ÚSTAV INFORMAČNÍCH STUDIÍ A KNIHOVNICTVÍ
FF UK V PRAZE

Jan Pokorný

Technické nástroje integrace informačních zdrojů

Verze 1.0

Praha

Říjen 2007

1 TECHNICKÉ NÁSTROJE INTEGRACE INFORMAČNÍCH ZDROJŮ

V této práci si ukážeme některé technologie, které lze využít při integraci informačních zdrojů. Není možné zde zmínit, natož podrobně popsat všechny dostupné technologie. Záměrně jsou zde vynechány konkrétní produkty jednotlivých výrobců software. Důvodem je velká proměnlivost nabídky a snaha neprosazovat konkrétní řešení. Chybí zde proto přehled vyhledávacích a databázových strojů, metavyhledávačů, portálových řešení a dalších informačních systémů, přestože mohou v konkrétních projektech vystupovat jako dílčí komponenty zajišťující svěřené funkce. Dnešní dynamické prostředí mnohdy vyžaduje modulová řešení (tzv. suity). Jsou to skládanky sdílených a integrovaných komponent různých výrobců, protože ve většině případů není možné najít produkt, který by beze zbytku splňoval všechny požadavky projektu.

1.1 VÝZNAM TECHNICKÝCH STANDARDŮ

Jedním ze základních předpokladů integrace heterogenních informačních zdrojů a služeb je schopnost vzájemné spolupráce neboli interoperabilita. Schopnost spolupráce v technickém smyslu nevzniká náhodně, ale je definována tím, nakolik dané technologie či služby odpovídají dohodnutým specifikacím. Souborům specifikací pro interoperabilitu říkáme standardy. Standardy musí být dostatečně volné, aby vyhovovaly všem zúčastněných, ale dost přísné, aby zajistily potřebnou kvalitu. Standardy musí být schváleny a přijaty buď velkou skupinou zainteresovaných stran (např. výrobců hardware) nebo národní či mezinárodní standardizační organizací, jako jsou International Organization for Standardization (ISO), American National Standards Institute (ANSI), European Committee for Standardization (CEN), InterNational Committee for Information Technology Standards (INCITS) či World Wide Web Consortium (W3C).

Některé standardy mohou být patentovány a chráněny, naopak některé jsou zcela volné a mohou být využity kýmkoli, kdo se pro ně rozhodne. Takovýmto volným standardům říkáme **otevřené standardy** a právě ony jsou nejvýznamnějšími hybateli technologického pokroku v systémové integraci.

Opakem otevřených standardů jsou tzv. **proprietární technologie**, které vznikají většinou v rámci jediného výrobce nebo jednoho projektu šité na míru. Proprietární technologie je uzavřená, vyznačuje se výraznými privátními specifiky, které brání bezešvé komunikaci a propojování s okolím.

Při projektování IS pro integraci informačních zdrojů stojíme vždy před základní otázkou, zda vyvinout proprietární řešení nebo využít technologie postavené na otevřených standardech. Proprietární technologie může být pro konkrétní situaci vhodnější, může být i levnější a krátkodobě efektivnější. Existují však vážné důvody, proč se přiklánět ke standardizovaným řešením:

- **nezávislost na technologii, produktu či službě jediného výrobce** – výrobce může zaniknout, ukončit podporu dané technologie či znevýhodnit obchodní podmínky
- **otevřené standardy umožňují svobodně vybírat jednotlivé komponenty a kombinovat je** – v integračním řešení můžeme používat dílčí technologie různých výrobců a propojovat je mezi sebou, čímž můžeme využít těch nejlepších vlastností
- **standardy v sobě obvykle odrážejí nejlepší možná řešení** – standardy vznikají na základě zkušeností odborníků v dané oblasti, před schválením jsou dlouhou dobu připomínkovány – pokud neexistují vážné důvody k proprietárnímu řešení, je nejjednodušší používat standardy, které jsou zárukou kompatibility s dalšími prvky v okolí, navíc nemusíme vynakládat finance na vývoj vlastních řešení
- **standardy jsou postupně vyvíjeny** – vznikající nové verze otevřených standardů jsou impulsem k progresivním změnám řešení u všech výrobců bez nutnosti přímé koordinace

Pokud již IS nepoužívá standardizované technologie na všech vrstvách, měl by být alespoň vybaven standardizovanými vstupy a výstupy, aby mohl systém spolupracovat s jinými IS. Příkladem může být knihovnický systém, který sice ukládá záznamy ve svém vlastním vnitřním formátu, ale na úrovni funkcí pro import a export dokáže zpracovat záznam v MARC 21.

Následující tabulka shrnuje nejvýznamnějšími otevřené standardy používané v IT v knihovnách:

Komunikační protokoly	HTTP, Z39.50, FTP
Protokoly speciálních služeb	OAI-PMH, openURL, EDI, SRU/SRW
Dotazovací jazyky	CCL, CQL, Z39.50, SQL
Formáty popisných dat	MARC 21/MARCXML, DC, MODS, MADS, METS
Předmětové systémy	Konspekt, MDT, DDT
Formáty plných textů	HTML/CSS, PDF, RTF
Formáty grafických objektů	JPEG, PNG, TIFF, DjVu, SVG
Formáty zvukových objektů	WAV, MP3, AU, MIDI
Formáty video objektů	AVI, MPG, DVD
Autentikace	LDAP, EAP, Shibboleth, SAML

1.2 TECHNOLOGIE PRO VYHLEDÁVÁNÍ A KOMUNIKACI SE VZDÁLENÝMI INFORMAČNÍMI ZDROJI

V kapitole **Chyba! Nenalezen zdroj odkazů.** jsme mluvili o centrálně a distribuovaně pojatých způsobech vyhledávání. Oba dva způsoby mají své výhody i nevýhody a jejich nasazení je často podmíněno nejen technickými požadavky, ale i politikami spolupracujících subjektů. Díky internetizaci knihoven, která zajistila většině knihoven permanentní připojení na internet, se v minulých letech výrazněji prosadily technologie distribuovaného vyhledávání, zejm. protokol Z39.50. V současné době je naopak cítit jistá tendence návratu k centralizovaným typům vyhledávání, což je způsobeno rozvojem elektronických archivů a digitálních repozitářů, které používají ke spolupráci s okolím protokol OAI-PMH. Protokol Z39.50 se navíc může zdát z dnešního pohledu poněkud zastaralý, protože na rozdíl od modernějších technologií neinklinuje k webovému pojetí služeb, které jsou postaveny na komunikačním protokolu HTTP a formátu XML. Standard SRU/SRW, který byl vyvinut jako XML nástupce protokolu Z39.50, se zatím v praxi příliš nerozšířil, nicméně mnoho vývojářů v něm vidí cestu dalšího vývoje.

Podívejme se proto na tyto protokoly podrobněji.

1.2.1 Z39.50

Komunikační a aplikační protokol Z39.50 pro vyhledávání a přejímání dat jsme v této práci již mnohokrát zmiňovali. V současné době totiž představuje nejrozšířenější technologii pro distribuované vyhledávání v knihovnických aplikacích a informačních zdrojích. Hlavní výhoda Z39.50 spočívá v tom, že nabízí zcela **nezávislou abstraktní vrstvu**, která umožňuje propojit v podstatě jakékoli informační zdroje mezi sebou či s integračním prvkem, a to nezávisle na jejich operačních systémech a na jejich databázové a aplikační vrstvě. Vzhledem ke svému masivnímu rozšíření a širokým možnostem využití nemá protokol Z39.50 pro distribuované vyhledávání v současné době konkurenci.

