Příloha č. 4 k výzkumné zprávě projektu VE20072009004

Tomáš Hanousek

OAI-PMH pro začátečníky

1. OAI – Přehled

OAI: Open Archives Initiative (OAI)

Smyslem platformy otevřených archivů je zpřístupnění na Webu dostupných materiálů prostřednictvím vzájemného sdílení metadat mezi repozitáři, jejich publikování a archivace. OAI platforma vzešla z komunity institucí zabývajících se publikováním vědeckých prací v elektronické podobě (e-prints), kde narůstala potřeba nalézt snadné a efektivní řešení pro zpřístupnění různorodých elektronických úložišť. OAI vyvíjí a prosazuje snadno použitelný výměnný rámec a připojuje k němu standardy, původně pro účely zpřístupnění vědeckých prací, nyní však již počítá s výměnou dalších digitálních dokumentů. Programové prohlášní OAI zní „Open Archives Initiative vyvíjí a prosazuje výměnné standardy s cílem umožnit efektivní šíření obsahu."

Mnohé komunity začínají nebo mají v úmyslu využívat přístup k otevřeným archivům. Internet a rostoucí počet dokumentů v digitální podobě rozšiřuje počet dalších možných repozitářů. Dokumenty mohou být zpřístupněny a využívány stále širším okruhem zájemců, kteří mají i další záměry, než jen realizaci původní myšlenky tvorby repozitářů. Navíc možnost zpřístupnění více repozitářů umožňuje budování nových typů služeb, které mohou lépe sloužit potřebám uživatelů. Dalším podnětem vyplývajícím z této snahy je možnost vytvářet nové nenákladné modely vzájemné komunikace mezi odbornými institucemi.

OAI organizace zahrnuje výkonný orgán pro řízení a metodický a technický výbor pro určování směru vývoje protokolu. Federace digitálních knihoven (The Digital Library Federation - DLF), Spolek pro síťové sdílení informací (the Coalition for Networked Information - CNI) a Národní vědecká nadace (National Science Foundation - NSF) financují OAI. Zatímco exekutiva a finance jsou amerického původu, úspěch OAI je pevně zakotvený na účasti komunity lidí z celého světa, obzvlášť z Evropy a Severní Ameriky. Nyní, když je k dispozici dobře navržená a stabilní druhá verze OAI-PMH protokolu, potřeba zachování řízení v rukách malé skupiny osob pro účely samostatného a rychlého rozhodování, je již možná méně důležitá ve srovnání s představou stability a odborným vedením orgány pro dohled nad standardy, např. ISO. Tato možnost byla diskutována uvnitř OAI.

OAI Protocol for Metadata Harvesting (OAI-PMH)

OAI-PMH protokol pro sklízení metadat definuje mechanismus pro sklízení metadatových záznamů z repozitářů. OAI-PMH poskytuje jednoduchý technický prostředek poskytovatelům dat pro účely zpřístupnění svých metadat službám založeným na obecně rozšířených standardech HTTP (Hypertext Transport Protocol) a XML (Extensible Markup Language). Metadata určená k harvestování mohou být nabízena v jakémkoliv formátu, na kterém se dohodne určitá komunita nebo libovolná skupina poskytovatelů dat a služeb. Pro základní úroveň výměny metadat musí být podporován alespoň nekvalifikovaný Dublin Core. Takto mohou být metadata z mnoha zdrojů uložena v jediné databázi a poskytované služby mohou zpřístupňovat centrálně sklizená nebo agregovaná data. Vazbu mezi metadaty a odpovídajícím obsahem OAI protokol nedefinuje. Je nutné si uvědomit, že OAI-PMH neposkytuje možnost prohledávání nad těmito daty, pouze jednoduše umožňuje sběr metadat do jednoho místa. V případě poskytování služeb, sklízení metadat musí být doplněno dalšími prostředky.

Zdá se, že snaha OAI o prosazení použití tohoto protokolu vzbuzuje velké naděje. Podpora nových předpisů v odborné komunikaci je zřejmě největší zmiňovanou výhodou. Asi nejsnadněji dosažitelný cíl je odhalování „skrytých zdrojů dat" a jejich nenákladná výměna. Ačkoliv je OAI-PMH protokol technicky velmi jednoduchý, vybudování systému s logicky promyšlenými službami respektující požadavky uživatelů zůstává složité. OAI-PMH protokol by se mohl stát běžnou součástí WWW infrastruktury, s garantovanou podporou stejně jako nyní HTTP protokol. Prokazatelný úspěch prvních implementací s použitím kombinace těchto dvou relativně jednoduchých protokolů vede k všeobecnému rozšíření vědeckými a paměťovými institucemi, výzkumnými organizacemi a vydavateli.

---

7 základních definic

Open Archive Initiative (OAI)
OAI je iniciativa, která vyvíjí a prosazuje výměnné standardy s cílem umožnit efektivní zpřístupnění obsahu.

Archiv
Termín „archiv" ve smyslu OAI vystihuje původní význam komunity institucí zabývajících se publikováním vědeckých prací v elektronické podobě, kde termín archiv je obecně přijatý jako synonymum pro repozitář odborných dokumentů. Nechť představitelé archivní profese s omluvou zaznamenají tuto striktní definici archivu s jiným významem, než se používá v jejich oboru – instituce, úřad, který metodicky a odborně uchovává archiválie dlouhodobého významu a historické hodnoty. OAI používá termín archiv v širším významu slova: jako repozitář (datové úložiště) pro uložené informace. Jazyk a termíny nikdy jednoznačně nevystihují význam a OAI s respektem žádá o shovívavost komunitu odborných archivářů při použití termínu archiv.

OAI Protocol for Metadata Harvesting (OAI-PMH)
OAI-PMH je protokol pro sklízení metadat pro účely jejich sdílení mezi službami.

Protokol
Protokol je soubor pravidel, která definují způsob komunikace mezi systémy. Příkladem dalších rozšířených protokolů používaných ke komunikaci mezi systémy na Internetu jsou FTP (File Transfer Protocol) a HTTP (Hypertext Transport Protocol).

Sklízení (Harvesting)
V kontextu OAI, sklízení se týká specificky shromažďování metadat z různých repozitářů do centrálního datového úložiště.

Poskytovatel dat (Data Provider)
Poskytovatel dat zahrnuje jeden nebo více repozitářů (web serverů), které podporují OAI-PMH protokol pro účely poskytnutí metadat.

Poskytovatel služeb (Service Provider)
Poskytovatel služeb posílá požadavky ke sklízení poskytovatelům dat a využívá sklizená metadata k vybudování služeb s přidanou hodnotou.

