Metavyhledávání a jeho typy – příspěvek k terminologické diskusi

PhDr. Helena Kučerová
Vyšší odborná škola informačních služeb, Praha 4

Postprint původní studie:

KUČEROVÁ, Helena. Metavyhledávání a jeho typy – příspěvek k terminologické diskusi. *Knihovna plus* [online]. 2011, č. 2 [cit. 2011-08-31]. Dostupné z: http://knihovna.nkp.cz/knihovnaplus112/kucer.htm. ISSN 1801-5948

Resumé:

V článku jsou představeny tři modely souběžného vyhledávání z více zdrojů: model sjednocující uživatelské rozhraní, model s jednotným vyhledávacím softwarem a model založený na jednotném indexovém souboru. Cílem je poskytnout věcný podkladový materiál pro návrhy české terminologie tohoto typu vyhledávání.

Klíčová slova: *modely vyhledávání informací – metavyhledávání – federativní vyhledávání – web scale discovery.*

Summary:

The article introduces three models describing parallel searching of multiple resources: a model unifying user interface, a model with centralized search engine, and a model based on a preharvested central index. The aim is to provide factual basis for proposals of Czech terminology in this type of information retrieval.

Keywords: *information retrieval models – metasearch – federated search – web scale discovery.*

Úvod

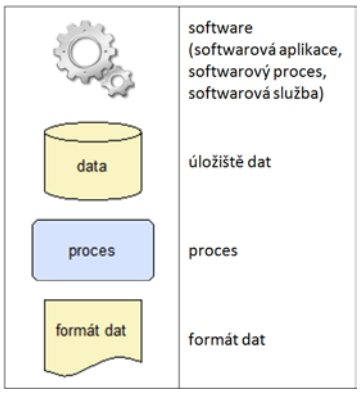
S nárůstem počtu zdrojů informací dostupných online[1] začali uživatelé vnímat jako problém rozdílná uživatelská rozhraní, autentizační procedury, dotazovací jazyky, prezentační formáty. Model druhé poloviny 20. století, kdy pro ně vyhledávání v odborných online zdrojích zajišťovali profesionální rešeršéři, nahradil všudypřítomný Internet, v jehož prostředí si každý může hledat informace sám. Vpád laiků do oblasti donedávna vyhrazené specialistům způsobil dramatickou změnu požadavků na způsob vyhledávání v online zdrojích – uživatelé vyžadují snadný přístup a simultánní prohledávání všech zdrojů z jednoho místa tak, jak si na něj navykli při vyhledávání v Internetu prostřednictvím Google. Softwarové aplikace a počítačové sítě jim již usnadnily publikování dokumentů, nyní od nich uživatelé totéž očekávají při vyhledávání informací. Požadované sjednocení způsobu vyhledávání z heterogenních zdrojů má však zároveň zachovat jejich samostatnost a umožnit jim další nezávislý vývoj podporující jejich specifika. Řečeno slovy Susan Feldmanové, současné vyhledávací technologie by měly „indexovat všechno a přitom to ponechat na svém místě, uspokojit potřebu získat vše z jednoho místa a přitom umožnit expertům, aby pokračovali v používání svých specializovaných nástrojů“[3]. To „všechno“ už zdaleka nejsou jen sbírky knihoven a archivů, ale i bibliografické databáze, autoritní báze, tezaury a další pomůcky k vyhledávání, e-knihy, elektronické časopisy, emailové archivy, dokumenty vygenerované pomocí kancelářských balíků (korespondence, zprávy, výkazy, tabulky), záznamy v transakčních databázích a datových skladech podnikových informačních systémů či studijních informačních systémů a četné další zdroje. Co do typů mohou zahrnovat jak text, tak i obrázky, audio, video, numerická data, primární zdroje i metadata v různých formátech, informace „zdarma“ (volně přístupné, veřejné) i informace za poplatek (licencované

zdroje), zdroje interní (podnikový intranet nebo institucionální repozitář) i externí, místní (lokální) i vzdálené zdroje (např. v cloudu, dostupné prostřednictvím webových služeb), soubory uložené na síťových serverech, na lokálních discích (PC, tablet, PDA) nebo na přenosných paměťových médiích (USB flash), zdroje volně dostupné na Internetu neboli povrchový web i tzv. neviditelný web.

Překotný vývoj v této oblasti doprovází řada nově se vynořivších pojmenování, se kterými přicházejí nejen vědci a odborníci z praxe, ale i marketingoví specialisté. Ta nejpoužívanější zde bez nároku na úplnost uvádíme v abecedním pořadí v jazyce jejich vzniku, tj. v angličtině, spolu s pracovními překlady do češtiny: *broadcast search* (plošné vyhledávání rozesíláním dotazu na více míst), *brokering* (zprostředkované vyhledávání), *centralized search* (centralizované vyhledávání), *cross search* (křížové vyhledávání), *distributed search* (distribuované vyhledávání), *enterprise search* (vyhledávání v podnikových informačních systémech a zdrojích), *federated search* (federativní vyhledávání), *harvested search* (vyhledávání založené na plošné archivaci distribuovaných zdrojů), *integrated search* (integrované vyhledávání), *metasearch* (metavyhledávání), *multisearch* (vyhledávání ve více zdrojích), *one-step / one-stop search* (jednorázové vyhledávání), *parallel search* (souběžné vyhledávání), *polysearch* (vyhledávání ve více zdrojích), *search portal / gateway* (vyhledávací brána), *unified search* (sjednocené vyhledávání), *web scale discovery* (o něm blíže v části „Varianta C – sjednocení indexu a centrální vyhledávání“). O tom, že vyhledávání z více informačních zdrojů současně už se stalo realitou i v prostředí českých knihoven, svědčí i nedávné zařazení termínu federativní vyhledávání do České terminologické databáze knihovnictví a informační vědy (TDKIV), a mimo jiné i diskuse o vhodném českém ekvivalentu pro nástroje typu „*discovery service*“ v elektronické konferenci Terminologie (květen 2011) a hlasování o něm v anketě elektronického časopisu Ikaros (červen 2011). Jakkoli si v tomto článku zatím netroufáme formulovat vlastní návrhy české terminologie pro tuto oblast, rádi bychom nabídli věcný podkladový materiál pro budoucí terminologické diskuse. Pokusíme se o zobecněný pohled na technologie v pozadí souběžného vyhledávání z více zdrojů, který by ukázal na profilující se typy tohoto způsobu vyhledávání a zkusil vystihnout rozdíly mezi nimi. Vycházíme z přesvědčení, že kvalitní terminologie má být postavena na dobře vymezených pojmech, jež mají reprezentovat naše pochopení podstaty toho, jak věci fungují.