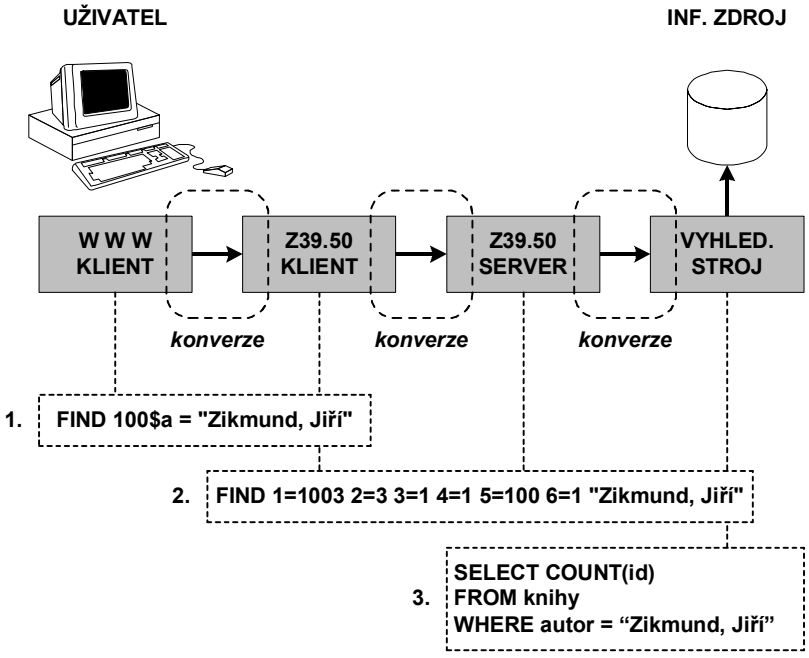
Aktuální verze Z39.50 pochází z roku 2003 (označována jako Z39.50-2003). V praxi je však nejrozšířenější starší verze **Z39.50-1995**, která byla v roce 1997 normalizovaná jako ISO 23950. Standard Z39.50 definuje kromě komunikační a aplikační funkcionality také vlastní dotazovací jazyk, takže zajišťuje nezávisle na jiných technologiích všechny funkce pro vyhledávání. Jádrem standardu Z39.50-1995 je definice 6 typů dotazů a 6 různých množin datových atributů. Pro oblast knihoven je nejvyužívanější typ dotazů Type-1 (kde jsou termy kombinované booleovskými operátory) a množina atributů Bib-1 (odpovídající bibliografické struktuře záznamů). Protože specifikace protokolu Z39.50 je poměrně rozsáhlá a otevřená, je

třeba ke vzájemné spolupráci definovat tzv. **profily Z39.50**, které říkají, jaké atributy a jaké možné hodnoty budou v konkrétních projektech minimálně podporovány. Součástí profilu mohou být také předepsané způsoby indexování jednotlivých polí, která jsou na Z39.50 atributy mapovány, aby všechny zúčastněné IS vracely srovnatelné výsledky.

Při komunikaci Z39.50 vystupuje informační zdroj jako Z39.50 server a prvek, který ho využívá, vystupuje jako Z39.50 klient. Komunikace mezi klientem a serverem je přesně specifikována protokolem, který kromě vyhledávání podporuje i další operace, včetně řazení, editace či mazání dat. Komunikace vypadá zjednodušeně takto:

- Systém, který vystupuje jako Z39.50 klient, nejprve převede konkrétní dotaz do syntaxe Z39.50 na základě mapování (např. z formátu MARC 21 nebo z parametrů zadaných uživatelem v uživatelském rozhraní). Dotaz je ve výsledku složen ze seznamu databází, ve kterých má být hledání provedeno, a z vyhledávacích výrazů (termů) s uvedenými hodnotami Z39.50 atributů pospojovaných booleovskými operátory.
- Klient se následně spojí se serverem a po úvodní výměně parametrů pošle na server dotaz Z39.50.
- Server dotaz přijme a následně ho přemapuje do interní syntaxe informačního zdroje, např. do SQL. Poté provede vyhledávání v příslušném systému správy bází dat.
- Vrácená množina výsledků je převedena do formátu, který klient očekává, a je odeslána buď jako celek nebo po částech zpět klientovi.
- Klient vrácené výsledky zpracuje a použije v daném kontextu na výstupu.

Na následujícím obrázku je příklad Z39.50 komunikace. Dotaz na pole 100$a formátu MARC 21 s výrazem „Zikmund, Jiří" je přemapován do syntaxe Z39.50 s typem dotazu Type-1 a množinou atributů Bib-1 specifikujícími použitá pole, strukturu, zarovnání atd. Server, který dotaz přijme, provede konverzi do syntaxe jazyka SQL a následně provede vyhledávání v databázi.



Obr. 1 – Komunikace Z39.50

Přesnou specifikaci normy Z39.50 lze získat na stránkách Knihovny amerického Kongresu[8], která se prostřednictvím agentury „Z39.50 Maintenance Agency" stará o další vývoj standardu Z39.50 a poskytuje potřebnou podporu. V ČR se o propagaci a koordinaci využívání Z39.50 stará sdružení ZIG-CZ (Z39.50 Implementors Group)[7], které zaštiťuje Státní technická knihovna. Toto sdružení v únoru 2004 vyhlásilo ve spolupráci s Národní knihovnou ČR **Z39.50 profil JIB**[9], který

vymezuje minimální požadavky na nastavení Z39.50 atributů u českých katalogů a bibliografických databází pro účely jednoduchého vyhledávání.

Ve snaze přizpůsobit protokol Z39.50 modernímu webovému pojetí služeb, zjednodušit ho a zachovat všechny jeho osvědčené výhody, vznikla iniciativa Z39.50 International - Next Generation (ZING), která připravila dva nové standardy SRU a SRW. Oproti Z39.50 představují SRU/SRW často levnější a univerzálnější řešení, zejm. s ohledem na vyhledávání v prostředí WWW, je však otázkou, nakolik je budou producenti AKS a databází ochotni implementovat do svých IS.

1.2.2 SRU/SRW

Informační technologie, na kterých je založen provoz informačních zdrojů, v poslední době směřují k maximálnímu zjednodušení a unifikaci. Prosazují se technologie, které se již dříve dobře osvědčily u služby WWW, tedy komunikační protokol HTTP a značkovací jazyky, především XML. Ukázalo se, že kombinace HTTP/XML je natolik univerzální a jednoduchá, že na ní lze postavit obecně jakékoli informační služby. Tento trend neminul ani protokol Z39.50. V listopadu 2002 vyvrcholila snaha již zmíněné iniciativy ZING a vznikla první verze protokolů SRU a SRW - nových standardů pro distribuované vyhledávání v informačních zdrojích. V současné době jsou k dispozici ve verzi 1.1.

SRU (Search/Retrieve via URL) je vyhledávací protokol, který zajišťuje vyhledávání a získávání dat z informačních zdrojů v prostředí WWW. Dotaz vyjadřuje standardizovaným dotazovacím jazykem CQL (Common Query Language) a kóduje ho do URL. Myšlenka SRU vychází ze starší technologie openURL, která taktéž k zasílání bibliografických údajů využívá URL. URL se skládá z adresy SRU serveru, který přijímá požadavky, a z části, která obsahuje samotný CQL dotaz v předepsané syntaxi. Adresa a vyhledávací část jsou odděleny znakem „?", jednotlivé parametry dotazu potom znakem „&". Parametry jsou tvořeny dvojicemi „klíč=hodnota", kde klíče předepisuje norma SRU. Podporovány jsou tři základní operace: explain (pro získávání informací o databázi a podporovaných vlastnostech), searchRetrieve (zajišťující vlastní vyhledávání) a scan (k procházení rejstříku databáze). SRU umožňuje zasílat data v kódování Unicode, což zajišťuje jeho univerzálnost. Jednoduchý SRU dotaz může vypadat např. takto:

http://z3950.loc.gov:7090/voyager?version=1.1&operation=searchRetrieve&query=olomouc

Výsledky jsou ze serveru vraceny ve formátu XML v požadovaném schématu (MARCXML, Dublin Core, MODS a další) a obsahují údaje v závislosti na volané operaci. Kromě vyhledávání podporuje SRU také řazení výsledků, specifikaci CSS stylu pro zobrazení, autentikaci a další podpůrné funkce. Celý SRU dotaz lze posílat zcela standardně metodou GET nebo POST jako běžné URL, takže nevznikají žádné komplikace při síťové komunikaci.

SRW (Search/Retrieve Web Service) je obdobou SRU. Komunikace však není realizována prostřednictvím URL, ale pomocí XML zpráv vyměňovaných mezi klientem a serverem prostřednictvím protokolu HTTP. Způsob prezentace XML zpráv je definován standardem webových služeb SOAP (Simple Object Access Protocol) podle agentury W3C. Z toho vyplývá, že jak SRW server, tak i SRW klient musí podporovat SOAP, konkrétně verzi minimálně 1.1. Syntax dotazu i odpovědi je stejný jako u SRU, s výjimkou názvů operací a možností volat CSS styl. Dotaz v podobě SRW může vypadat takto:

```
<SOAP:Envelope xmlns:SOAP="http://schemas.xmlsoap.org/soap/envelope/">
  <SOAP:Body>
    <SRW:searchRetrieveRequest xmlns:SRW="http://www.loc.gov/zing/srw/">
      <SRW:version>1.1</SRW:version>
      <SRW:query>(dc.author exact "svoboda" and  dc.title >= "buzuluk")</SRW:query>
      <SRW:startRecord>1</SRW:startRecord>
      <SRW:maximumRecords>10</SRW:maximumRecords>
      <SRW:recordSchema>info:srw/schema/1/mods-v3.0</SRW:recordsSchema>
    </SRW:searchRetrieveRequest>
  </SOAP:Body>
</SOAP:Envelope>
```

Vývoj a podporu obou standardů zajišťuje agentura „SRU Maintenance Agency" při Knihovně amerického Kongresu. Přestože je název jiný, jedná se ve skutečnosti o stejnou agenturu jako pro protokol Z39.50. Na stránkách agentury lze získat další informace i přesné specifikace standardů.