2. Historie a vývoj OAI-PMH

Podstata OAI spočívá ve vývoji repozitářů pro ukládání vědeckých prací v elektronické podobě, tzv. archivů. Tyto repozitáře byly zřízeny pro účely vzájemné výměny výsledků vědeckého výzkumu dříve, než byly recenzovány a publikovány. První z nich byl *xxx* (později *arXiv*), který se původně od roku 1991 věnoval fyzice vysokých energií a rozrostl se do té míry, že pokryl celou oblast fyzikálních věd a příbuzné oblasti matematiky, nelineárních věd a výpočetní techniky. *CogPrints* se zaměřil na psychologii a jazykovědu. *Networked Computer Science Technical Reference Library* (zájmová technická knihovna výpočetní techniky - *NCSTRL*) zpřístupnila technické příspěvky týkající se výpočetní techniky stejně jako xxx nebo vlastní repozitáře spolupracujících vědeckých sborů. Podobně *RePEc* je k dispozici přispěvatelům z oblasti ekonomie s možností vložení svých pracovních verzí dokumentů do příslušných zájmových archivů, nebo pokud nejsou k dispozici, tak do *EconWPA* archivu Washingtonské university. K tomu *Networked Digital Library of Theses and Dissertations* (zájmová digitální knihovna diplomových a doktorantských prací - *NDLTD*) vybudovala digitální knihovnu diplomových a doktorantských prací v elektronické podobě rozvíjenou přímo studenty členských institucí.

Mechanismus naplňování těchto repozitářů spočíval ve všech případech ve vložení příspěvků samotnými autory. OAI pro potřeby tohoto tutoriálu definovala tzv. „e-prints" jako dokumenty archivované autory samými. Webová rozhraní umožnila lidem ovlivňovat tyto repozitáře, kteří mnohdy přispěli k nalezení řešení problémů. Pro různé repozitáře byla navržena různá rozhraní, takže uživatelé byli nuceni učit se ovládat různá uživatelská rozhraní v případě, že chtěli přistupovat k různým repozitářům a tam hledat řešení svých vědeckých problémů. Zatímco tzv. „Guildford protokol" podporuje vzájemnou součinnost mezi repozitáři RePEc archivů z oblasti ekonomie, NCSTRL repozitáře technických knihoven mají implementován tzv. „Dienst protokol". Tyto podskupiny protokolů umožňují implementaci široké palety služeb pro koncové uživatele včetně vyhledávání a listování napříč repozitáři sdružených do skupin. NDLTD digitální knihovna diplomových a doktorantských prací vytvořila metodiku pracovních postupů pro vkládání dokumentů a vyvinula XML DTD pro diplomové a doktorantské práce v elektronické podobě. Nicméně, některá sdílená metadata jsou podporována napříč tímto rozmanitým prostředím. Formují se stále další samostatné iniciativy v oblasti nových způsobů vědecké komunikace. Mnoho klíčových hráčů zabývajících se touto problematikou zjišťuje, že problém vzájemné součinnosti je čím dál tím více důležitý.

Santa Fe konvence

Byly stanoveny dva klíčové problémy vzájemné součinnosti tzv. e-prints archivů: za prvé koncoví uživatelé přišli do styku s různorodým vyhledávacím rozhraním, které komplikovalo výzkum, a za druhé neexistoval strojový způsob sdílení metadat. Zkoumané řešení spočívalo ve vyhledávání napříč elektronickými archivy a sklízení archivních metadat v případě poskytnutí služeb centrálního vyhledávání. V červenci 1999 Paul Ginsparg, Rick Luce a Herbert Van de Sompel z Národní laboratoře z Los Alamos sezvali skupinu technických odborníků zabývajících se publikováním vědeckých prací v elektronické podobě na listopadové setkání do Novomexického Santa Fe. Paul Ginsparg zabývající se arXiv a Herbert Van de Sompel působící v té době navíc na Univerzitě v Gentu, navrhli vytvoření univerzální služby pro archivaci autorské vědecké literatury (UPS - Universal Preprint Service). UPS by se tak stalo stěžejní a volně dostupnou vrstvou vědeckých informací, nad kterou by mohly vyrůst jak volné, tak komerční služby. Prvním krokem vedoucím k tomuto záměru byla identifikace nebo vytvoření

spolupracujících technologií a stanovení rámců pracovních postupů pro šíření vědeckých prací v elektronické podobě. Tyto záměry byly ohlášeny širšímu publiku pod hlavičkou „The Open Archives initiative aimed at the further promotion of author self-archived solutions – Iniciativa otevřených archivů OAI zaměřená na prosazení a podporu archivních řešení publikování vlastních autorských odborných prací".

Cílem setkání v Santa Fe bylo prodiskutovat problémy vzájemné součinnosti systémů, dohodnout se na rozpracování a prosazení modelu služby digitální knihovny založené na hlavních již existujících e-prints repozitářích (např. xxx/arXiv, *CogPrints*, NCSTRL, RePEc, NDLTD) a zároveň založit fórum pro další práci na vzájemné komunikaci mezi systémy elektronických archivů. Během přípravy na toto setkání byly již některé základní práce provedeny. Herbert Van de Sompel inicioval projekt simulující některá hlediska vzájemné součinnosti distribuovaných elektronických archivů. Thomas Krichel (Univerzita v Surrey, RePEc) experimentoval s konverzí dat z existujících iniciativ publikování vědeckých prací do metadatového formátu *ReDIF* používaného v systémech pro zpřístupnění prací z oblasti ekonomie RePEc. Michael Nelson (NASA Langley) použil data mj. z CogPrints, NASA, NCSTRL, RePEc a xxx k vytvoření různě strukturovaných archivů na základě svých návrhů tzv. chytrých objektů použitých na hloupé elektronické archivy (Smart Object Dumb Archives). Cílem této práce nebylo stanovení předpisů pro architekturu UPS, ale usnadnit debatu na toto téma během setkání.

Výzvy a navržená řešení

- Křížové prohledávání (nebo také distribuované hledání) nebo sklízení? -

Klíčovou úlohou v prvotní fázi byla volba hlavního směru vývoje architektury UPS systému. Byly možné dva přístupy – křížové prohledávání digitálních archivů založené na protokolu jako je Z39.50 nebo sklízení metadat do jednoho nebo více centrálních datových úložišť a poskytování dalších služeb předpokládající přesun velkého objemu dat a přizpůsobení jejich formátu pro účely zpřístupnění pomocí uživatelského rozhraní.

Zkušenosti digitálních archivů vypovídaly o tom, že křížové prohledávání nefunguje zcela dobře, alespoň pouze částečně, neboť vyhledávací služba je závislá na úrovni rychlosti a věrohodnosti prohledávaných serverů. Např. NCSTRL shledala, že distribuované prohledávání malého počtu uzlů je schůdné, ale výkon systému s více než 100 uzly byl již velmi pomalý. Resource Discovery Network (RDN) ve Velké Británii objevil, že služba dokonce jen s pěti zdroji pro křížové prohledávání byla málo výkonná a slabá v poskytnutých možnostech uživatelského rozhraní. Vývojáři proto hledali přijatelné řešení centralizované databáze. Čím více serverů je vzájemně prohledáváno, tím je větší šance, že některý ze serverů bude pomalý nebo nedůvěryhodný.