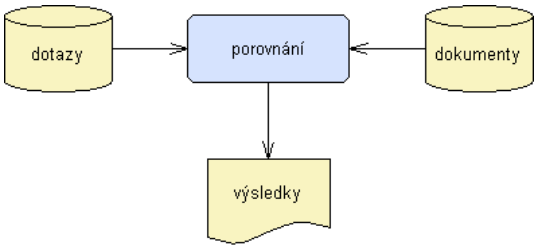
Obecné schéma fungování rešeršního systému

V konstrukci obecného schématu rešeršního systému vycházíme především z [2, 6], v jednotlivých případech jsme využili i přístupy z [1, 4, 8, 9, 11]. Východiskem pro nás bude procesní model vyhledávání informací (obr. 2 a obr. 3), jehož prvky využijeme v dalších schématech na obr. 4-8. Tento model se soustředí na zachycení klíčových procesů a jejich vzájemných vztahů. Během procesů jsou transformovány, vyhledávány a organizovány datové objekty. Plné šipky znázorňují následnost procesů, šipky s přerušovanou čarou představují tok dat z a do jejich úložišť. V diagramech jsou použity následující symboly (viz obr. 1):



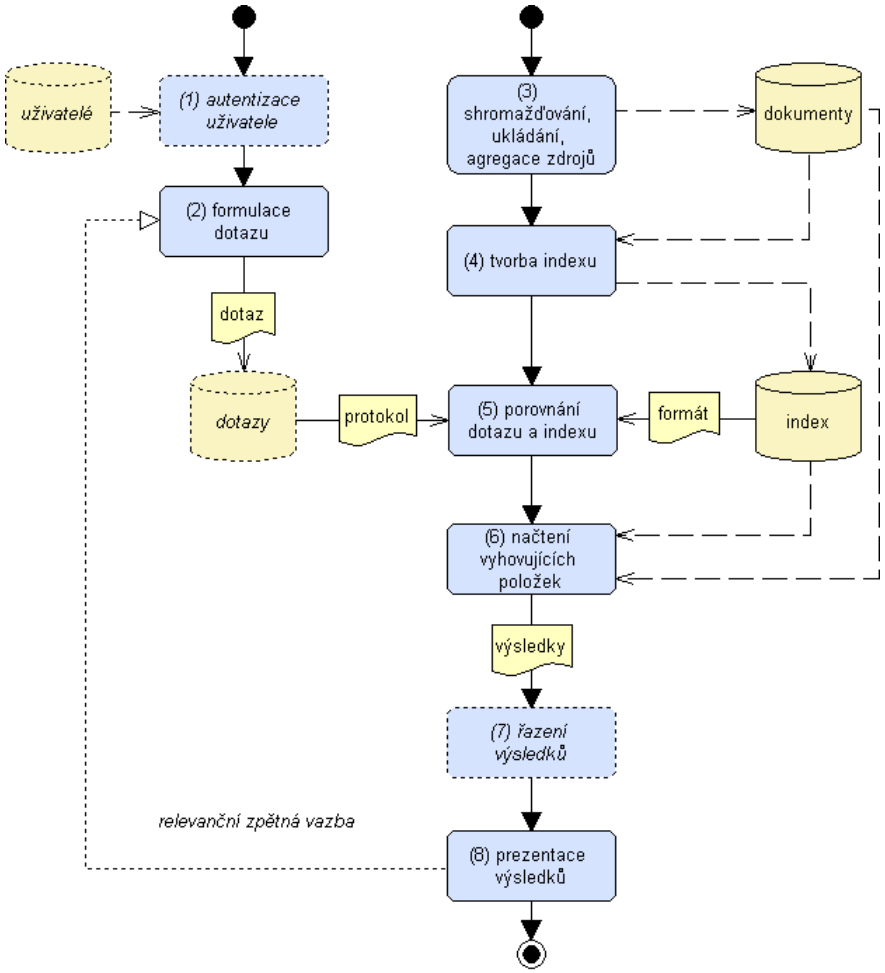
Obr. 1. Použité grafické symboly

Vyjdeme z obecného modelu na obrázku 2, který ukazuje, že technologické pozadí všech rešeršních systémů je v zásadě velmi jednoduché – jeden zpracovatelský proces se dvěma vstupy a s jedním výstupem. Obsah dotazů (a jejich prostřednictvím informační potřeby uživatele) se porovnává s obsahem prohledávaných zdrojů; množina, jež odpovídá (je relevantní vzhledem k) dotazu, je uživateli prezentována coby výsledek vyhledávání. Porovnávací proceduru v současné době realizuje ve většině případů počítačový program nad množinou jazykových výrazů, které pomocí přirozeného nebo umělého jazyka reprezentují obsah dotazů a dokumentů.