1.2.3 OAI-PMH

Zatímco protokoly Z39.50 a SRU/SRW slouží k prohledávání vzdálených informačních zdrojů, pomocí nichž můžeme budovat distribuované IS (např. virtuální souborné katalogy), protokol OAI-PMH reprezentuje zcela opačnou filosofii – je to nástroj, který umožňuje integračnímu systému získávat (sklízet) popisné údaje (metadata) z jednotlivých informačních zdrojů a centralizovaně je ukládat. V kontextu integrace informačních zdrojů umožňuje budování **fyzicky centralizovaných úložišť popisných údajů** informačních objektů a následné vytváření vyhledávacích indexů. Vyhledávání pak probíhá přímo v integračním IS na vybudovaných indexech, bez účasti vzdálených informačních zdrojů.

Protokol **OAI-PMH** (Open Archives Initiative Protocol for Metadata Harvesting) slouží pro sklizeň popisných dat, nikoli pro sklizeň popisovaných informačních objektů jako takových. Primární informační objekty, pokud v elektronické podobě existují, zůstávají v úložišti vzdáleného informačního zdroje a jsou přenášeny až na vyžádání na základě linkování z popisného záznamu.

Veškerá komunikace protokolem OAI-PMH probíhá formou XML zpráv, které jsou posílány prostřednictvím HTTP mezi klientem (tzv. harvesterem) a serverem (tzv.repozitářem). Každý metadatový objekt je v repozitáři opatřen jednoznačným identifikátorem, který musí odpovídat specifikaci URI. Tento identifikátor je celosvětově jedinečný, protože kromě ID samotného objektu je jeho součástí doménové jméno příslušného serveru (např. oai:nkp.cz:340076). Standard OAI-PMH dále specifikuje, že povinným metadatovým schématem je nekvalifikované Dublin Core. Kromě toho však samozřejmě dovoluje podporovat i další metadatová schémata.

Protokol OAI-PMH nabízí pro účely sklizně šestici základních příkazů: **Identify** (slouží pro získání technických i popisných informací o repozitáři, které následně klient využije k nastavení dalších operací pro sklízení), **ListMetadataFormats** (slouží k vypsání seznamu všech podporovaných metadatových schémat), **ListSets** (slouží k vypsání seznamu všech skupin záznamů, které mohou být v repozitáři vytvořeny), **GetRecord** (slouží k získání konkrétního záznamu v požadovaném formátu podle uvedeného jednoznačného identifikátoru), **ListIdentifiers** (vrací seznam všech identifikátorů záznamů, který lze omezit uvedeném schématu, skupiny či časového rozpětí) a **ListRecords** (zajišťuje vlastní sklizeň, tj. vrací celé metadatové záznamy).

První stabilní verze protokolu OAI-PMH je verze 2.0 zveřejněná v červnu 2002 po několikaletém vývoji této technologie v rámci iniciativy Open Archives. Z původního záměru vytvořit řešení pro kooperaci mezi elektronickými archivy vědeckých prací se postupně vyvinula technologie otevřeného přístupu k jakýmkoli elektronickým zdrojům v oblasti digitálních knihoven, elektronických archivů nebo bibliografických či plnotextových databází. Dnes prostřednictvím protokolu OAI-PMH nabízejí svá data stovky informačních zdrojů po celém světě. Vzniká velká řada projektů, které sklízená metadata využívají pro tvorbu citačních rejstříků nebo centrálních indexů virtuálních digitálních repozitářů, pro integrované prohledávání informačních databází společně s webovými zdroji či pro následné zpřístupnění přes Z39.50, kdy budované IS slouží jako dílčí agregátory.

1.3 TECHNOLOGIE NA ZAJIŠTĚNÍ OCHRANY AUTORSKÝCH PRÁV A LICENCÍ

Jak jsme již zmínili v kapitole **Chyba! Nenalezen zdroj odkazů.**, je u řady informačních zdrojů vyžadována ochrana autorských práv, dodržování licenčních podmínek a uplatňování speciálních režimů distribuce dokumentů. Provozovatelé, producenti, nakladatelé či autoři se v licenčních smlouvách zaměřují zejm. na kontrolu **pořizování kopií dokumentů**, kontrolu **modifikace dokumentů** a na kontrolu **distribuce dokumentů/řízení přístupu**. Tytéž snahy nalezneme i u zdrojů poskytujících sekundární informace. Abychom mohli v integračním systému uživatelům přinášet primární dokumenty a obsahy databází, musíme tato práva a požadavky zajistit.

Zatímco u papírových kopií dokumentů platí jasná pravidla daná fyzickou podstatou nosiče, zajištění práv u elektronických dokumentů je problematičtější. Je třeba nasadit technologie, které dokáží řídit kopírování a modifikování dokumentů a kontrolovat přístup k jejich obsahu. Problém je třeba řešit ve dvou rovinách:

- na úrovni zabezpečení přístupu k dokumentu v integračním systému (ochrana **online**)
- na úrovni práce s dokumentem v rukách uživatele (ochrana **offline**)

Jinými slovy, ochrana online zabezpečuje **přístup k dokumentu**, ochrana offline pak zabezpečuje **přístup k obsahu** dokumentu. Nejprve je třeba zajistit, aby se k daným informačním zdrojům na integračním serveru dostali pouze ti uživatelé, kteří k tomu mají oprávnění. Když oprávněný uživatel dokument získá a stáhne si ho na svůj lokální počítač, je třeba dále zajistit, aby uživatel nemohl dokument neoprávněně kopírovat nebo ho dále distribuovat, přivlastnit si autorství, popř. aby po vypršení oprávnění nemohl s dokumentem dále pracovat. V mnoha případech je cílem distribuce dokumentu pouze transport k uživateli pro účely **jednorázového tisku**, po kterém by měl dokument **expirovat**.

Ochrana online spočívá hlavně v zajištění bezpečné autentikace (viz kapitola **Chyba! Nenalezen zdroj odkazů.**) a celkového zabezpečení integračního systému (viz kapitola **Chyba! Nenalezen zdroj odkazů.**), tedy serveru a připojených prvků. V praxi je bohužel pozornost často zaměřena pouze tímto směrem.

Ochranu offline, tedy ochranu samotných dokumentů po stažení z integračního serveru, kdy již nelze nijak ovlivnit další distribuci, zajišťují **technologie dokumentových formátů**. V řadě realizovaných řešení je tento aspekt zcela opomíjen a ochrana obsahu dokumentů je ignorována. Dokumenty jsou tak často dále distribuovány ilegálními cestami, využívány neoprávněnými osobami a nekontrolovaně modifikovány a kopírovány. V dokumentech je proto nutné pro zajištění práv a licencí zajistit trvalou informaci o nositeli autorských práv, řídit přístup k obsahu dokumentu, řídit práva k manipulaci s dokumentem, řídit čas expirace s následným znemožněním přístupu a zajistit kompaktní formátování dokumentu.

> Řada technologií na ochranu autorských práv a licencí využívá šifrovacích algoritmů, které se dělí na symetrické a asymetrické.
>
> **Symetrické šifrovací algoritmy** používají stejný klíč pro šifrování i pro dešifrování. Klíč musí zůstat utajen, aby se neoprávněné osoby nemohly dostat k šifrovanému obsahu. Klíč musí být často střídán a musí být dostatečně náhodný.
>
> **Asymetrické šifrovací algoritmy** používají k šifrování a dešifrování vždy různé klíče, z nichž jeden je veřejný a druhý soukromý. Soukromý klíč je číselná hodnota chráněná heslem, umožňující autorovi podepisovat dokumenty. Veřejný klíč autora se přibaluje k podpisu a slouží k identifikaci a ověření autentičnosti původce/podpisu na straně příjemce. Veřejný klíč příjemce může autor použít k přidělení oprávnění pracovat s šifrovaným dokumentem. Veřejné klíče nemusí být utajovány, protože bez příslušných soukromých klíčů nemohou odkrýt obsah dokumentů.