Dalším problémem je stanovení toho, které cílové servery zahrnout do systému vzájemného prohledávání. Popisy sbírek, kde jsou dostupné, nemusí být konzistentní napříč repozitáři, pokud nejsou navrženy pro elektronickou výměnu mezi systémy. Navíc určitý čas zabere koncovému uživateli zdlouhavé zkoumání chování WWW rozhraní. Rozdíly v syntaxi dotazovacích jazyků a variant vyhledávacích atributů (mezi servery, postupem času) představují překážky ve složitosti použití, buď pro koncového uživatele nebo pro prohledávací software, nebo oba. Další technický problém představuje uspořádané třídění výsledků z distribuovaných serverů, rozdílná velikost a typ dat uložených na cílových serverech mohou výsledky deformovat. Návrh uživatelského rozhraní je velice složitý v případě, že zobrazená metadata jsou distribuována napříč

velkým počtem repozitářů. Bylo navrženo, že řešením je dostat všechny metadatové záznamy společně do jednoho místa.

UPS model představený na setkání v Santa Fe demonstroval meziarchivní digitální knihovnu poskytující služby založené na souboru metadat sklizených z různých digitálních archivů. Jeho architektura byla navržena technickou knihovnou výpočetní techniky NCSTRL s použitím modifikované verze protokolu Dienst. Tímto způsobem počet prohledávaných uzlů byl redukován na jediný, poskytující významné výhody ve výkonnosti a funkčnosti. Poskytované služby mohou používat jediný dotazovací jazyk, soubor vyhledávacích atributů a třídící algoritmus. Navíc příprava formátu dat umožňuje vybudování jednoduššího způsobu procházení záznamy.

- Poskytovatelé dat (Data Providers) a služeb (Service Providers) -

UPS architektura stanovuje dvě logické role: „poskytovatele dat" a „poskytovatele služeb". Poskytovatelé dat ovládají datové skladiště a zpřístupňují zdrojové objekty uložené v repozitáři způsobem jejich vystavení pro účely sklízení metadat popisujících tyto zdrojové objekty z repozitáře. Existují tvůrci a provozovatelé metadat a repozitářů zdrojových objektů. Poskytovatelé služeb sklízejí metadata od poskytovatelů dat. Využívají sklizená metadata pro účely poskytnutí služeb s přidanou hodnotou nad těmito daty. Konkrétním typem nabízené služby může být např. vyhledávací nástroj nebo systém porovnání korespondujících recenzí. Je třeba si uvědomit, že jedna organizace poskytující data může mít obě role – nabízet data pro sklízení i poskytovat služby koncovým uživatelům. Klíčový posun v architektuře systému spočívá v přesunu od poskytnutí rozhraní jen koncovým uživatelům k systému podporujícímu obě rozhraní – rozhraní pro koncového uživatele i strojové rozhraní pro sklízení metadat.

Začátek protokolu

Jméno pro službu zpřístupnění dokumentů v digitální podobě UPS (Universal Preprint Service) bylo rychle změněno, částečně z důvodu vyvarování se možných potíží vyplývajících z faktu, že UPS je již mj. zavedená obchodní značka pro komerční službu doručování balíků, a částečně také proto, že se zjistilo, že ne všechny elektronické dokumenty existují v tištěné podobě. Pracovní rámec, na základě kterého tato univerzální služba byla vyvinuta, byl navržen OAI.

Z jednání a experimentů vyplynulo, že pro účely podpory sklízení metadat musí panovat shoda v těchto bodech:

- transportní protokol – např. HTTP nebo FTP
- metadatový formát – např. Dublin Core nebo MARC
- základ pro zajištění kvality metadat – povinný soubor elementů, jmenné a předmětné konvence apod.
- intelektuální vlastnictví a užívací práva – *co kdo* může dělat *s čím*

Výchozí shoda v klíčových bodech umožnila vyvinout protokol pro sklízení metadat nazvaný *Santa Fe Convention* na počest setkání, kde bylo této dohody dosaženo.

Historie verzí OAI-PMH protokolu

	Santa Fe konvence	OA-PMH v. 1.0/1.1	OAI-PMH v. 2.0
povaha	experimentální	experimentální	stabilní
příkazy	Dienst	OAI-PMH	OAI-PMH
požadavky	HTTP GET/POST	HTTP GET/POST	HTTP GET/POST
odpovědi	XML	XML	XML
přenos dat	HTTP	HTTP	HTTP
metadata	OAMS	nekvalifikovaný Dublin Core	nekvalifikovaný Dublin Core
zdrojové objekty	E-prints	objekty typu dokument	resources
model	sklízení metadat	sklízení metadat	sklízení metadat

Santa Fe konvence byla první etapa OAI-PMH protokolu pro sklízení metadat. Byla založena na modelu UPS, službě RePEc/SODA, model poskytovatele dat, protokolu Dienst a práci skupiny Santa Fe. Zaměření Santa Fe konvence bylo přizpůsobit zpřístupnění dokumentů v elektronické podobě.

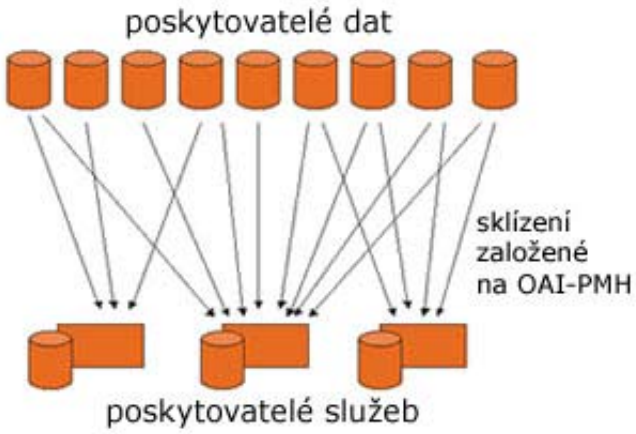
OAI-PMH protokol verze 1.0 zavedl jako základ pro výměnu metadat soubor elementů nekvalifikovaného Dublin Core. Protokol byl sestaven na základě Santa Fe konvence, setkání Svazu digitálních knihoven při Univerzitě Cornell a podpory alfa-testerů. Zaměření bylo rozšířeno o podporu výzkumu objektů typu dokument. Protokol byl založen na specifikaci snadné součinnosti založené na modelu sklízení metadat, rozšířeném a snadno využitelném transportním protokolu HTTP s využitím metod požadavků GET i POST a odpovědí formátovaných v XML. Je třeba si uvědomit, že se nejedná o protokol pro vyhledávání, protokol je spíše zaměřen na modelu sklízení metadat. OAI-PMH 1.1 byla opravená verze specifikace 1.0 s možností specifikace XML schématu. Obě verze byly v podstatě experimentální.

OAI-PMH verze 2.0 je hlavní revize protokolu a není kompatibilní s předchozími verzemi 1.x. Byla vystavena na OAI-PMH 1.x, podpoře skupiny implementátorů a techniků OAI a zpětné odezvě alfa-testerů. V tomto případě došlo opět k rozšíření protokolu, a to o metadata popisující zdrojové objekty. Jedná se stále o specifikaci založené na snadné součinnosti a modelu sklízení metadat. Verze 2.0 již není jen experimentální, ale jedná se o stabilní protokol. OAI se zavázala k tomu, že případné další revize protokolu budou již zpětně kompatibilní.