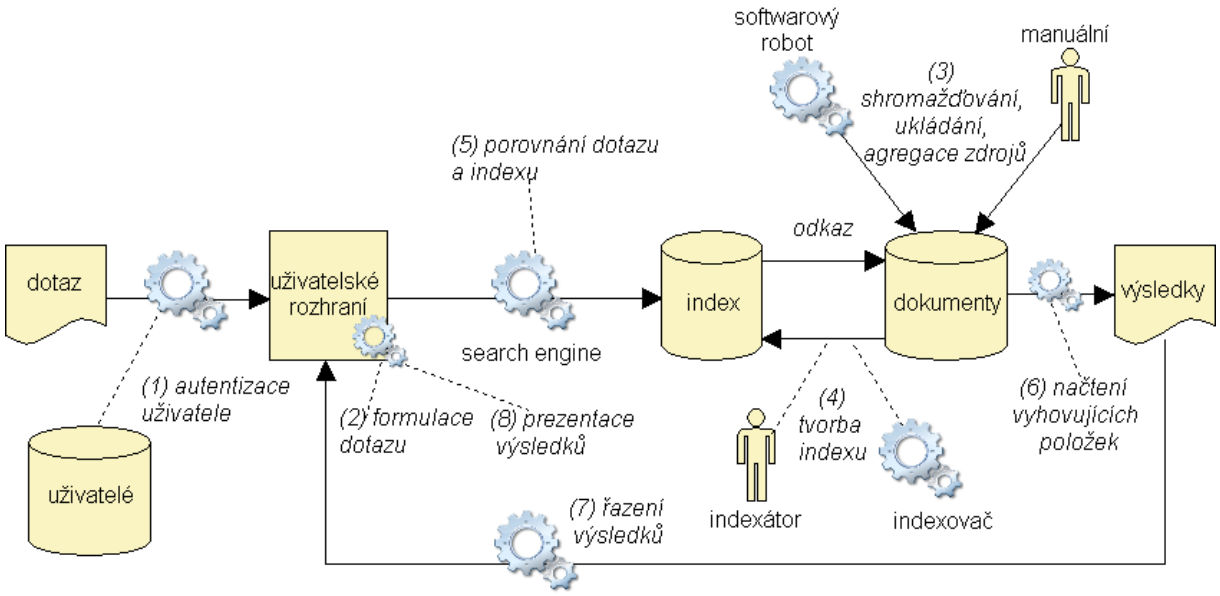
Obr. 2. Obecný model vyhledávání informací

Pro pochopení technologie vyhledávání bude vhodné tento jednoduchý rámec podrobněji analyzovat a doplnit jej o další prvky. Diagram na obrázku 3 znázorňuje detaily obou vstupních toků: první tok představuje zadání dotazu k vyhledávání, druhý tok zajišťuje vytvoření prohledatelné kolekce údajů. Tok směřující k zadání dotazu začíná procesem ověření oprávnění uživatele pracovat s příslušným zdrojem, k čemuž je potřebné mít k dispozici soubor dat o uživatelích. Po provedení autentizace (1) následuje formulace rešeršního dotazu (2), který je jazykovým vyjádřením informační potřeby uživatele. Dotaz je pomocí domluveného komunikačního protokolu předán v jazyce a formátu srozumitelném vyhledávacímu softwaru k vyhodnocení. Druhá větev diagramu začíná shromažďováním primárních informačních zdrojů (3) do úložiště, jež může mít různé názvy – např. archiv, databáze, knihovna, repozitář; v našem diagramu je obecně označeno jako „dokumenty“. Následuje proces tvorby indexového souboru (4), jehož podstatou je vytvoření množiny metadat odkazujících do základního souboru dokumentů. Kvalita tohoto procesu klíčovým způsobem ovlivňuje úspěšnost vyhledávání, protože, jak je zřejmé z diagramu, vyhledávací program porovnává obsah dotazu nikoli se samotnými dokumenty, ale právě s obsahem indexového souboru. Aby bylo možné srovnat obsah dotazu a indexu (5), je třeba zajistit kompatibilitu jejich datových formátů. Po této operaci následuje načtení potenciálně relevantních položek (6) z indexu a případně i ze souboru dokumentů, na něž indexový soubor odkazuje. Ty se seřadí (7) a ve formátu použitelném pro uživatele se prezentují (8). Obvyklým postupem je tzv. iterativní vyhledávání, kdy se na základě vyhodnocení obsahu výsledků vyhledávání pokračuje modifikací původního dotazu, což se označuje jako relevanční zpětná vazba.



Obr. 3. Obecný model vyhledávání informací – podrobnější pohled

Ne všechny procesy a komponenty znázorněné na obrázku 3 musí být přítomné v každém případu vyhledávání. Na nepovinné části v našem diagramu upozorňuje jejich ohraničení přerušovanou čárou. Autentizace uživatelů a s ní související databáze údajů o nich není např. nutná při vyhledávání pomocí webových vyhledávačů. U autentizace navíc ještě různé systémy řeší různě její zařazení do sledu aktivit – například v některých systémech je uživateli umožněno volné dotazování, jsou mu poskytnuta metadata a autentizace je vyžadována až ve chvíli, kdy by chtěl získat přístup k plným textům vyhledaných dokumentů. Další nepovinnou komponentou je úložiště dotazů – některé systémy dotazy používají pouze v procesu aktuálního vyhledávání, poté je odstraňují z paměti; v současné době však přibývají aplikace, v nichž se dotazy archivují a jsou pak kupříkladu nabízeny uživatelům jako doplněk či alternativa k jejich formulacím dotazu („Měli jste na mysli…“), případně jsou využity při personalizaci a při aktivních nabídkách („Uživatelé, kteří hledali to, co vy, hledali také…“). Poslední nepovinnou komponentou je řazení výsledků. Teoreticky je sice možné uživateli předat výsledek vyhledávání v surové podobě tak, jak byly výsledky načteny ze zdroje, v praxi však dnes každý profesionální vyhledávací systém musí být vybaven mechanismem minimálně pro formální řazení (např. podle abecedy), přičemž z uživatelského pohledu je samozřejmě nejžádanější řazení podle věcné relevance.



Obr. 4. Vyhledávání z jednoho zdroje

Schéma na obrázku 4 znázorňuje všechny podstatné komponenty rešeršního systému a jejich vazbu na procesní model z obrázku 3: data o uživatelích, uživatelské rozhraní, index a primární dokumenty. Doplněni jsou aktéři, kteří zodpovídají za realizaci příslušných procesů, tj. lidé a software.