1.3.1 Označení původce dokumentu a vlastníka autorských práv

Nejběžnějšími technologiemi, které umožňují do dokumentů zapisovat a následně v nich identifikovat původce, jsou metadata, digitální podpisy a vodoznaky. Všechny tyto metody lze využít v integračním systému jako součást procesu publikování a distribuce.

Metadata jsou popisné údaje o obsahu a charakteru dokumentu a mohou být uložena přímo v datovém souboru nebo externě mimo soubor. Možnost zápisu metadat přímo do datových souborů se objevuje v nových verzích téměř všech moderních dokumentových formátů, a to jak textových, obrazových, zvukových tak i ve videoformátech. Metadata mohou kromě jiného obsahovat údaje o původci dokumentu, nositeli autorských práv, možnostech užití, roku vzniku atd. a lze je proto využít jako pomůcku pro ochranu autorských práv. Příkladem mohou být metadata ve formátu MS Word, PDF, JPG (metadata EXIF nebo IPTC), MP3 (metadata ID3) atd. Metadata jsou editovatelná, což samo o sobě nezajišťuje účinnou ochranu, protože je může

uživatel přepsat nebo odstranit. Výjimkou jsou formáty schopné nastavit vnitřní práva pro práci se souborem a tak i znemožnit editace metadat, jako je např. PDF.

Vodoznaky u elektronických dokumentů jsou založeny na podobném principu jako vodoznaky na papíře. Jedná se o aplikaci grafického vzoru, obrázku nebo textu na pozadí nebo popředí textového nebo obrazového dokumentu, často s vysokou mírou průsvitnosti, aby vodoznak neznemožnil prohlížení dokumentu. Jejich účelem je označit a identifikovat dokument a vlastníka autorských práv. Při správně zvolené aplikaci jsou relativně těžko odstranitelné, a proto z hlediska uchovávání informací o autorských právech značně efektivní. Nevýhodou aplikace vodoznaků je změna původního obsahu, případně i snížená čitelnost. Vodoznaky lze do dokumentů snadno vkládat v běžných textových/grafických editorech nebo ve speciálně k tomu vybavených systémech.

Pokročilejší a účinnější variantou vodoznaku je **digitální vodoznak**, jehož prostřednictvím je do obrazového, ale i zvukového nebo video dokumentu vložena informace o autorských právech v podobě okem neviditelného nebo uchem neslyšitelného šumu. Důležité je, že vložený digitální vodoznak nijak nenarušuje kvalitu obsahu daného dokumentu. K výhodám této technologie patří snadná detekce a relativně vysoká odolnost vodoznaku, který se neztrácí ani při výrazných modifikacích dokumentu, jako je komprese nebo převzorkování (odolnost závisí na použitých metodách vkládání). Různá aplikační řešení navíc umožňují prohledávat určené části internetu pomocí robotů a zjišťovat, nevyskytují-li se někde na cizích serverech naše dokumenty v rozporu s licenčními podmínkami. Možnosti kontroly nad nepovolenou distribucí a neoprávněným použitím dokumentů jsou proto poměrně vysoké. Vývojem technologií digitálního vodoznaku se zabývá řada IT firem, nejznámější z nich je Digimarc s technologií MyPictureMarc[2], kterou podporuje i Adobe Photoshop. Firma Digimarc navíc nabízí i zajímavou technologii digitálního vodoznaku pro textové dokumenty na principu úpravy šířky mezer mezi slovy a řádkování.

Digitální podpisy představují asi nejelegantnější způsob označování a identifikace vlastníka autorských práv dokumentu. Technologie digitálního podpisu vznikly pro účely ověřování identity původce dokumentu, důvěryhodnosti informací a integrity dokumentu (možnosti sledovaní změn v dokumentu po jeho vzniku). Při použití digitálního podpisu se do dokumentu vloží jedinečný otisk se zašifrovanými číselnými údaji. Digitální podpis může informovat každého z uživatelů o subjektu, který daný dokument vytvořil, a umožnit tak kontakt s nositelem autorských práv nebo řešení případných sporů týkajících se autorství.[4] Digitální podpis vychází většinou z asynchronní koncepce tzv. soukromého a veřejného klíče, kterou můžeme dále rozdělit do dvou skupin. První skupina technologií vyžaduje existenci tzv. certifikační autority, tedy jakéhosi garantovaného vydavatele platných podpisů. Druhá skupina se obejde bez certifikační autority a je založena na tzv. přímé důvěře jednotlivých účastníků výměny dokumentů. Pro účely ochrany autorských práv jsou vhodnější technologie s účastí certifikační autority kvůli nezávislé garanci platnosti a pravosti podpisů, což může být vhodné při možných soudních sporech. Nejvíce rozšířené standardy pro digitální podpisy jsou DSS (Digital Signature Standard) používající algoritmus DSA (Digital Signature Algorithm) a PKCS (Public Key Cryptography Standards) používající mimo jiné nejpoužívanější šifrovací algoritmus RSA (Rivest, Shamir & Adelman). Hotová řešení pro podepisování dokumentů nabízejí např. firmy Adobe, Microsoft, Entrust, VeriSign a GeoTrust.

1.3.2 Řízení přístupu k obsahu dokumentu

Základem ochrany dokumentu před neoprávněným užitím je **šifrování jeho obsahu**. Slouží jak pro účely distribuce dokumentu, kdy není zajištěna zabezpečená komunikace a dokument se může dostat do rukou neoprávněných osob, tak pro účely práce s dokumentem offline. Úspěšnost šifrování je dána mírou **snížení šance prolomit ochranu**, tj. nalézt dešifrovací klíč. Šifrovacích mechanizmů je velké množství a stále vznikají dokonalejší (vývoj složitějších algoritmů je umožněn zejm. nárůstem výkonu počítačů). Šifrování a následné dešifrování se provádí pomocí hesla nebo jiného kódu, kterým může být např. bezpečnostní certifikát, podle toho, zda-li je použitá šifrovací metoda symetrická nebo asymetrická. Nástroje a řešení na šifrování dokumentů nabízejí např. firmy Adobe, OpenFile Systems nebo Microsoft.

Dokument může být šifrován již na straně vzdáleného informačního zdroje nebo ho může šifrovat až integrační systém. O odpovědnosti za obsah v integračních systémech rozhoduje způsob distribuce dokumentů. Získává-li uživatel dokument přímo ze vzdáleného informačního zdroje (integrační systém uživatele pouze přesměroval), nese odpovědnost za zabezpečení obsahu vzdálený zdroj. Pokud ale integrační systém vystupuje jako přímý distributor dokumentů, tj. uživatel získává dokumenty jeho prostřednictvím, musí integrační systém zajistit ochranu vlastními prostředky.

1.3.3 Řízení práv k manipulaci s dokumentem

Jakmile uživatel získá elektronický dokument do svých rukou, může s ním provádět řadu operací, které mohou být v rozporu s licenční politikou. Integrační systém by proto měl před vlastní distribucí dokument ošetřit, aby k porušování licencí nedocházelo. Je to jediný způsob, jak získat důvěru vydavatelů, producentů a provozovatelů v bezpečnou digitální distribuci dokumentů a navazující služby.

Klíčovými předměty ochrany jsou:

- **tisk** – uživatel může dokument vytisknout na tiskárně nebo mohou jiné aplikace využívat tiskový výstup k importu obsahu (aplikace využívající tiskové ovladače)
- **kopírování a vyjímání obsahu dokumentů** – uživatel může přes schránku OS kopírovat obsah do jiných aplikací (např. textového editoru) nebo mohou k obsahu přistupovat jiné aplikace a vytěžovat ho
- **změny v obsahu dokumentu** (editace) – uživatel může měnit texty, obrázky či jiné objekty v dokumentu, čímž může změnit nejenom podobu ale smysl dokumentu
- **změny v uspořádání dokumentu** – uživatel může přidávat, vyjímat či otáčet stránky nebo měnit jejich pořadí, čímž porušuje původní podobu originálního dokumentu

Výše uvedené operace s dokumentem je možné kontrolovat nastavením uživatelských práv, pokud to formát dokumentu umožňuje. Jediná operace, kterou nelze spolehlivě ošetřit, je sejmutí dokumentu na obrazovce jako snímek (screenshot), protože zobrazení dokumentu na obrazovce je předpokladem čtení a obrazovku lze snímat prostředky OS. Restrikce by měly být vždy založeny na šifrování dokumentu, aby byly účinné i v aplikacích jiných výrobců software, které nemusí tyto ochranné technologie respektovat. Stejně jako u řízeného přístupu lze šifrovat heslem (synchronní šifrování) nebo certifikátem (asynchronní šifrování).