Pružné rozšíření

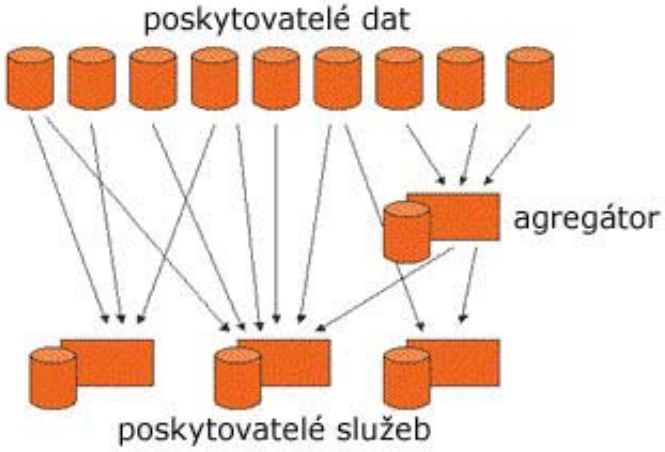
Vzhledem k tomu, že OAI-PMH je jednoduchý protokol založený na HTTP a XML, umožňuje rychlé rozšíření. K dispozici je velké množství softwarových pomůcek. Systém může být nasazen v různých variantách konfigurace tak, jak je prezentováno na následujícím schématu. Zdrojové metadatové a fulltextové objekty jsou typicky volně přístupné, není to ale podmínka. OAI-PMH může být také použit jen pro sdílení metadat v rámci uzavřené skupiny anebo v komerčních aplikacích.

Více poskytovatelů služeb



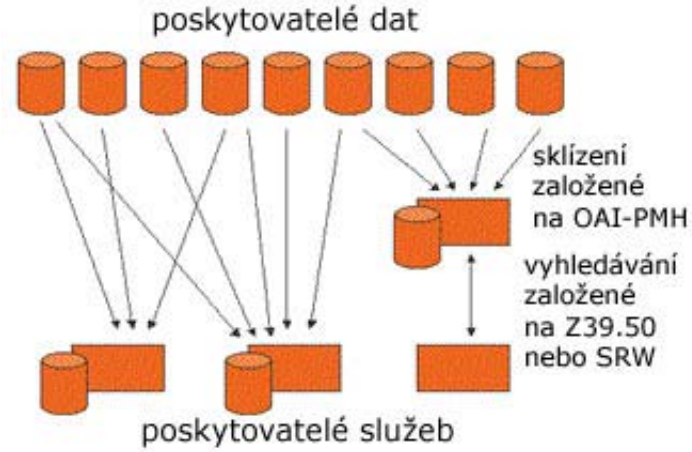
Více poskytovatelů služeb může sklízet metadata od více poskytovatelů dat.

Agregátoři



Agregátoři mohou sklízet metadata od více poskytovatelů dat (agregovat metadata) a poskytovat služby nad těmito daty i zároveň metadata nabízet ke sklízení.

Sklízení kombinované s vyhledáváním



Sklízení může být doplněno o vyhledávaní založené např. na Z39.50 nebo SRW.

Shrnutí

Původním návrhem bylo vytvoření samostatných řešení, objevila se však potřeba vzájemné součinnosti. Jako odpověď na tuto potřebu, setkání v Santa Fe vedlo k důsledné podpoře OAI iniciativy, která prosadila vzájemnou součinnost systémů prostřednictvím OAI-PMH protokolu sklízení metadat jako otevřeného standardu a zasadila se o rozšíření informací o OAI-PMH. OAI-PMH je mechanismus snadného sklízení metadatových záznamů z jednoho systému do druhého – od poskytovatelů dat k poskytovatelům služeb. Více poskytovatelů služeb může sklízet metadata od více poskytovatelů dat. Tím je zajištěno široké rozšíření metadat. OAI-PMH není protokol pro vyhledávání, ale slouží jako podpora pro služby založené na vyhledávacích mechanismech. Jedná se o základní vrstvu, na které je možné vybudovat další služby s přidanou hodnotou.

Vývoj za poslední tři roky vedl ke sdílení všeobecného popisu jakéhokoliv zdrojového objektu, již ne jen vědeckých prací v elektronické podobě. Ačkoliv je nekvalifikovaný Dublin Core stanoven jako základ pro vzájemnou výměnu metadat, OAI-PMH protokol může být rozšířen o libovolný další metadatový formát zapsaný ve formátu XML. Vzhledem k tomu, že se jedná o protokol založený na HTTP protokolu pro požadavky (a zároveň ovládání přístupů, kompresi, chybová hlášení apod.) a XML pro odpovědi, je implementace OAI-PMH přátelská vůči WWW použití a zároveň přátelská i ve vztahu k bezpečnostním firewall serverům. Umožňuje poskytovatelů služeb vyžádat si konkrétní metadatové záznamy založené na časovém údaji, příslušnosti k sadě nebo metadatovém formátu, nebo všechny záznamy. OAI může být snadno rozšířeno, protože se jedná o jednoduché řešení vybudované s využitím existujících technologií s podporou mnoha vývojových softwarových nástrojů.

Klíčové definice

E-print
Dokument publikovaný samotným autorem. V tomto významu je termín běžně používaný, obsahem tohoto dokumentu je výsledek vědeckého nebo jiného odborného výzkumu.

Objekt typu dokument (Document-like object)
Jedná se o datový celek v digitální podobě, který je srovnatelný s papírovým dokumentem. Termín označuje relativně jednoduchý zdrojový objekt, který by neměl obsahovat např. multimediální části nebo interaktivní služby.

Resource
Resource je zdrojový objekt, který může být čímkoliv, co lze identifikovat – např. elektronický dokument, obrázek, služba (např. dnešní předpověď počasí) nebo soubor dalších zdrojových objektů. Ne všechny zdrojové objekty jsou přenositelné po síti, např. člověk, společnost, svázané knihy v knihovně, které také mohou být považovány za zdrojové objekty.

XML
XML je zkratka pro Extensible Markup Language. XML je rozšiřitelný značkovací jazyk pro tvorbu dalších jazyků. Definuje význam popisovaných dat. XML dokument může být

validován oproti DTD nebo schématu definujícím elementy vytvořeného jazyka. Pro velký počet metadatových formátů sloužících k zápisu metadatových záznamů existují XML mapování.

DTD
DTD je zkratka pro Document Type Definition – definice dokumentu. DTD je formální specifikace struktury dokumentu.

Dublin Core
Dublin Core (DC) je metadatový formát definovaný za základě mezinárodní dohody. Dublin Core metadatový soubor elementů definuje 15 základních elementů pro účely jednoduchého popisu zdrojových objektů. Použití všech 15 elementů je doporučeno, žádný z nich však není povinný. DC formát může být rozšířen o další volitelné elementy, kvalifikátory a slovníkové výrazy.

