství knih čekajících na skenování v oblastním archivu převyšuje životnost několika knižních skenerů, takže tato zařízení bude třeba nadále obměňovat nebo doplňovat. Přestože jsou knižní skenery určeny k intenzivnímu provozu, jsou poměrně poruchové a náhradní díly včetně jejich výměny jsou drahé. Proto se podle našich zkušeností vyplatí požadovat při nákupu co nejdelší záruční dobu, alespoň 5 let. Použití knižního skeneru pro snímání jiného materiálu, než knih, je nehospodárné. Také případné pořizování snímků knih na zakázku mimo archiv by při velkém množství materiálu bylo příliš nákladné.

Nesmí se podcenit zaučení obsluhy skeneru a následná kontrola snímků v co nejbližším časovém odstupu. V plzeňském archivu tuto práci provádí zpravidla mimo badatelské hodiny a dny obsluha badatelny. Nevyplatí se opravovat chyby až poté, co je odhalí badatelé.

Úprava snímků

Ořez lze nastavit přímo na skeneru. Jakákoli další úprava snímků – jas, kontrast, vyvážení barev, softwarové vyrovnání zakřivení knihy – se ukázala jako neefektivní. Hromadná úprava snímků přináší riziko přehlédnutí jejich znehodnocení, každá následná oprava je velmi nákladná.

Ukládání snímků

Po špatných zkušenostech s neprověřeným softwarem (tzv. „modul Datový sklad“) a levnými síťovými diskovými poli (NAS) jsme pořídili výkonná disková pole s možností velkého rozšíření kapacity. Snímky ukládáme pomocí klasického souborového systému, zálohujeme na jiné diskové pole a na pásky. Předpokládáme, že pro tzv. dlouhodobé či trvalé uložení snímků jednou využijeme Národní digitální archiv.

Označování snímků

Snímky pojmenováváme kombinací čísel i slov tak, aby byly jednoznačně a snadno identifikovatelné i pro člověka, bez nutnosti použití nějaké databáze. Veškerá manipulace se snímky je prováděna „ručně“ bez speciálního informačního systému, lze je tak přímo za využití běžných programů pro prohlížení snímků prezentovat např. v badatelnách. Změnu tohoto systému zatím neplánujeme, i když se ukázalo, že má i své nevýhody.

Popisování snímků

Činnost velmi náročná na čas a na odbornost osob pořizujících metadata, zvláště u dosud neinventarizovaných fondů. Při dlouhodobější spolupráci lze částečně využít brigádníky, např. studenty archivnictví nebo historie, brigáda časem může přejít v klasický pracovní poměr.

Chceme-li být dostatečně rychlí, musíme dobře zvážit rozsah popisných dat. Je-li to možné, je třeba co nejlépe promyslet a zpracovat metodiku, včetně příkladů. Nemusí být komplikovaná, stejně nepostihne vše a v praxi se objeví nepředpokládané situace, pozor spíše na efektivitu. Je třeba se smířit s tím, že časová a personální kapacita archivu dnes nestačí na rozsáhlý popis veškerého

materiálu. Je však dobré takový popis alespoň umožnit, a to jak do budoucna, tak s využitím již dříve vypracovaného popisu některých dokumentů, třeba v papírové podobě.

Import dat z existujících tabulek a databází z různých zdrojů (okresních archivů) přináší časově náročné úpravy dat, pozor též na spárování se snímky, které zpravidla pořizuje někdo jiný, než metadata, u nás většinou i na jiném pracovišti a se značným časovým odstupem. Pokud metadata stejného typu pořizuje více editorů, nevyhneme se nejednotnosti, která ve sloučené databázi nepůsobí dobře. Proto je výhodné pověřit jednoho stálého člověka revizí takových dat.

Aby byla i do budoucna využita informační hodnota digitálního archivu, je podle mého názoru dobré nešetřit prostředky a čas na propojení s jinými zdroji dat, u archivního materiálu především s co nejobsáhlejším místním rejstříkem a se souvisejícími archivními pomůckami. Ani v tomto případě ale nebudou zřejmě nikdy dostatečně uspokojeny nároky teoretiků na vyčerpávající popis ani představy laické veřejnosti o snadnosti získání hledaných informací.

Aby nebyla za několik let vynaložená práce znehodnocena, od začátku se též vyplatí popis strukturovat tak, aby byl v souladu s perspektivními mezinárodními standardy. Spoléhat se na dlouhodobé trvání případných unikátních, s ničím nekompatibilních národních standardů se mi v této souvislosti jeví jako naivní. Stejně tak jsme, v tomto případně vskutku bolestně, získali nedůvěru k uzavřeným softwarovým aplikacím s neznámou, neprůhlednou a nepružnou datovou strukturou.

Úskalí vývoje software

Nakonec několik poznatků ze samotné tvorby softwarového projektu, který je použit mimo jiné jako základ webu Porta fontium. Pracovní skupina archivářů a informatiků navrhla zcela jinou datovou strukturu, než u předchozího software, který SOA v Plzni spoluvytvářel. Je třeba do ní převést velké množství dat z různých databází a tabulek a umožnit vkládání dalších vazeb mezi nimi. Přitom je nutno počítat s tím, že rozsah informací není předem dán a vždy se bude měnit podle charakteru a komplexnosti popisu vkládaných dat. Pozornost je třeba věnovat nejen importům, ale i budoucím exportům dat do jiných systémů a portálů, opět v rozsahu a formátu, který není předem zcela znám.