Některé operace je však často výhodné pro dokumentové formáty uživatelům nebo jiným aplikacím umožnit. Patří sem např. zajištění přístupu k metadatům dokumentu i v případě, že je dokument šifrován, možnost vkládat uživatelských poznámek apod. Tyto vlastnosti podporuje např. formát PDF zmiňovaný v kapitole 1.3.6.

1.3.4 Řízení expirace platnosti a znehodnocení dokumentu

Přístup k obsahu dokumentu lze limitovat také z časového hlediska. Uživatel může získat oprávnění využívat obsah dokumentu pouze do určitého okamžiku. Tento okamžik označujeme jako expirace. Podpora řízení expirace dokumentu nachází uplatnění v celé řadě případů. Pro účely služeb integrovaného systému jsou to zejm.:

- **přístup k obsahu dokumentu pro účely pořízení tištěné kopie** – expirace je nastavena na krátkou dobu nebo pro jednorázové otevření, aby měl uživatel možnost dokument vytisknout (elektronická kopie slouží pouze pro účely distribuce)
- **realizace digitálních výpůjček** – jedná se o elektronickou podobu klasických výpůjček např. na 30 dní, poté je přístup k obsahu uživateli odepřen
- **vynucení aktualizace dokumentu online** – u dokumentů, jejichž obsah má význam pouze v případě, že je aktuální, lze omezením doby platnosti donutit uživatele pořizovat si pravidelně aktuální kopie

Technologie řízení expirace dokumentů jsou většinou založeny na **šifrování** obsahu a na otisku **časového razítka** (time stamp), což je časový záznam digitálně podepsaný důvěryhodnou třetí stranou, který vyjadřuje okamžik expirace. Na podobném principu je založena ochrana software

s časově platnými licencemi nebo ochrana dočasných práv na využití určitých, většinou placených služeb.

S expirací je spjat problém možného přenastavení systémového času na koncovém zařízení, kdy uživatel po vypršení oprávnění předstírá, že okamžik expirace ještě nenastal. Tento problém se pokouší řešit řada technologií, nejčastěji pomocí ověření skutečného času u tzv. autority časových razítek, která garantuje pravdivost časového údaje. Tyto metody však bohužel vyžadují účast vzdálených systémů, tedy předpokládají, že bude uživatel online a že ověření časového údaje povolí.

Účinné řízení expirace lze u dokumentů dosáhnout pomocí DRM systémů (Digital Rights Management), mezi které patří např. produkt OpenFile Publisher od firmy OpenFile Systems, PageRecall od firmy Authentica, WMDRM od firmy Microsoft nebo řešení LiveCycle od firmy Adobe.

1.3.5 Zajištění kompaktního formátování dokumentu

Snaha o zajištění kompaktního a neměnného formátování textu a grafiky v dokumentech je vždy výsledkem kompromisu. Na jedné straně se snažíme vyhovět licenčním podmínkám poskytovatelů a uživatelům zajistit dokument, který se za všech okolností zobrazí stejně a čitelně, na druhé straně musíme vyhovět různým potřebám uživatelů (poruchy zraku, obliba jiných stylů) a koncových zařízení (různá úroveň zobrazovacích schopností). Zatímco v letech, kdy se objevovaly první projekty pro zpřístupňování dokumentů přes internet, byly nejběžnějšími dokumentovými formáty HTML a prostý text, dnes jsou dokumenty šířeny mnoha různými formáty. Velkého rozšíření se dočkal formát PDF, stále častěji se objevují textové formáty odvozené z XML (např. RSS). Mění se také spektrum používaných koncových zařízení - stále více uživatelů používá k přístupu na web jinou techniku, než jsou počítače (mobily, PDA).

Při hledání vhodného výstupního formátu z IS pro integraci informačních zdrojů je vhodné posoudit následující otázky:

- **Je třeba zabránit změnám formátování z licenčních důvodů?** Tento požadavek bývá spojen s požadavkem na neměnnost obsahu a věrohodnost dokumentu. V takovém případě je vhodné použít kompaktní formáty s pevným stránkováním a rozložením, jako je např. PDF nebo obrazové formáty. Ty zároveň zajišťují, že je formátování a rozložení textu stejné při zobrazení i při tisku. Alternativou mohou být formáty podporující různé „zámky" zakazující změny v dokumentu (např. Microsoft Word a funkce Zámek), ty ale většinou neřeší problém změny velikosti zobrazovacího okna a následné přeformátování.
- **Jedná se o dokumenty převážně textové?** U textových dokumentů není problém v dynamickém přeformátování textu podle šířky zobrazovacího zařízení, proto pro tento případ dobře poslouží formáty průtokového typu, např. HTML nebo RTF.
- **Obsahují dokumentu velké množství grafiky nebo komplikované rozložení?** Obsahuje-li dokument velké množství tabulek, obtékání grafických objektů apod., může změna šířky zobrazovacího okna text zdeformovat a narušit rozložení s negativním vlivem na čitelnost, srozumitelnost i význam. Zde se nabízí více řešení. Kromě použití kompaktních formátů, jako je PDF, lze objekty zajistit proti změnám prostřednictvím CSS, definování absolutních nebo minimálních rozměrů objektů apod. Ale vzhledem k tomu, že integrační systém využívá velké množství různorodých informačních zdrojů s rozmanitým spektrem formátů, je vhodné zvolit jednoduchý způsob, který půjde aplikovat univerzálně bez nutnosti vytvářet adaptéry pro konkrétní případy.
- **Budou dokumenty dodávány i na mobilní zařízení?** Mobilní zařízení jako PDA nebo mobilní telefony s OS mají relativně malou zobrazovací plochu. Pro tyto zařízení je vhodné formátování textu upravit, eliminovat grafiku, případně grafické objekty zobrazovat pouze na vyžádání. Široce podporovaným formátem pro tato zařízení je kromě použití prostého textu a HTML formát Mobi PRC s možností zabezpečení, který podporují např. firmy Nokia, Sony Ericsson, Siemens a Motorola.

Pokud může integrační systém ovlivnit výstupní formát dokumentů, lze pro textové dokumenty, které není třeba chránit, obecně doporučit formát HTML. Pro dokumenty s komplikovaným

rozložením nebo pro dokumenty vyžadující ochranu je obecně nejvýhodnější formát PDF a pro mobilní zařízení formát Mobi PRC.

1.3.6 Formát PDF pro komplexní řízení ochrany dokumentu

K dispozici je dnes několik široce podporovaných dokumentových formátů, které podporují všechny výše zmíněné vlastnosti dokumentů a pro které jsou nabízeny ucelené sady aplikačních nástrojů pro editaci, indexaci a prohlížení dokumentů v těchto formátech. Jedná se např. o formát LIT od firmy Microsoft nebo formát Mobi PRC pro mobilní zařízení od firmy MobilPocket. Řada výrobců také používá různé modifikace otevřeného formátu OEB (Open E-book), který rozšiřují o vlastní metody šifrování a zabezpečení. Tyto technologie jsou poměrně dynamicky rozvíjeny díky vytvářejícímu se trhu s elektronickými knihami a časopisy, kde je zabezpečení obsahu dokumentu klíčovým předpokladem komerčně úspěšného prodeje titulů.