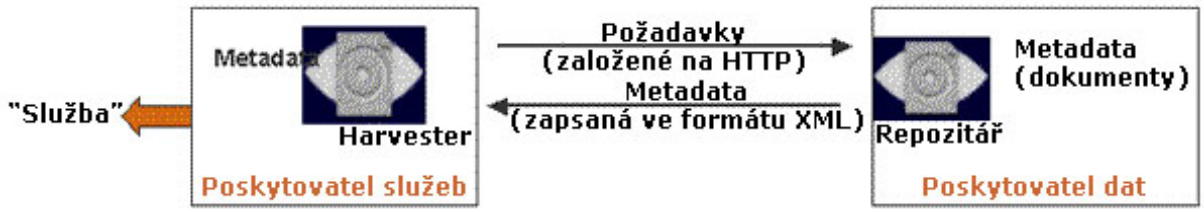
Součinnost (Interoperability)
Součinnost je schopnost systémů, služeb a organizací pracovat společně nad problémy nebo úkoly vedoucí ke společným nebo rozdílným cílům. V technické oblasti je součinnost podporována otevřenými standardy vzájemné komunikace mezi systémy a standardy pro popis zdrojových objektů. Součinnost je zde míněna hlavně ve významu zpřístupnění zdrojových objektů.

3. Hlavní technické myšlenky OAI-PMH

Shrnutí hlavních myšlenek OAI

- celosvětové spojení odborných archivů
- volný přístup do archivů (alespoň na úrovni zpřístupnění metadat)
- logicky navržená uživatelská rozhraní digitálních archivů a poskytovatelů služeb
- jednoduchá implementace s použitím rozšířených protokolů založených např. na HTTP, XML, DC

Základní funkčnost OAI-PMH



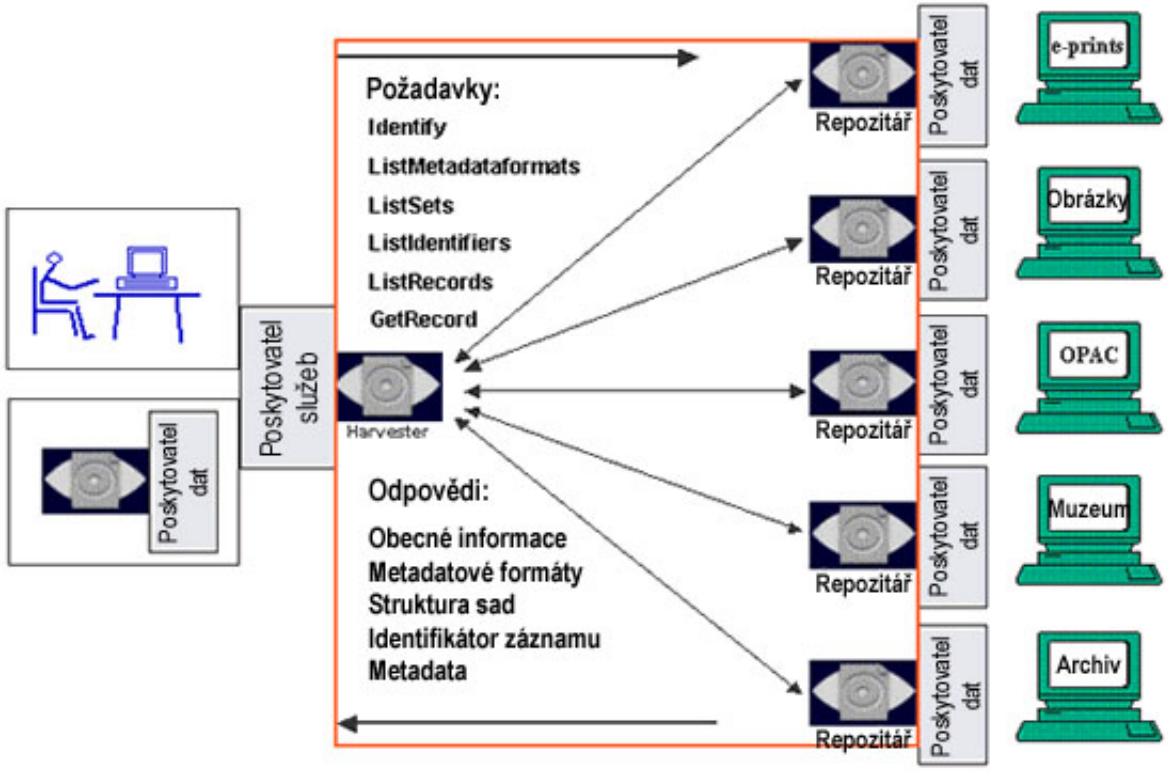
OAI: obecný přehled

Jsou definovány dvě skupiny „účastníků": poskytovatel dat a poskytovat služeb.

Poskytovatelé dat (otevřené archivy, datová úložiště) umožňují volný přístup k metadatům a mohou, ne nezbytně nutně, zpřístupňovat celé texty a další zdrojové objekty (např. obrázky). OAI-PMH umožňuje jednoduchou implementaci řešení poskytovatelů dat.

Poskytovatelé služeb využívají OAI rozhraní ke sklízení metadat poskytovatelů dat a jejich ukládání. Ke sklízení se nepoužívá on-line vyhledávací dotazy adresované poskytovatelům dat, ale služby založené na sklízení dat prostřednictvím OAI-PMH. Poskytovatelé služeb mohou vybrat určité sady záznamů od poskytovatelů dat (např. podle hierarchie nebo časové značky). Poskytovatelé služeb nabízejí služby založené na sklízení metadat a mohou je rozšířit o další služby s přidanou hodnotou (vyhledávání, zpřístupnění zdrojových dat – textů, obrázků, zvukových nahrávek, videozáznamů apod.).

OAI-PMH: celkový pohled na strukturu modelu



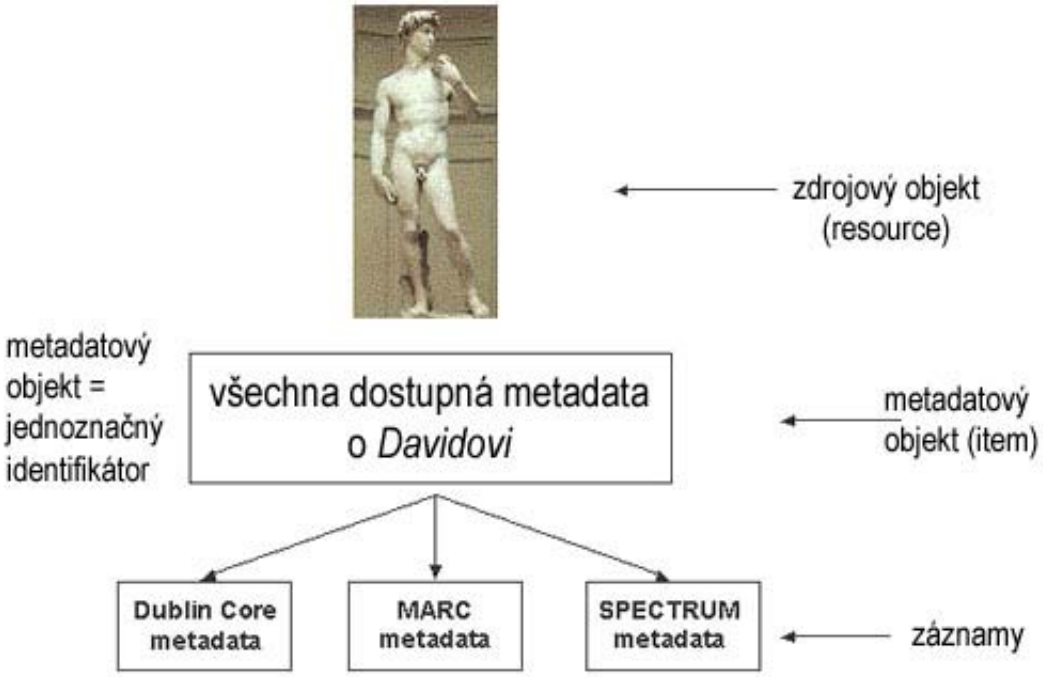
OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) protokol je založen na HTTP protokolu. Argumenty dotazu jsou předávány jako parametry metod **GET** nebo **POST**. OAI-PMH podporuje šest typů příkazů, např. příkaz `http://archive.org?verb=ListRecords&from=2002-11-01` vrátí seznam záznamů (sklidí záznamy z repozitáře) s časovou značkou od 1.11.2002.