Pro uživatele bezesporu nejdůležitější prvek vyhledávacího systému představuje uživatelské rozhraní. V našem schématu ho vidíme jako komponentu, která hraje roli jednak na vstupu do vyhledávacího procesu tím, že umožňuje formulaci uživatelova dotazu (2), zformátuje ho v souladu s vyhledávacím protokolem a předá k vyhodnocení vyhledávacím softwarem, jednak na výstupu, kdy jsou jejím prostřednictvím interpretovány a prezentovány výsledky vyhledávání uživateli (8). Tyto aktivity jsou samozřejmě rovněž podporovány specializovanými softwarovými aplikacemi.

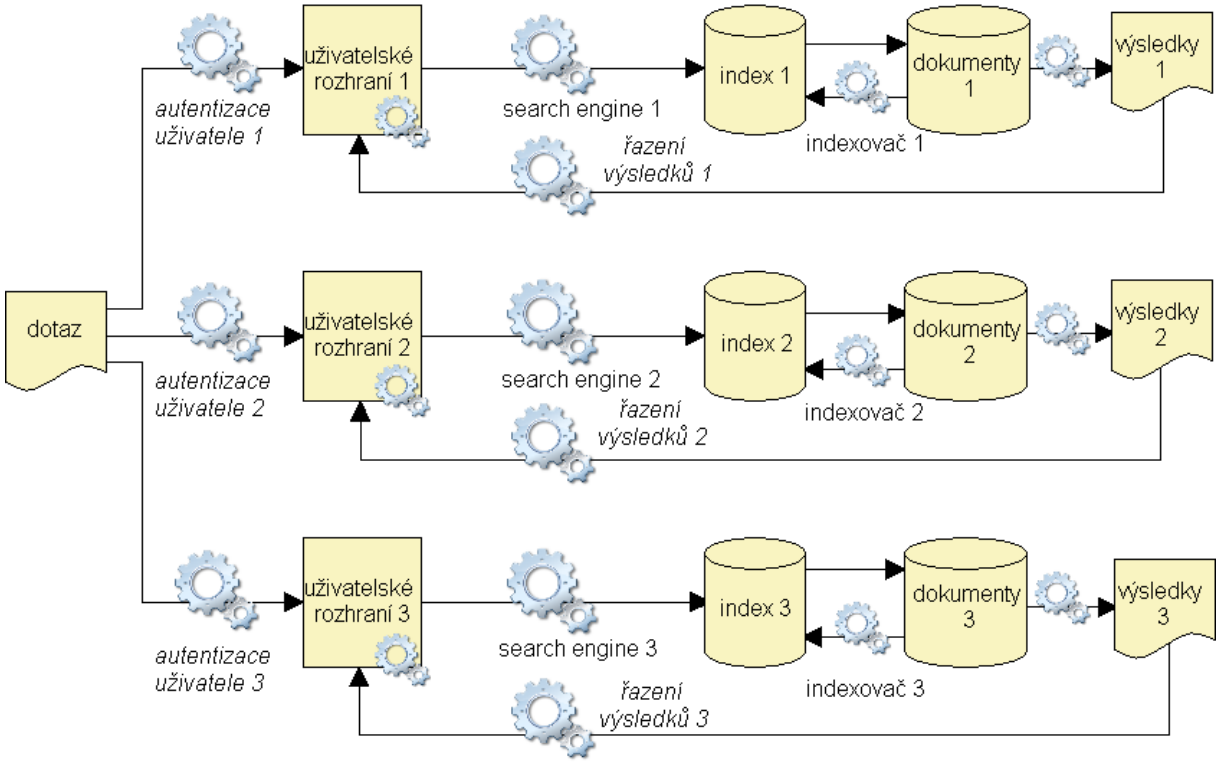
Další významnou komponentou je indexový soubor a dva typy aktérů, kteří ho mohou vytvořit (4) – indexovač a indexátor. Indexovač je softwarová aplikace, indexátor je osoba. Indexátorem může být profesionál (např. katalogizátor), ale i laik, který k videu, jež umístí na YouTube, přidá několik tagů. Na tomto místě je vhodné připomenout, že existují různé typy indexů. Za krajní póly je možné považovat tzv. strukturovaný nebo též metadatový index (např. katalog knihovny) a index plnotextový. Strukturovaný index současně s údajem o dokumentu obsahuje informaci, jaké strukturní části dokumentu se údaj týká (např. jméno autora). Jeho obsah mohou tvořit i údaje, které se přímo v dokumentu nevyskytují (např. třídníky MDT). Může být vytvořen jak intelektuálně indexátorem, tak automaticky pomocí indexovače. Plnotextový index je generován výhradně automaticky z textu dokumentu. V praxi je možné se setkat i s různými hybridními přístupy, kdy například plnotextový index generovaný automaticky ze slov dokumentu je kombinován s termíny z intelektuálně zpracovaného tezauru, nebo je obohacen prostředky zpracování přirozeného jazyka. Tato heterogenita indexových souborů představuje spolu s heterogenitou primárních zdrojů významný problém k řešení.

Shromáždění primárních dokumentů (3) je možné rovněž realizovat jednak manuálně (např. prostřednictvím akvizitéra v knihovně nebo i samotných autorů, jak je to obvyklé v současných aplikacích Webu 2.0), jednak automaticky pomocí softwarového robota zvaného *crawler, spider* nebo *harvester*.

Klíčovým činitelem procesu vyhledávání informací je software. Ten zajišťuje dva základní typy operací: transformaci a členění [2, s. 3]. V případě transformace jde především o transformace komunikovaných zpráv a o reformátování dotazů a záznamů v souladu s používanými jazyky a protokoly. Operace členění spočívá jednak v rozdělení prohledávaného souboru na části relevantní a nerelevantní – vyhovující a nevyhovující dotazu (vlastní vyhledávací algoritmus) a jednak v řazení množiny výsledků. Jak je zřejmé z diagramu, úlohy vyhledávacího softwaru jsou vykonávány na různých místech procesu: nejdůležitější je úloha porovnání dotazu a indexu (5), kterou realizuje tzv. *search engine* (vyhledávací stroj). Následuje načtení vyhovujících položek (6) a řazení výsledků (7). Všechny zde uvedené úlohy mohou být vykonány jednou komplexní aplikací[2], nebo je může provádět ve vzájemné interoperabilitě více různých programů.