Nejflexibilnější a nejvíce podporovanou technologií pro zajištění ochrany autorských práv a vyhovění různým licenčním a bezpečnostním požadavkům je však bezesporu dokumentový formát PDF (Portable Document Format) od firmy Adobe. Formát PDF ve verzi 1.6 a technologie s ním spojené mají několik klíčových výhod, které z nich dělají favorita pro bezpečnou distribuci dokumentů v rámci integračního systému:

- formát PDF je hned po HTML nejpoužívanějším dokumentovým formátem na webu
- prohlížeč PDF je nezávislý na OS a instalovaných písmech
- prohlížeč PDF je bezplatný, bez požadavku na investice do SW u příjemce dokumentu
- PDF zachovává vzhled a integritu originálních dokumentů
- texty mohou být do PDF vloženy jako texty, křivky nebo bitmapy
- do PDF formátu je možno vkládat bitmapové obrázky i vektorové objekty různých typů
- metadata formátu PDF jsou uchovávána v XML
- PDF umožňuje dynamické přeformátování textu pro účely přenosných zařízení
- export do PDF je podporován pro většinu jednoduchých i profesionálních kancelářských a grafických editorů, včetně DTP a CAD systémů
- PDF podporuje plnotextové prohledávání, vkládání poznámek, záložek a datových polí
- PDF umožňuje vkládání jakýchkoli elektronických příloh, včetně multimediálních
- PDF umožňuje zabezpečení dokumentu proti neoprávněnému prohlížení, tisku, kopírování obsahu, včetně nastavení času expirace, po kterém již není možné dokument použít
- zabezpečení je postaveno na šifrování obsahu dokumentu a/nebo jeho příloh
- dokumenty v PDF lze digitálně podepsat a tím označit původce dokumentu
- dokumenty v PDF lze dodatečně opatřit vodoznaky, pozadím, záhlavím a zápatím
- PDF plní i archivní funkci se zárukou, že půjde dokument otevřít i po letech nezávisle na hardware a OS (formátová podmnožina PDF/A)
- PDF plní také funkci elektronické obálky pro bezpečnou distribuci objektů

Firma Adobe nabízí pro formát PDF ucelené aplikační řešení v podobě dokumentového serveru Adobe LiveCycle Document Server[1]. Tento systém zajišťuje tvorbu dokumentů (např. skenováním, převodem z jiných formátů, vkládáním údajů z databází nebo XML do šablon), správu a zabezpečení a řízenou distribuci dokumentů jak v elektronické, tak tištěné podobě (PDF, tisk, fax, e-mail, web, bezdrátový přenos). Dokumenty může vzájemně sdružovat, šifrovat, nastavovat úroveň přístupu a digitálně podepisovat. Jako modulární systém s otevřeným rozhraním ho lze bezešvě integrovat do ucelených integračních řešení. Představuje proto ideální stavební kámen při budování IS pro integraci informačních zdrojů.

PDF řešení lze použít jako univerzální výstup z integračního systému pro jakýkoli typ dokumentů i strukturovaných záznamů, protože do tohoto formátu lze snadno převádět jakýkoli tiskový výstup nebo formátovat data dynamicky z libovolných datových zdrojů. Díky možnosti zabezpečení PDF tak můžeme bezpečně distribuovat plné texty, referáty, záznamy i grafické objekty.

1.4 TECHNOLOGIE VZDÁLENÉHO PŘÍSTUPU

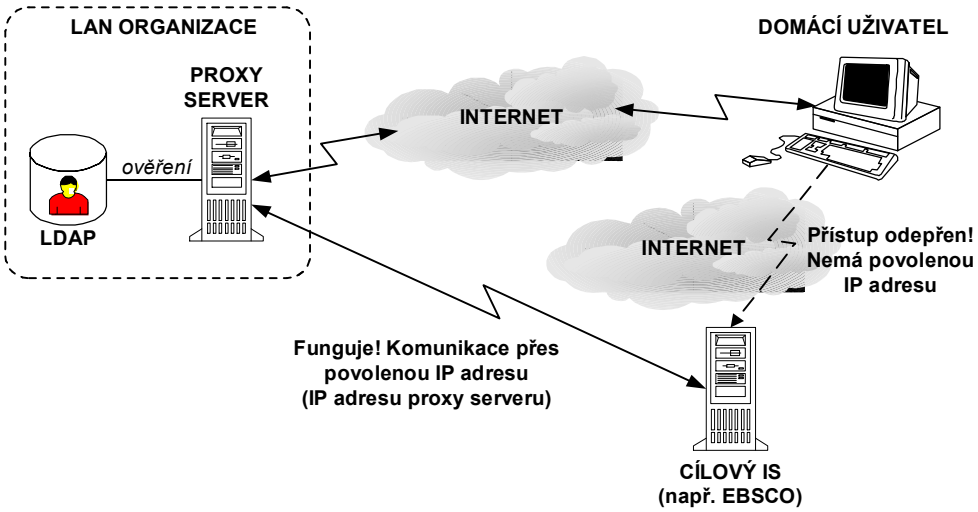
Provozovatelé často zpřístupňují své informační zdroje na základě licenčních ujednání s důrazem na dodržování autorských a vydavatelských práv. Při integraci informačních zdrojů je proto třeba uspokojivě vyřešit problém, jak těmto právům a licenčním podmínkám dostát, ale jak zároveň neznemožnit uživatelům s informačními zdroji **pracovat kdykoli a odkudkoli**.

Informační zdroje, zvláště pokud jsou komerční, se většinou snaží zamezit neoprávněným přístupům k obsahu a vyžadují různé druhy autentikace. Nejčastěji se v praxi setkáváme s autentikací na základě ověření uživatelského jména a hesla, s autentikací na základě IP adresy počítače nebo s kombinací obojího (více informací o autentikaci – viz kapitola **Chyba! Nenalezen zdroj odkazů.**). Uživatelská jména a hesla pro přístup k informačním zdrojům nesmějí instituce svým uživatelům sdělovat, IP adresy jsou vázány na rozsahy dané instituce. Výsledkem je, že uživatelé nemohou přistupovat k obsahu informačních zdrojů mimo prostory instituce, např. ze zaměstnání nebo z domova. Řešení tohoto problému nabízejí technologie vzdáleného přístupu.

Technologie vzdáleného přístupu musí být pro uživatele snadno použitelná a zároveň bezpečná. Uživatelé nejsou většinou ochotni nebo schopni na svá zařízení nic instalovat, navíc koncových zařízení existuje celá řada a jsou platformově velmi rozmanitá. Nabízí se však několik typů vzdáleného přístupu, které se v současné době díky velké poptávce po těchto technologiích rychle vyvíjejí, zdokonalují a vůči koncovému uživateli zjednodušují. Pro účely integrace informačních zdrojů jsou nejvyužívanější technologie vzdáleného přístupu: proxy server, VPN, terminálový přístup a Shibboleth.

1.4.1 Proxy server

Proxy server je počítač nebo specializovaný software, který umožňuje uživatelům nepřímé připojení k jinému serveru. V případě integrace informačních zdrojů hraje proxy server důležitou úlohu v tom, že funguje jako prostředník mezi uživatelem a cílovým informačním zdrojem, překládá jeho požadavky a vůči cílovému informačnímu zdroji vystupuje jako klient. Přijatou odpověď posílá následně zpět uživateli.



Obr. 2 – Vzdálený přístup k informačnímu zdroji přes proxy server

Vzdálený informační zdroj, ke kterému uživatel přistupuje prostřednictvím proxy, tak komunikuje pouze s proxy serverem a netuší, jaký uživatel ho stínově využívá – všechny počítače, které proxy server využívají, se jeví jako jeden počítač s IP adresou proxy serveru. Z hlediska autentikace pak stačí, aby byla všechna oprávnění na straně vzdáleného informačního zdroje nastavena pouze pro daný proxy server, např. na základě jeho IP adresy či loginu a

hesla. Proxy server však musí v takovém případě garantovat, že na něj budou přistupovat pouze oprávnění uživatelé, např. čtenáři konkrétní knihovny. Uživatelé se tak musí pro využití proxy serveru autentikovat, typicky uživatelským jménem a heslem, nejlépe šifrovaně. K tomu účelu se většinou využívá adresářových služeb přístupných přes LDAP, aby nemusela pro potřeby proxy vznikat samostatná databáze uživatelů.

Nasazením proxy serveru je v integračním systému dosaženo:

- **ochrany soukromí koncových uživatelů**
- **zvýšení výkonu komunikace** (pokud proxy server funguje zároveň jako cache a odpovědi si ukládá do vyrovnávací paměti)
- **zvýšení bezpečnosti** (uživatelé nemusí znát login a heslo ke vzdáleným zdrojům, navíc komunikace může být šifrována)
- **tvorby centralizovaných statistik přístupů** ke všem zdrojům
- **vzdálený přístup k informačním zdrojům odkudkoli**, protože již nehraje úlohu IP adresa uživatele

Funkci proxy serveru může plnit integrační systém již v principu na základě své architektury, např. pokud funguje jako paralelní vyhledávač, který požadavky uživatelů distribuuje na vzdálené servery a vystupuje tak jako prostředník v procesu vyhledávání.