Odpovědi jsou zapsány ve formátu XML. OAI-PMH podporuje jakákoliv metadata formátovaná v XML syntaxi. Základním formátem pro součinnost je při nejmenším Dublin Core.

Chybová hlášení jsou založena na HTTP.

Poskytovatelé dat mohou mít definovanou logickou strukturu skupin záznamů pro sklízení metadat v různých úrovních podrobnosti poskytovateli služeb. Časová značka označuje poslední změnu souboru metadat. Spolu s definicí skupin záznamů slouží časová značka poskytovateli služeb k volbě vhodného stupně podrobnosti pro sklízení.

OAI-PMH podporuje řízený tok přenosu dat.



Základní definice

Harvester:
klientská aplikace provozovaná poskytovatelem služeb, která generuje OAI-PMH požadavky za účelem sklízení metadatových záznamů z repozitářů

Repository:
síťově přístupný server (repozitář) provozovaný poskytovatelem dat schopný provést OAI-PMH požadavky – server poskytující metadatové záznamy prostřednictvím OAI-PMH protokolu

Resource:
zdrojový objekt (např. obrázek) popsaný metadaty, jejich vlastnosti nejsou v OAI-PMH definovány, mohou být nebo nemusí digitální

Item:
metadatový objekt uložený v repozitáři, který je tvořen metadatovými záznamy, má jednoznačný identifikátor

Record:
metadatový záznam v nějakém konkrétním formátu (např. DC, MARC) odvozený z metadatového objektu

Identifier:
jednoznačný identifikátor metadatového objektu uloženého v repozitáři

Set:
nepovinná volitelná sada (skupina) definovaná za účelem seskupení metadatových objektů společného charakteru v repozitáři.

Podrobný popis protokolu

Record:
Metadatový záznam v konkrétním formátu popisující zdrojový objekt. Záznam má 3 části: povinnou hlavičku, povinná metadata a volitelné další údaje o zdroji. Každá část obsahuje níže uvedené parametry.

hlavička (povinná)
identifier - identifikátor (povinný, jen 1 výskyt)
datestamp - časová značka (povinná, jen 1 výskyt)
setSpec elementy – údaje o příslušnosti záznamu k sadám (volitelné: žádný, jeden nebo více výskytů)
volitelný stavový atribut pro smazané metadatové objekty

metadata (povinná)
zapsaná ve formátu XML s uvedenou kořenovou značkou, repozitáře jmenného prostoru musí podporovat alespoň Dublin Core, mohou podporovat další formáty

údaje o zdroji (volitelné)
práva
původ

Datestamp - časová značka:
Časová značka označuje datum poslední změny metadatového záznamu. Jedná se o povinný parametr každého metadatového objektu. Rozlišujeme dvě možné úrovně granularity:
RRRR-MM-DD nebo RRRR-MM-DDThh:mm:ssZ, kde „T" odděluje zápis datumu a času a „Z" označuje tzv. Zulu Time resp. UTC celosvětový univerzální čas.

Časová značka umožňuje výběrové sklízení metadat s použitím argumentů **from** (od) a **until** (do). OAI-PMH aplikace jsou založené na mechanismu přírustkové aktualizace, umožňují sklízení metadat na základě časového údaje (datum nebo datum a čas) vzniku, poslední modifikace nebo smazání záznamu.

Schéma metadat:
OAI-PMH podporuje vytěžení metadat z repozitáře v různých formátech. Parametry metadatových formátů jsou:
- id řetězec specifikující formát metadat (**metadataPrefix**)
- URL schématu metadata pro účely příp. validace
- URI XML jmenného prostoru (globální identifikátor metadatového formátu)

Repozitáře musí umožňovat vytěžování metadat alespoň ve formátu Dublin Core. Prostřednictvím OAI-PMH mohou být definovány a přenášeny další metadatové formáty. Všechna poskytnutá metadata musí odpovídat příslušné specifikaci jmenného prostoru. Základem metadatového formátu Dublin Core je sada 15 elementů. Všechny tyto elementy jsou volitelné a mohou se opakovat.

Sada elementů metadatového formátu Dublin Core:

Title (Název)	Contributor (Přispěvatel)	Source (Zdroj)
Creator (Autor)	Date (Datum)	Language (Jazy k)
Subject (Předmět a klíčová slova)	Type (Typ zdroje)	Relation (Vztah)
Description (Popis)	Format (Formát)	Coverage (Pokrytí)
Publisher (Vydavatel)	Identifier (Identifikátor zdroje)	Rights (Správa autorských práv)

Sady záznamů:
Sady umožňují logické členění repozitářů. Tato vlastnost je volitelná. Archivy nemusí mít definovány žádné sady. Nejsou dány žádné doporučení pro implementaci sad. Sady nemusí nutně vyjadřovat obsah repozitáře. Nejsou striktně hierarchické. Je důležité a nezbytné definovat sady záznamů na základě dohod v rámci komunit, ve kterých se předpokládá realizace sklízení metadat.

- funkce: výběrové sklízení (parametr **`set`**)
- aplikace: tématicky zaměřené brány (portály), nástroje pro vyhledávaní disertačních prací, a další
- příklady definice sad na základě:
    - typů publikací (diplomové práce, odborné články apod.)
    - typů dokumentů (texty, zvukové nahrávky, videozáznamy, obrázky apod.)
    - obsahu (historie, zdravotnictví, biologie apod.)

Formát požadavku:
Požadavky musí být odesílány pomocí http protokolu metodou **`GET`** nebo **`POST`**, repozitáře musí poporovat obě tyto metody. Každý požadavek se skládá z URL repozitáře a argumentů zapsaných ve tvaru klíč=hodnota oddělených znakem „&". Jedním z těchto klíčů musí být tzv. sloveso verb=typ_požadavku, kde typ_požadavku je příkaz, např. **`ListRecords`** pro výpis záznamů. Použití dalších dvojic klíč=hodnota závisí na zvoleném příkazu.

Příklad odeslání požadavku metodou **`GET`**:
`http://archive.org/oai?verb=ListRecords&metadataPrefix=oai_dc`
Repozitář s URL http://archive.org/oai by měl vrátit metadatové XML zaznamy zapsané v metadatovém formátu Dublin Core.