Pro zjednodušení budeme v následujících schématech odhlížet od procesu shromažďování zdrojů a v procesu tvorby indexu budeme uvažovat pouze automatické generování, nikoli manuální tvorbu indexu. Nebudeme se zabývat dopodrobna ani technickou realizací autentizace uživatele[3].

Díky existenci počítačových sítí je možné realizovat kromě centralizovaného umístění všech komponent vyhledávacího procesu i jejich distribuci v různých uzlech sítě. Tato distribuce se může týkat nejen jednotlivých komponent (např. vyhledávací program je umístěn v jiném uzlu sítě než prohledávané dokumenty), ale může nastat i v rámci komponenty samotné (tradičním typem jsou tzv. distribuované databáze). Názorně vyhledávání v distribuovaném prostředí ukazuje obrázek 5.



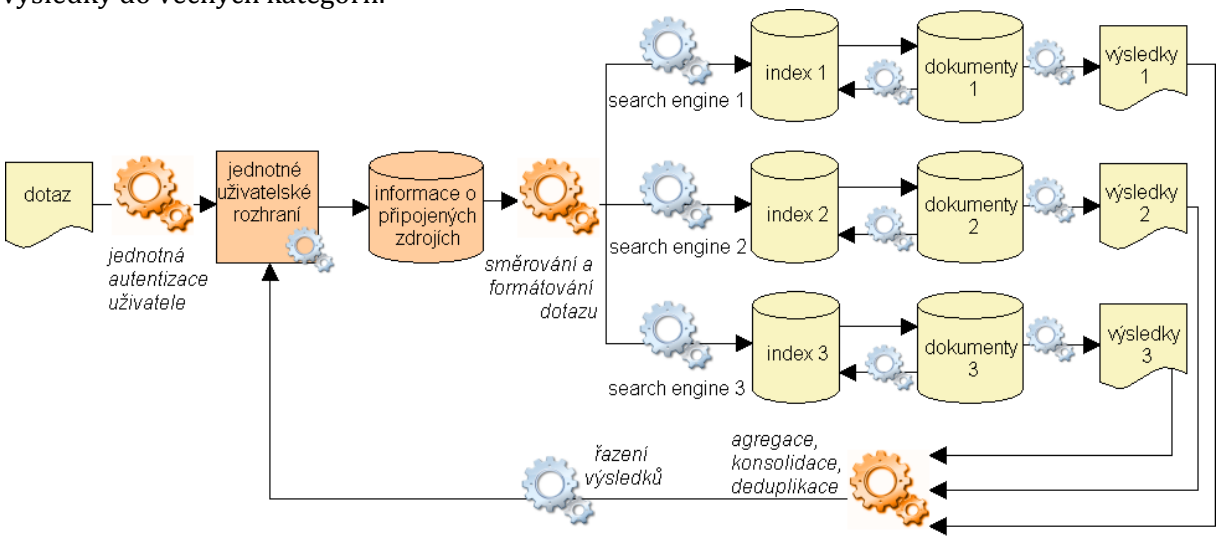
Obr. 5. Vyhledávání z více zdrojů

Obtížnost vyhledávání z více zdrojů nespočívá jen v množství operací na vstupu – jak je vidět z obrázku, uživatel musí svůj dotaz postupně zadávat do každého systému znovu a zpravidla se i pokaždé znovu do každého dalšího systému přihlašovat. Problém čeká i na výstupu – celkové vyhodnocení výsledku dotazu musí provést sám uživatel manuálně,

distribuované softwarové aplikace mu v tom nejsou nijak nápomocny. Nabízejí se dvě řešení: 1) sjednotit distribuované zdroje fyzicky do jednoho centrálního úložiště – touto cestou se ubírají agregované databáze (např. souborný katalog) a datové sklady; 2) ponechat distribuovaným zdrojům jejich samostatnost a sjednotit je virtuálně prostřednictvím jednotného uživatelského rozhraní. Nadále se budeme věnovat tomuto druhému přístupu a postupně si ukážeme možnosti sjednocování jednotlivých prvků distribuovaného vyhledávacího modelu – uživatelského rozhraní, vyhledávacího softwaru a indexového souboru. Uvidíme, že zatímco koncovému uživateli se každý rešeršní systém tohoto typu jeví stejně (jedno uživatelské rozhraní, jedna sada výsledků), technologie v pozadí takového virtuálního celku může být různá.