Výběr platformy pro vývoj

U intranetu a webových stránek se nám vyplatilo použít open source platformu pro správu obsahu Drupal. Díky neustálému vývoji a bezpečnostním aktualizacím komponent systému je tak zajištěno udržení stability a bezpečnosti v dalších letech a také možnost rozšiřitelnosti, protože po celém světě se vyvíjí nepřeberné množství rozšiřujících modulů, které je možno poměrně snadno využít.

V případě aplikace pro zpřístupňování dokumentů je otázka, zda by nebylo lepší použít nějaký rozšířený existující specializovaný open source software používaný u nás nebo v zahraničí pro zpřístupňování digitalizovaných dokumentů. My jsme se zalekli jejich komplikovanosti a zároveň poměrně úzké specializace, možných problémů s nasazením a případným rozšiřováním funkcí, nutným např. k převodu stávajících elektronicky pořízených dat, a obtížnosti sehnat firmu, která by byla schopná nám pomáhat s implementací a úpravou takového software, protože sami jsme v tom zkušenosti neměli a tlačil nás čas. Dnes už je situace na trhu trochu lepší, kromě toho jsme zjistili, že

mnoho firem uvádí, že umí námi zvolený obecnější systém Drupal, ale je problém sehnat firmu, která by byla schopná v něm programovat speciální moduly.

Abychom nebyli svázáni možnostmi a funkcemi existujícího software, bylo by jistě jednodušší vytvářet program zcela od začátku, na míru svých požadavků, např. s využitím nějakého PHP frameworku, jako to udělal Moravský zemský archiv v projektu Acta Publica nebo SOA Třeboň se svým digitálním archivem. Přišli bychom však o výhodu vývoje podstatných částí systému komunitou, tedy z prostředků jiných institucí a firem, včetně zajišťování dostatečně častých bezpečnostních aktualizací a snadné rozšiřitelnosti.

Z hlediska naší spolupráce s univerzitou a využití studentů informatiky by bylo lepší zaměřit se na vývoj software v programovacím jazyce Java nebo .NET, v PHP dnes už programuje málo studentů.

Kromě systému Drupal využíváme pro zobrazování snímků image server IIP Image, což se ukázalo jako šťastná volba, kvalitativně převyšující načítání celých snímků jako na webu Acta Publica nebo Monasterium. Navíc používání obrazového formátu JPEG2000 s pyramidovou strukturou není tak těžkopádné a náročné na souborový systém, jako uchovávání předzpracovaných snímků pro zobrazování podobnou technologií Zoomify (v případě on-line generování sady snímků pro Zoomify by zase hrozila neúnosně pomalá odezva webové aplikace).

K dalším výhodám zvoleného image serveru patří, že lze v jednotném uživatelském prostředí zobrazovat i snímky ze vzdáleného serveru, což se chystáme brzy zprovoznit i na webu Porta fontium.

Vyskytly se jen problémy se snímky vygenerovanými programem Photoshop, lepší je využívat např. knihovnu Kakadu.

Problémy a rizika

Nečekaně velké problémy přinesl import databáze obcí bývalého modulu Matriky. Struktura databáze byla tak nepřehledná a nekonzistentní, že si s ní neporadili ani v MZA, ani jeden náš studentský tým a byla objasněna až letos ve spolupráci s dalším studentským týmem. Kvůli tomu se výrazně zdrželo zveřejnění podstatné části systému, protože na rejstřík obcí jsou navázána prakticky všechna ostatní data. Aplikace Porta fontium je proto dlouhou dobu dočasně provozována v omezeném režimu nevyužívajícím všechny její možnosti. Náročný byl také relativně brzký přechod na novou verzi Drupalu a kompromisy s tím spojené.

Samotné vedení softwarového projektu je dosti náročné na čas a na disciplínu administrátorů i přispívajících vývojářů z komerčních firem nebo z řad studentů, kteří zde mají za úkol prakticky si vyzkoušet právě týmovou práci na reálném softwarovém projektu. Nelze než si přiznat, že v této oblasti máme ještě hodně co dohánět.

Přesto lze zatím spolupráci s každý rok jiným týmem vývojářů, ať už z řad studentů či komerčních firem, považovat za úspěšnou. Velkým rizikem, které se nám zatím nepodařilo odstranit, se ale jeví přílišná závislost projektu na omezeném množství informatiků SOA v Plzni. Pokud by se tento stav nedařilo zlepšit, bylo by odpovědné začít včas připravovat přechod a převod všech dat na vhodný

rozšířenější open source systém, v ideálním případě společně s dalšími archivy, které by takovou potřebu měly.

Ani zprvu slibně započatá větší týmová spolupráce s odbornými archiváři při vývoji aplikace se následně z různých důvodů nerozvíjela tak, jak by nám bylo milé.

Případnou spolupráci s dalšími partnery, informatiky jiných archivů nebo paměťových institucí patrně nelze v takto rané fázi vývoje projektu zatím předpokládat, i když bychom za ni byli rádi. Vzhledem k malé cílové skupině bude stálá finanční a personální podpora alespoň jednoho archivu nezbytná vždy, což je zřejmé třeba na příkladu projektu Open source spisové služby, která má po ztrátě podpory ze strany Ministerstva vnitra i přes svou relativní rozšířenost velké problémy dostát svého původního cíle, tedy kompatibility s národními standardy.

Doufejme, že i další léta budou otevřena efektivnímu zpřístupňování a tím i ochraně našich archiválií a s nimi zůstane v nějaké formě uchována a využívána i mnohaletá odborná práce našich archivářů.

Děkuji za pozornost

Petr Kocourek, Státní oblastní archiv v Plzni, informatik@soaplzen.cz