1.4.2 VPN

VPN (Virtual Private Network) je technologie, která typicky slouží pro připojení vzdáleného počítače do lokální sítě organizace. I když se může počítač nacházet ve fyzicky nezávislé síti kdekoli na světě, může prostřednictvím VPN komunikovat se síťovými prostředky lokální sítě, jako by byl v lokální síti fyzicky přítomen. K navázání VPN spojení je třeba VPN server, který má přístup na internet i do lokální sítě organizace, a VPN klient, který se přes internet připojí k VPN serveru a prostřednictvím něho pak do lokální sítě. VPN server vystupuje jako síťová brána, která s klientem vytvoří zabezpečený síťový tunel, v rámci něhož pak probíhá požadovaná síťová komunikace. Podle konkrétně použité technologie jsou tak přenášená data většinou šifrována a zapouzdřena, často silnými šifrovacími algoritmy (např. DES, Triple DES nebo AES). Uživatelé pomocí VPN získají IP adresu instituce, která VPN provozuje, čímž získají přístup ke všem zdrojům, které jsou zprovozněny v dané lokální síti.

Klienti VPN jsou dnes již součástí všech operačních systémů, navíc existuje řada VPN klientů určených pro konkrétní síťové technologie, např. firmy Cisco. Připojení k VPN serveru je však nutné na straně klienta nakonfigurovat, což vyžaduje určitou technickou zdatnost. Proto se připojení VPN využívá zejm. na akademické půdě, méně již ve veřejných knihovnách.

1.4.3 Terminálový přístup

Terminálový přístup zajišťuje vzdálené přihlášení ke vzdálenému počítači, typicky serveru. Uživatel pomocí speciálního programu, který spouští na svém počítači, pracuje na vzdáleném počítači stejným způsobem, jako kdyby u něho fyzicky seděl – na vstup jsou přenášeny příkazy z klávesnice a myši, na výstupu se vracejí generované obrazovky. Terminálový přístup může být zajištěn pouze v textovém režimu (např. telnet) nebo v grafickém režimu (např. Vzdálená plocha v MS Windows). Velkou výhodou je, že vzdálený počítač může pracovat pod zcela jiným operačním systémem než počítač uživatele. Uživatel může využívat různé aplikace a zdroje, které jsou na vzdáleném počítači k dispozici – proto se terminálový přístup také označuje jako sdílení pracovní plochy (Desktop Sharing) nebo VNC (Virtual Network Computing). Komunikace při terminálovém přístupu bývá navíc šifrovaná, takže práce je bezpečná.

Nasazení technologie terminálového přístupu je vhodná u integračních projektů s malou uživatelskou základnou, protože je značně náročná na výkon terminálového serveru, zvláště pokud pracuje v grafickém režimu. Výhodou tohoto řešení je garance programového vybavení uživatele, protože uživatelské prostředí je realizováno na serveru.

1.4.4 Shibboleth

Technologie Shibboleth je představitelem tzv. **distribuované autentikace**. Zajišťuje přístup k zabezpečeným lokálním i vzdáleným informačním zdrojům na základě důvěryhodného ověření identity uživatele, navíc mezi zúčastněnými IS vytváří prostředí **SSO**. Technologie je navržena a dále rozvíjena konsorciem Internet2.

Prostředí Shibboleth vyžaduje spolupráci poskytovatelů obsahu (informačních zdrojů) a informačních systémů zajišťujících ověřování identity uživatelů. Jinými slovy, aby mohla být technologie Shibboleth nasazena, musí ji podporovat jak příslušné informační zdroje, tak instituce, které zajišťují ověřování identity svých uživatelů. Lze však také nasadit tzv. Shibboleth agregátory, které technologii Shibboleth podporují a které na základě IP adresy nebo loginu a hesla poskytují autentikaci pro ty informační zdroje, které technologii Shibboleth nepodporují.[5] Takovým agregátorem je např. produkt EZproxy od firmy Useful Utilities.

> Systém pro správu identity uživatelů (Identity Management System) slouží k poskytování služeb potřebných pro identifikaci osob působících v určitém systému a pro řízení přístupu identifikovaných uživatelů k těm komponentám systému, ke kterým je přístup nějakou formou omezený. Jedná se tedy o komplex služeb, který poskytuje na základě definovaných politik bezpečné a automatizované řešení správních úkonů efektivně řešících přístup uživatelů k omezeným a mnohdy citlivým informačním zdrojům.[3]

Protože je toto prostředí založeno na vzájemné důvěře, jsou zřizovány tzv. federace poskytovatelů obsahu a poskytovatelů identit. Základem technologie Shibboleth je služba Where Are You From (WAYF), která zajišťuje přesměrování autentikačních požadavků od poskytovatelů obsahu na příslušného poskytovatele ověření identity uživatele. Uživatel, který chce přistoupit k zabezpečenému informačnímu zdroji, je přesměrován na službu WAYF, kde si zvolí svého poskytovatele identity a u něho se autentikuje. Informační zdroj následně získá výsledek autentikace, případně minimální sadu atributů nutných pro autorizaci služby informačního zdroje. Poté může oprávněný uživatel začít se zdrojem pracovat.

Technologicky je Shibboleth založen na otevřených standardech SAML a LDAP, čímž je dána jeho snadná implementovatelnost ve stávajících prostředích IS. Softwarové nástroje Shibboleth jsou k dispozici jako open source pro všechny běžné operační systémy, čímž je zajištěn předpoklad postupného rozšíření mezi provozovatele informačních zdrojů. V současné době již Shibboleth podporují provozovatelé Ebsco, ProQuest, Elsevier, Blackwell, JSTOR, OCLC, Thomson Gale, CSA, Ovid Technologies a mnoho dalších.

1.4.5 Technologie SSO

Single sign-on (SSO) představuje skupinu speciálních technologií, které umožňují uživateli využívat **jeden autentikační kód** (běžně uživatelské jméno a heslo) pro přístup do více nezávislých IS. Typickým představitelem těchto technologií jsou centralizované adresářové služby typu LDAP nebo Active Directory, které slouží k ukládání autentikačních a autorizačních politik jednotlivých IS při společné databázi uživatelů. V kombinaci s distribuovanou autentikací může SSO umožnit uživateli **autentikovat se pouze jednou** a následně automaticky získávat přístup do více nezávislých IS. Tuto rozšířenou formu SSO reprezentují např. protokoly SAML či již zmíněný Shibboleth - systém SSO přijme autentikační token jiného IS, ověří ho a následně umožní uživateli využít požadovanou službu či informační zdroj, aniž by se musel uživatel znovu přihlašovat.

V kontextu integrace informačních zdrojů je SSO předpokladem integrace často různorodých autentikačních mechanizmů, kterými jsou informační zdroje zabezpečeny.

Kromě zvýšení komfortu práce koncových uživatelů, kteří si díky SSO nemusí pamatovat desítky přihlašovacích jmen a hesel, přinášejí tyto technologie mnoho výhod i pro samotný integrační systém. Jedná se zejména o **sjednocení autentikačních a autorizačních politik** v rámci celého integračního systému, zlepšení možností pro **provádění bezpečnostních auditů**,

provozování **jediného úložiště identit** uživatelů pro všechny komponenty integračního systému a **snížení zátěže na podporu** koncových uživatelů v otázkách zapomenutých hesel, nesprávně zapisovaných hesel či nepřístupných účtů.

K dispozici je řada řešení, které je možné bezplatně použít v integračních projektech. Kromě již zmíněných technologií LDAP, Shibboleth a SAML se jedná zejm. o populární open source řešení Central Authentication Service (CAS) vyvinuté na univerzitě v Yale, projekt CoSign podporovaný sdružením National Science Foundation Middleware Initiative (NMI) či široce oblíbená technologie Kerberos.

1.5 TECHNOLOGIE PRO VYTVÁŘENÍ OBSAHU INFORMAČNÍCH ZDROJŮ

Instituce, které provozují IS pro integraci informačních zdrojů, často sami vytvářejí informační objekty, které publikují v podobě elektronických dokumentů nebo informačních databází. Stávají se tak sami producenty, případně provozovateli dílčích informačních zdrojů, které mohou být objektem zájmu jejich vlastního nebo jiného integračního systému. Z toho důvodu by měly volit takové publikační technologie, které umožní snadnou a bezešvou integraci. Podívejme se v krátkosti na současné možnosti publikování elektronických informací.