Varianta A – sjednocení uživatelského rozhraní a paralelní vyhledávání

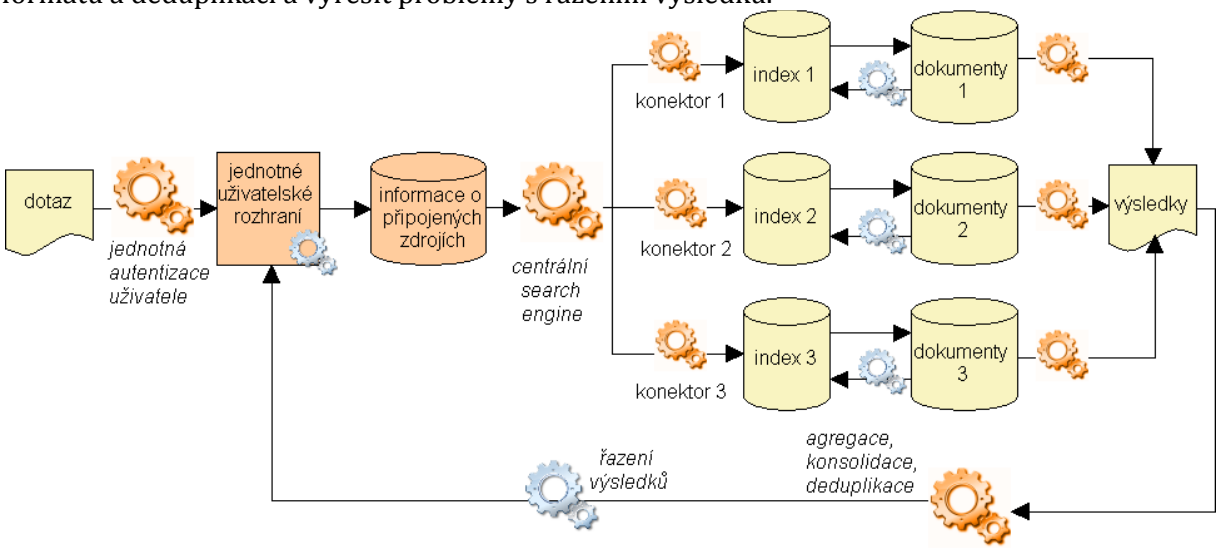
Tato technologie je známá a detailně zdokumentovaná v databázovém prostředí pod názvem federativní vyhledávání už od 80. let 20. století [5, 8]. Podstatou řešení je program, který převezme za uživatele úlohu zadávání dotazů různým zdrojům a sjednotí výsledky vyhledávání. Vyhledávací systém je zapotřebí doplnit na vstupu o další datový objekt – centralizované informace o připojených zdrojích. Typickým obsahem jsou metadata identifikující jednotlivé zdroje a popisující jejich vyhledávací protokol a jejich topologii. Další přidanou komponentou je softwarová aplikace zajišťující směrování, formátování a rozesílání dotazů distribuovaným vyhledávacím strojům. Tato aplikace obchází jednotlivá nativní uživatelská rozhraní a nahrazuje jejich aktivity – musí tedy umět komunikovat se všemi připojenými zdroji. Komunikační řetězec naznačený na obrázku 6 se může ještě prodloužit o různé zprostředkovatelské překladače či mezivrstvy, příkladem může být vyhledávání ve virtuálním souborném katalogu s použitím výměnného (komunikativního) formátu pomocí protokolu Z39.50. Samotné vyhledávání realizují i nadále nativní (proprietární) vyhledávací stroje. Poslední rozšiřující komponentou je software, který má za úkol fyzicky sjednotit dílčí výsledky. To vyžaduje sjednocení jejich formátu a odstranění duplicitních výskytů stejného obsahu, tj. buď jejich obsah sloučit do jednoho záznamu, nebo duplicitní záznamy odstranit[4]. Po sjednocení výsledků vyvstává potřeba jejich opětovného členění. Vzhledem k heterogenitě (zejména sémantické) zdrojů vznikají problémy s řazením podle věcné relevance. Přitom důležitost tohoto řazení výsledků stoupá úměrně zvýšenému množství položek vyhledaných ze sjednocených zdrojů – uživatel, který stojí o několik relevantních odkazů, není ochoten hledat je ve stovkách či tisících hitů. Roste význam faset a klastrů podporujících iterativní zpřesňování a zkvalitňování výsledků vyhledávání: fasety nabízejí kritéria pro filtrování a klastry seskupují výsledky do věcných kategorií.



Obr. 6. Sjednocení uživatelského rozhraní (sjednocení vstupů i výstupů)

Varianta B – centralizovaný vyhledávač a paralelní vyhledávání

V tomto řešení je k sjednocenému uživatelskému rozhraní a informacím o připojených zdrojích přidán centralizovaný vyhledávač (centrální *search engine* na obr. 7), který obchází nejen nativní uživatelská rozhraní, ale i nativní vyhledávací stroje. Aby se mohl zhostit jejich úlohy, musí „umět" vyhledávací algoritmy všech připojených zdrojů. Potřebuje ještě mechanismus, který zajistí fyzické připojení centrálního vyhledávacího stroje k jednotlivým distribuovaným zdrojům – to zajišťují softwarové aplikace zvané konektory. Tím, že souběžné vyhledávání probíhá v režii jednoho centrálního vyhledávače, se poněkud usnadňuje sjednocování výsledků vyhledávání, i v tomto případě je ovšem nutné provést sjednocení formátů a deduplikaci a vyřešit problémy s řazením výsledků.



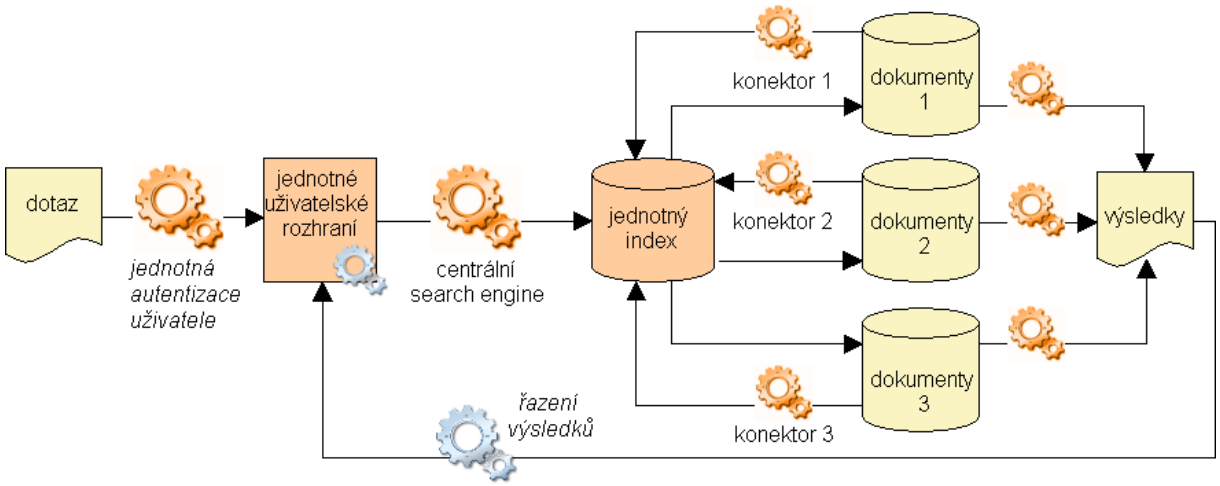
Obr. 7. Sjednocení vyhledávače (search engine)