Obecně repozitář na základě takto formulovaných požadavků vrací XML dokumenty v kódování UTF-8 nebo http chybová hlášení. Vrácený XML dokument lze validovat pomocí XML schématu, které je součástí OAI-PMH protokolu.

Protokol musí umožňovat zápis speciálních znaků: např. „**:**" (oddělovač čísla portu serveru) může být zapsán "**%3A**".

Poznámka:
Metoda GET kóduje požadavek do URL, server zpracovává argumenty parsováním proměnné WWW serveru QUERY_STRING. Metoda POST je vhodná pro přenos požadavku většího objemu dat libovolného typu (text, binární data apod.). Server pak zpracovává parametry požadavku ze standardního vstupu proudu dat stdin. Tato metoda je vhodná i v případě použití parametrů příkazu zapsaných s použitím diakritiky a speciálních znaků.

Odpověď:
Odpovědi jsou formátovány pomocí HTTP protokolu. Obsah musí být typu text/xml, tedy text zapsaný ve formátu XML. Chybová hlášení jsou založena na HTTP chybových stavových kódech a odvozena od OAI-PMH chyb, např. mohou být vráceny kódy chyby 302 (přesměrováno) nebo 503 (služba není dostupná). OAI-PMH umožňuje volitelně přenos dat v komprimovaném tvaru. Identifikace kompresní metody je povinná. Odpověď musí být správně formátovaný XML dokument s uvedením následujících značek:

1. deklarace XML
   (**<?xml version="1.0" encoding="UTF-8" ?>**)
2. kořenový element pojmenovaný **OAI-PMH** s třemi atributy
   (**xmlns**, **xmlns:xsi**, **xsi:schemaLocation**)
3. tři podřízené elementy (potomky)
    1. **responseDate** (datum a čas odpovědi v UTC – mezinárodním univerzálním čase)
    2. **request** (požadavek, na základě kterého se generuje tato odpověď)
    3. a) **error** (chyba, jestliže nastane chyba nebo výjimka)
       b) odpověď na OAI-PMH příkaz

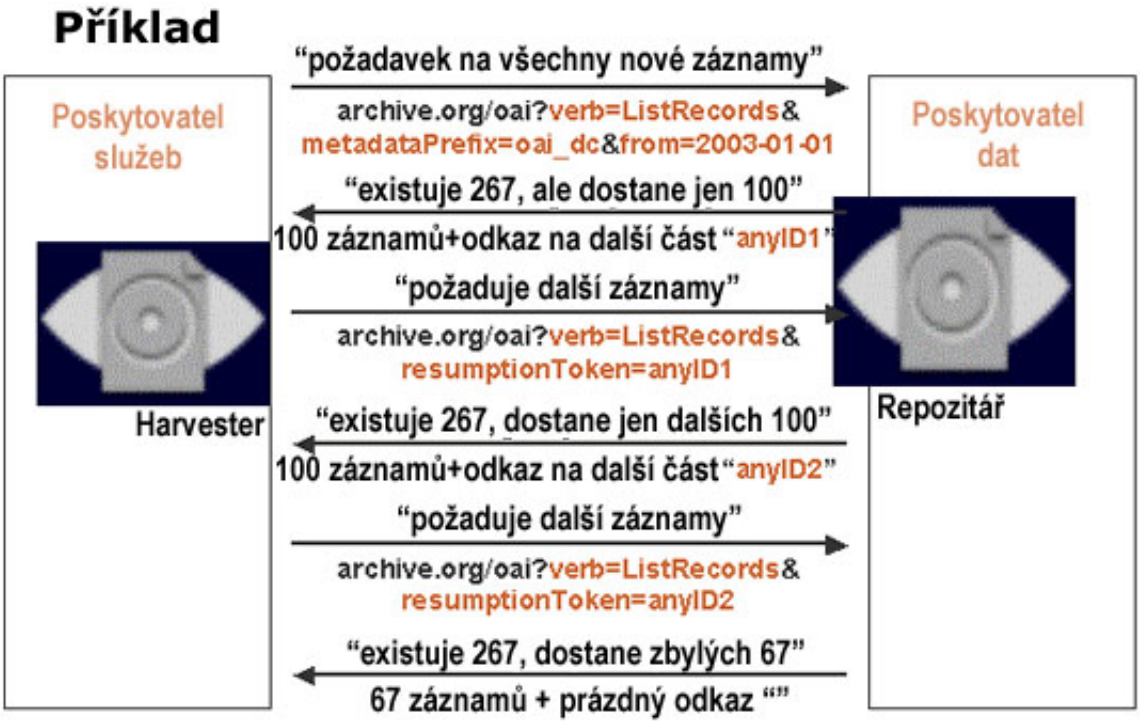
Řízení toku dat:
Čtyři typy požadavků vrací seznam údajů (mj. metadatové záznamy). Tři z nich mohou vyvolat odpověď „seznam velkého objemu".

OAI-PMH podporuje dělení (dávkování). O tom, zda bude odpověď rozdělena na části a jakým způsobem, rozhoduje repozitář – poskytovatel dat.

Odpověď na požadavek obsahuje:
nekompletní seznam
pokračovací značka – odkaz na další část seznamu (klíč resumptionToken)
+ volitelně datum vypršení platnosti odpovědi,
velikost kompletního seznamu,
ukazatel

Požadavek na pokračování ve sklizení další části seznamu stejného příkazu:
pokračovací značka (odkaz na další část seznamu) jako parametr
všechny další parametry jsou vynechány

Odpověď obsahuje odkaz na další část seznamu (což může být poslední) - pokračovací značku (parametr resumptionToken). Pokud je poskytnuta poslední část seznamu, hodnotou pokračovací značky je prázdný řetězec.



Chyby a obsluha výjimek:
Repozitáře musí označit OAI-PMH chyby vložením jednoho nebo více elementů **`error`**.
Definované chybové identifikátory jsou tyto:

```
BadArgument – neplatný argument
badResumptionToken – neplatný ukazatel na další část seznamu
badVerb – neplatný příkaz
cannotDisseminateFormat – formát nelze rozšířit
idDoesNotExist – identifikátor neexistuje
noRecordsMatch – požadavku neodpovídá žádný záznam
noMetaDataFormats – neznámý metadatový formát
noSetHierarchy – sada není definována
```

Typu požadavků

Je definováno šest různých typů požadavků:

```
Identify
ListMetadataFormats
ListSets
ListIdentifiers
ListRecords
GetRecord
```

Poskytovatel dat není povinný používat všechny typy příkazů, avšak repozitář musí mít implementovány všechny. V závislosti na typu příkazu jsou dány povinné a volitelné argumenty. Každý typ příkazu je níže podrobněji popsán.