1.5.1 Digitální repozitáře

Zamýšlí-li instituce budovat sbírku statických elektronických dokumentů, je vhodným řešením digitální repozitář. Digitální repozitář slouží jako **organizované úložiště digitálních objektů** a nemá žádné ambice řídit jejich vznik či následnou editaci. S objekty zachází jako s uzavřenými entitami, může k nim však připojovat popisná **metadata** vytvářená buď ručně nebo na základě extrakce údajů z příslušných dokumentových formátů (např. z XMP metadat formátu PDF nebo z EXIF metadat formátu JPEG). Součástí metadat by měl být vždy jednoznačný identifikátor objektu, volitelně pak různé formy věcného popisu, např. klasifikační schémata či abstrakta.

Digitální repozitář by měl v první řadě umět vysoce organizovaně ukládat a evidovat všechny objekty a plnotextově i strukturovaně v nich vyhledávat. Má tak mnoho společných rysů se systémy DMS (Document Management System), které se využívají ve firemní oblasti jako jeden z nástrojů podnikových intranetů. Digitální repozitář by měl být také vybaven rozhraním OAI-PMH pro sklízení metadat jinými informačními systémy, případně dalšími formami distribuce.

Tvorba a editace digitálních objektů je typicky realizována mimo digitální repozitář, často jako výsledek digitalizace klasických fondů nebo publikační činnosti dané komunity (např. zveřejňování závěrečných prací na vysokých školách).

1.5.2 Publikační systémy

Pokud chce instituce spojit fázi **tvorby** elektronických dokumentů s fází **publikování** v jednom systému, je vhodné použít systém pro správu obsahu. Tyto systémy se obecně označují zkratkou CMS (Content Management System) nebo v případě zúžení na webový obsah jako WCM (Web Content Management).

Součástí CMS bývá redakční modul, který umožňuje provádět rychlé změny obsahu dokumentů (většinou HTML dokumentů) bez znalosti programování a formátovacího jazyka. Zároveň často podporuje postupy redakčního schvalování, čímž umožňuje týmovou spolupráci více autorů při tvorbě obsahu. Vytvářený obsah je automaticky plnotextově indexován pro účely vyhledávání, zároveň je možné dokumenty opatřovat popisnými metadaty. Dokumenty jsou začleněny do přísně organizované struktury složek a podsložek, což přináší další možnosti v navigaci k obsahu. Přestože standardním rozhraním CMS bývá webové rozhraní, k dispozici bývají i další možnosti přístupu k obsahu, např. RSS kanály či rozhraní OAI-PMH. Další volitelnou součástí CMS mohou být nástroje umožňující interakci s uživateli, např. uživatelská fóra, ankety, možnost vkládat komentáře k článkům atd. CMS se pak přibližuje architektuře portálů.

Systémy CMS jsou vhodné k tvorbě dynamických webů, kde se předpokládá spolupráce týmu autorů, k publikování článků ve formě elektronických časopisů apod.

1.5.3 Databáze

K publikování **strukturovaných dat**, reprezentujících ať již primární či sekundární informace, jsou vhodným řešením aplikace využívající nějaký systém pro řízení bází dat (DBMS). Klasickým příkladem takových řešení jsou automatizované knihovní systémy, ve kterých se budují katalogy knihoven či jiné bibliografické databáze, nebo plnotextové databáze článků. Typicky tyto aplikace obsahují editační i prezentační část, takže může instituce v jednom systému informační objekty vytvářet i v nich vyhledávat. Architektura těchto aplikací většinou přísně odděluje aplikační a databázovou vrstvu, což má příznivý vliv na výkon, flexibilitu i bezpečnost.

Při výběru databázových systémů by měl být kladen důraz na interoperabilitu všech forem komunikace, směrem do systému i směrem do okolí. Tu typicky zajišťují standardizované jazyky vyšší úrovně pro manipulaci a definici dat, např. SQL nebo QBE. Klíčovými hledisky při výběru databázového systému by měla být robustnost a zotavitelnost při chybách bez ztráty dat. Moderní systémy DBMS proto často podporují správu jednotlivých datových transakcí a technologie zajišťující integritu dat.

Nejvyužívanějšími DBMS jsou v současné době z komerční oblasti technologie Oracle Database 10g, Sybase Adaptive Server Enterprise 15, Microsoft SQL Server 2005, IBM DB2 9 či IMB Informix 10. Jsou však k dispozici k vysoce kvalitní DBMS šířené nekomerčně, např. MySQL 5, PostgreSQL 8 nebo SQLite 3.

1.6 SHRNUTÍ

Při projektování IS můžeme vybírat z bohaté nabídky informačních technologií, které nám dnešní trh nabízí. Nejsou to pouze komerční technologie, čím dál častěji se objevují kvalitní technologie šířené zdarma jako GPL či freeware. Pokud se rozhodneme pro vývoj vlastních proprietárních řešení, měli bychom k tomu mít vážný důvod – ztrácíme tak totiž řadu výhod otevřených standardů a ověřených technologií. Použité technologie by měly být dostatečně otevřené a flexibilní, abychom je mohli v případě potřeby nahrazovat jinými. Ideální je nasazení modulární architektury složené ze samostatných komponent integrovaných mezi sebou do funkčního celku. Při výběru technologie musíme posuzovat celkové náklady na vlastnictví, nejen pořizovací cenu. Na celkových nákladech se totiž podílí i různé udržovací poplatky, mzdové náklady na odborníky a náklady na podporu, cena za hardware apod. po celou dobu životního cyklu.

V IS pro integraci informačních zdrojů hrají nejvýznamnější úlohu technologie pro vyhledávání a komunikaci se vzdálenými informačními zdroji, technologie na zajištění ochrany autorských práv a licencí, technologie vzdáleného přístupu a technologie pro vytváření obsahu. Protože informační technologie se rychle a dynamicky mění, je nutné stále sledovat trendy a vývojové směry, zejm. z hlediska standardizace.

LITERATURA

1. Amos software. *Adobe LiveCycle Policy Server* [online]. 2004 [cit. 2005-04-10]. Dostupný z WWW: <http://www.amsoft.cz/Produkty/Adobe/server/policy/overview.html>.
2. Digimarc. *MyPictureMarc : Communicate your copyrights with MyPictureMarc* [online]. 2004 [cit. 2005-04-10]. Dostupný z WWW: <http://www.digimarc.com/watermark/mypicturemarc/>.
3. HANÁČEK, Petr, STAUDEK, Jan. Správa identity. In HRUŠKA, Tomáš. *DATAKON 2005 : Brno, 22.-25. 10. 2005*. [s.l.] : [s.n.], 2005. s. 123-146. Dostupný z WWW: <www.fi.muni.cz/usr/staudek/vyuka/security/d05_idm_tutorial_text.pdf>. ISBN 80-210-3813-6.
4. KREJČÍ, Richard. *Elektronický podpis v PDF* [online]. Praha : Grafika Publishing, 2002 [cit. 2005-04-10]. Dostupný z WWW: <http://www.grafika.cz/art/pdf/pdfpodmis.html>.
5. PAVLÍK, Jiří. Shibboleth - elegantní technologie pro vzdálený přístup k dabatázím. In *INFORUM 2006: 12. konference o profesionálních informačních zdrojích : Praha, 23. - 25.5. 2006*. [s.l.] : [s.n.], 2006. s. 3. Dostupný z WWW: <http://www.inforum.cz/inforum2006/pdf/Pavlik_Jiri.pdf>.
6. *Z39.50 implementation experience*. Edited by Paul Over, William E. Moen. 1st edition. Washington : U.S. Government Printing Office, 1995. 123 s. s. NIST special publication 500-/ computer systems technology.
7. *Z39.50 Implementors Group* [online]. 2000 , 2002 [cit. 2007-03-29]. Dostupný z WWW: <http://www.stk.cz/ZIG/>.
8. Z39.50 MAINTENANCE AGENCY. *The Z39.50 Document* [online]. 2004 [cit. 2007-03-29]. Dostupný z WWW: <http://www.loc.gov/z3950/agency/document.html>.
9. *Z39.50 profil JIB* [online]. 2004 [cit. 2007-03-29]. Dostupný z WWW: <http://info.jib.cz/dokumenty/profiljib.pdf>.
10. ŽABIČKA, Petr. OAI-PMH: Protokol pro metadatovou interoperabilitu. In *Automatizace knihovnických procesů : Liberec 2003*. 1. vyd. Praha : ČVUT, 2003. s. 43-48. Dostupný z WWW: <knihovny.cvut.cz/akp2003/sbornik/05_zabicka.pdf>.