Varianta C – sjednocení indexu a centrální vyhledávání

Tento model se osvědčil při konstrukci webových vyhledávačů. Ve sjednocování vyhledávacích komponent postupuje o další krok: místo nativních indexů jednotlivých zdrojů je vytvořen centrální index, který je prohledáván centrálním vyhledávacím strojem. Jak je ze schématu na obrázku 8 patrné, tento technologický model už se liší od vyhledávání v jednom zdroji pouze v jedné komponentě, kterou jsou distribuované samostatné heterogenní zdroje. Obdobně jako potřeboval centralizovaný vyhledávač konektory k prohledávaným zdrojům, potřebuje takové konektory centrální indexovač. Centrální index může být vytvořen dvěma způsoby: buď ho indexovač vygeneruje přímo z primárních dokumentů (tzv. sklizeň – *harvesting* nebo též prolézání – *crawling* zdrojů), a tím vlastně zopakuje aktivity, které už proběhly při tvorbě nativních indexů, nebo je možné převzít a sloučit už existující nativní indexy (tato operace bývá označována jako sklizeň metadat – *metadata harvesting*).

Producenti softwaru nabízející tuto technologii knihovnám pro ni používají název *discovery* vyhledávání, případně obtížně přeložitelný novotvar *web scale discovery*. *Discovery* by bylo možné do češtiny přeložit jako objev, nález, ale i zpřístupnění. *Web scale* (doslova webový rozsah) poukazuje na podobnost s „google vyhledáváním" a na obrovské množství zdrojů, jež je možné do centrálního indexu zahrnout. Na rozdíl od tzv. katalogů nové generace, které používají technologii *„discovery layer"* k vytvoření jednotného indexu pro všechny lokálně vlastněné knihovní kolekce (navíc se jedná zpravidla pouze o strukturovaný index na monografické či seriálové bibliografické úrovni), *web scale discovery* umožňuje sjednotit index jak pro lokální, tak pro vzdálený obsah (např. pro licencované online databáze, ale i pro

zdroje volně dostupné v prostředí Internetu), a to včetně indexace plných textů. Slovo *scale* navíc evokuje termín *„scalability“*, tedy škálovatelnost neboli rozšiřitelnost.



Obr. 8. Sjednocení indexu

Závěr

Představení tří různých modelů souběžného vyhledávání informací z více zdrojů přímo vybízí k jejich srovnání. V případě varianty A je kritickým místem pro efektivnost vyhledávání heterogenita jednotlivých zapojených vyhledávacích strojů. Platí, že kvalitu výsledků a rychlost odezvy určuje nejslabší článek. U varianty B je nejdůležitější výkonnost a robustnost centrálního vyhledávacího stroje a kvalita konektorů. Jak varianta A, tak varianta B nevyžadují vytvářet redundantní datová úložiště. Vyhledávání v distribuovaných zdrojích probíhá vždy v reálném čase nad aktuálními daty. Varianta C umožňuje díky centrálnímu indexu výrazně zrychlit proces vyhledávání a řazení výsledků. Jako jediná z představených modelů vyžaduje generování redundantního datového úložiště, které musí být navíc schopno pokud možno průběžné aktualizace, protože jinak by hrozilo, že se vyhledávání bude realizovat v zastaralém indexovém souboru. Kvalita indexového souboru obecně představuje kritické místo tohoto modelu.

Vzhledem k tomu, že všechny tři představené modely jsou alternativou, nikoli nahrazením dosavadního způsobu vyhledávání, bylo by možné uvažovat i o srovnání souběžného vyhledávání obecně s vyhledáváním v původním nativním prostředí jednotlivých zdrojů. Přínosem souběžného vyhledávání je bezesporu usnadnění pro koncového uživatele – jednotné uživatelské rozhraní, jednotná autentizace, jednotný seznam výsledků. Dalším přínosem je enormní nárůst počtu zdrojů, které může tímto způsobem uživatel prohledávat. V tomto případě kvantita může přejít v novou kvalitu, je totiž pravděpodobné, že v tak ohromném množství zdrojů se skutečně podaří objevit informace, které by jinak zůstaly skryty. To platí do jisté míry i pro možnosti zpracování výsledků vyhledávání získaných z distribuovaných zdrojů – kupříkladu ad hoc generované klastry ze sjednocených výsledků dotazů přinášejí přidanou hodnotu, kterou by opět nebylo možné získat z jednotlivých zdrojů. Problémem souběžného vyhledávání je nutnost přidávat do rešeršního systému další komponenty a procesy, protože v každém novém komunikačním mezičlánku zákonitě dochází k šumu, který se může projevit jak zkreslením významu, tak ztrátou informace. Kromě toho každý zpracovatelský proces působí časovou ztrátu. Určitou nevýhodou zejména pro profesionály zvyklé na práci s individuálními zdroji by mohlo být, že souběžné vyhledávací systémy se jeví jako černá skříňka, objevují se i námitky vůči slučování výsledků rozdílné kvality a důvěryhodnosti.

Srovnání modelu vyhledávání z jednoho zdroje s variantami A, B a C souběžného vyhledávání ukazuje, jak se postupně mění technologie zpřístupňování elektronického obsahu. Zatímco modely na obrázcích 4 a 5 představují obsah pevně spojený s proprietárním vyhledávacím softwarem a uživatelským rozhraním, na obrázku 6 je vidět možnost oddělení obsahu od nativního uživatelského rozhraní; na obrázku 7 už je osamostatněn obsah i od vyhledávacího softwaru a konečně obrázek 8 ukazuje „čistá" data, která jsou k dispozici libovolnému vyhledávači, který se k nim dokáže připojit. Zdá se, že vize sémantického webu a propojených dat *(linked data)* zde nachází jednu ze svých konkrétních realizací.