- Identify -

funkce
popis (představení) archivu

příklad
archive.org/oai-script?**verb=Identify**

parametry
žádné

chyby / výjimky
`badArgument` – neplatný argument
(např.. archive.org/oai-script?verb=Identify**&set=biology**)

formát odpovědi

Element	*Příklad*	*Výskyt*
repositoryName (jméno repozitáře)	My Archive	1
baseURL (základní URL)	http://archive.org/oai	1
protocolVersion (verze protokolu)	2.0	1
earliestDatestamp (prvotní časová značka – datum spuštění)	1999-01-01	1
deleteRecords (možnosti mazání záznamů)	no (záznamy nejsou mazány), transient (záznamy mohou být dočasně smazány), persistent (záznamy jsou smazány trvale)	1
granularity (úroveň granularity)	RRRR-MM-DD, RRRR-MM-DDThh:mm:ssZ	1
adminEmail (Emailová adresa administrátora systému)	oai-admin@archive.org	+
compression (podpora komprese dat)	deflate, compress	*
description (popis)	oai-identifier, eprints, friends, …	*

Výskyt: 1 = povinný, pouze 1 výskyt; + = povinný, více výskytů; * = volitelný, 0 nebo více výskytů

- **ListMetadataFormats** -

funkce
vrací seznam dostupných metadatových formátů archivu

příklad
archive.org/oai-script?**verb=ListMetadataFormats**&
identifier=oai:HUBerlin.de:3000218

parametry
`identifier` (identifikátor, nepovinný parametr)

chyby / výjimky
`badArgument` – neplatný argument
`idDoesNotExist` – identifikátor neexistuje

e.g. archive.org/oai-script?verb=ListMetadataFormats
&identifier=really-wrong-identifier

noMetadataFormats

- ListSets -

funkce
vrací seznam dostupných sad v repozitáři

příklad
archive.org/oai-script?**verb=ListSets**

parametry
resumptionToken - odkaz na další část seznamu (vyhrazený parametr)

chyby / výjimky
badArgument – neplatný argument
badResumptionToken – neplatný odkaz na další část seznamu
e.g. archive.org/oai-script?verb=ListSets
&resumptionToken=any-wrong-token
noSetHierarchy – sady nejsou v repozitáři definovány

- ListIdentifiers -

funkce
zkrácený zápis **ListRecords**, vrací jen hlavičky záznamů

příklad
archive.org/oai-script?**verb=ListIdentifiers**&
metadataPrefix=oai_dc&**from=2002-12-01**

parametry
from - od data (volitelný parametr)
until - do data (volitelný parametr)
metadataPrefix - prefix metadatového formátu (povinný parametr)
set – sada (volitelný parametr)
resumptionToken - odkaz na další část seznamu (vyhrazený parametr)

chyby / výjimky
badArgument - chybný argument (např. ?**&from=2002-12-01-13:45:00** – datum a čas má být oddělen znakem „T“)
badResumptionToken – neplatný odkaz na další část seznamu
cannotDisseminateFormat – metadata v požadovaném formátu nejsou dostupná
noRecordsMatch – požadavku neodpovídá žádný záznam
noSetHierarchy – sada není definována

- ListRecords -

funkce
sklízení záznamů z repozitáře

příklad
archive.org/oai-script?**verb=ListRecords**&
metadataPrefix=oai_dc&**set=biology**

parametry
`from` - od data (volitelný parametr)
`until` - do data (volitelný parametr)
`metadataPrefix` - prefix metadatového formátu (povinný parametr)
`set` **–** sada (volitelný parametr)
`resumptionToken` - odkaz na další část seznamu (vyhrazený parametr)

chyby / výjimky
`badArgument` - chybný argument
`badResumptionToken` – neplatný odkaz na další část seznamu
`cannotDisseminateFormat` – záznamy v požadovaném formátu nejsou dostupná
`noRecordsMatch` – požadavku neodpovídá žádný záznam
`noSetHierarchy` – sada není definována

- GetRecord -

funkce
vrací konkrétní metadatový záznam z repozitáře

příklad
archive.org/oai-script?**verb=GetRecord**&
identifier=oai:HUBerlin.de:3000218&
metadataPrefix=oai_dc

parametry
`identifier` – identifikátor (povinný)
`metadataPrefix` - prefix metadatového formátu (povinný)

chyby / výjimky
`badArgument` - chybný argument
`cannotDisseminateFormat` – záznam v požadovaném formátu není dostupný
`idDoesNotExist` – záznam s tímto identifikátorem neexistuje

Příklad č. 1: odpověď na požadavek *ListIdentifiers*

Tento příklad demonstruje odpověď repozitáře (poskytovatele dat) na příkaz *ListIdentifiers* s uvedením časového rozsahu, metadatového formátu, názvu sady a poskytovatele dat.

Example: http://edoc.hu-berlin.de/OAI-2.0?
verb=ListIdentifiers&from=2002-01-06&until=2002-01-08&
metadataPrefix=oai_dc&set=doctypes:dissertations

```
<?xml version="1.0" encoding="UTF-8"?>
<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/"
         xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
         xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
                             http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2002-10-22T17:49:49+01:00</responseDate>
  <request verb="ListIdentifiers" from="2002-01-03" until="2002-01-08" metadataPrefix="oai_dc"
           set="doctypes:dissertations">http://edoc.hu-berlin.de/OAI-2.0</request>
  <ListIdentifiers>
    <header>
      <identifier>oai:HUBerlin.de:3000819</identifier>
      <datestamp>2002-01-08</datestamp>
      <setSpec>doctypes</setSpec>
      <setSpec>doctypes:dissertations</setSpec>
      <setSpec>dnb</setSpec>
      <setSpec>dnb:dnb33</setSpec>
    </header>
    <header>
      <identifier>oai:HUBerlin.de:3000831</identifier>
      <datestamp>2002-01-07</datestamp>
      <setSpec>doctypes</setSpec>
      <setSpec>doctypes:dissertations</setSpec>
      <setSpec>dnb</setSpec>
      <setSpec>dnb:dnb27</setSpec>
    </header>
  </ListIdentifiers>
</OAI-PMH>
```

Příklad č. 2: odpověď na požadavek *GetRecord*

Tento příklad demonstruje odpověď repozitáře na příkaz *GetRecord* – požadavek vrácení konkrétního záznamu specifikovaného identifikátorem.

Example: http://edoc.hu-berlin.de/OAI-2.0?
verb=GetRecord&identifier=oai:HUBerlin:3000819&
metadataPrefix=oai_dc

```
<?xml version="1.0" encoding="UTF-8"?>
 <OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/"
          xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
          xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
                              http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
   <responseDate>2002-11-27T14:57:01+01:00</responseDate>
   <request verb="GetRecord" metadataPrefix="oai_dc"
            identifier="oai:HUBerlin.de:3000819">http://edoc.hu-berlin.de/OAI-2.0</request>
   <GetRecord>
     <record>
       <header>
         <identifier>oai:HUBerlin.de:3000819</identifier>
         [...]
       </header>
       <metadata>
         <oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
                    xmlns:dc="http://purl.org/dc/elements/1.1/"
                    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
                    xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
                                        http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
           <dc:title>EinfluÃŸ genetischer Variationen im Tumor Nekrose [...]</dc:title>
           <dc:creator>SchÃ¼tzÃ¶ffel, Antje</dc:creator>
           [...]
       </metadata>
     </record>
   </GetRecord>
 </OAI-PMH>
```