Literatura:

1. BRIN, Sergey; PAGE, Lawrence. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*. 1998, vol. 30, issues 1-7, s. 107-117. Rozšířená verze: BRIN, Sergey; PAGE, Lawrence. The anatomy of a large-scale hypertextual Web search engine. In *Proceedings of the Seventh World Wide Web Conference (WWW7), April 14-18 1998*. Brisbane, 1998 [online] [cit. 2011-07-11]. Dostupné z WWW: <http://www7.scu.edu.au/1921/com1921.htm/>.

2. BUCKLAND, Michael K.; PLAUNT, Christian. On the construction of selection systems. *Library Hi Tech*. 1994 [cit. 2011-07-11], vol. 12, no. 4 (48), s. 15-28. ISSN 0737-8831. Dostupné z 6WWW: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.52.5322/>.

3. FELDMAN, Susan. The Answer Machine: Are we there yet? *Searcher*. 2011, vol. 19, no. 1 (January/February), s. 18-27. ISSN 1070-4795.

4. GALAMBOŠ, Leo. Vyhledávání na webu. In *DATAKON 2007: sborník databázové konference*. Brno : Masarykova univerzita, 2007, s. 17-24 [cit. 2011-07-11]. Dostupné z WWW: <http://www.datakon.cz/datakon08/sbornik_datakon07.pdf/>.

5. HEIMBIGNER, Dennis; McLEOD, Dennis. A federated architecture for information management. *ACM Transactions on Office Information Systems.* 1985, vol. 3, no. 3 (July), s. 253-278 [cit. 2011-07-11]. ISSN 1046-8188. DOI 10.1145/4229.4233.

6. LARSON, Ray R. Information retrieval systems. In *Encyclopedia of library and information sciences*. 3rd ed. Taylor & Francis, 2010, s. 2553-2563 [cit. 2011-07-11]. DOI 10.1081/E-ELIS3-1200440222.

7. POKORNÝ, Jaroslav; SNÁŠEL, Václav; KOPECKÝ, Michal. *Dokumentografické informační systémy*. 2. přeprac. vyd. Praha : Karolinum, 2005. Kap. 6.3, Architektury vyhledávacích strojů, s. 134-137. ISBN 80-246-1148-1.

8. SHETH, Amit P.; LARSON, James A. Federated database systems for managing distributed, heterogenous, and autonomous databases. *ACM computing surveys*. 1990, vol. 22, no. 3, s. 183-236. ISSN 0360-0300.

9. SONG, Min; SONG, Il-Yeol; CHEN, Peter P. Design and development of a cross search engine for multiple heterogeneous databases using UML and design patterns. *Information Systems Frontiers*. 2004, vol. 6, no. 1 (March), s. 77-90. ISSN 1387-3326. DOI 10.1023/B:ISFI.0000015876.14848.8c.

10. VAUGHAN, Jason. Web Scale Discovery Services. *Library Technology Reports*. 2011, vol. 47, no. 1 (January), s. 5-61. ISSN 0024-2586.

11. VICKERY, Alina; VICKERY, Brian Campbell. Information retrieval. In *Information science in theory and practice.* 3rd rev. and enl. edition. Berlin, New York : Walter de Gruyter – K. G. Saur, 2004. Chapter 5, s. 116-132. Print ISBN 978-3-598-11658-2. eBook ISBN 978-3-598-44008-3. DOI 10.1515/9783598440083.

12. ZENG, Marcia Lei; CHAN, Lois Mai. Metadata interoperability and standardization – a study of methodology, part II: achieving interoperability at the record and repository levels. *D-Lib Magazine*. 2006, vol. 12, no. 6 (June) [cit. 2011-07-11]. ISSN 1082-9873. Dostupné z WWW: <http://www.dlib.org/dlib/june06/zeng/06zeng.html/>.

13. ZENG, Marcia Lei; QIN, Jian. *Metadata*. New York : Neal-Schuman, 2008. Kap. 6.5.1, Metadata retrieval, s. 233-237. ISBN 978-1-55570-635-7.

Poznámky:

[1] Podle viceprezidentky pro vyhledávací technologie firmy IDC Susan Feldmanové většina organizací včetně středně velkých provozuje minimálně 50-100 různých informačních repozitářů nebo informačních kolekcí [3]. Typická odborná či akademická knihovna dnes kromě katalogu nabízí svým čtenářům desítky až stovky kolekcí elektronických zdrojů – namátkou například NK ČR provozuje 16 různých portálů (jeden z nich, specializovaná oborová brána Knihovnictví a informační věda, nabízí 37 různých zdrojů), v sekci Katalogy a databáze NK ČR nabízí 15 zdrojů a v sekci Licencované online zdroje dalších 60 zdrojů; Portál elektronických zdrojů Univerzity Karlovy obsahuje na 250 zdrojů.

[2] Aby terminologických potíží nebylo málo, často je taková komplexní aplikace rovněž nazývána *search engine* (takový přístup je uplatněn např. v [1]).

[3] Případný zájemce může najít aktuální informace o řešení této problematiky např. v příspěvku V. Jansy z konference Infos 2011: JANSA, Václav. Čtenář, jeho elektronická identita a služby knihovny. In *Infos 2011 :* Zborník z 36. medzinárodného informatického sympózia o postavení a úlohách pamäťových inštitúcií v oblasti rozvoja kultúry, vedy, techniky a vzdelávania, Stará Lesná – Vysoké Tatry, 9.-12. 5. 2011, s. 89-99 [cit. 2011-07-11]. Dostupné z WWW: <http://vili.uniba.sk/AK/INFOS2011.pdf>.

[4] Tato operace je velmi důležitá, protože obsah souběžně prohledávaných zdrojů často není disjunktní, ale částečně se překrývá. Kupříkladu text citovaného článku z časopisu Searcher [3] je kromě „volného" Internetu (http://www.infotoday.com/searcher) dostupný ve čtyřech kolekcích EBSCOhost (Academic Search Complete, Business Source Complete, Computers and Applied Sciences Complete a Library, Information Science & Technology Abstracts with full text) a ve dvou zdrojích Wilson (Library Literature and Information Science Full Text a OmniFile FullText Select), bibliografický záznam bez plného textu je kromě uvedených zdrojů obsažen ještě v databázi ProQuest Technology Collection.