PROBLEMATIKA VĚCNÉHO POŘÁDÁNÍ INFORMACÍ A JEJICH ZPŘÍSTUPNĚNÍ

Marie Balíková
Národní knihovna ČR

Cílem tohoto příspěvku je uvést do problematiky věcného pořádání informací a jejich zpřístupnění. Zabývá se základními charakteristikami tradičních pořádacích systémů a možnostmi jejich integrace a harmonizace. Jednotlivá dílčí témata budou rozpracována v příspěvcích, které budou publikovány v dalších číslech časopisu.

Věcné pořádání informací představuje dílčí proces vstupního zpracování informací, který vychází z obsahové analýzy dokumentu. Výsledkem obsahové analýzy dokumentu je selekční obraz dokumentu, tj. sestava obsahových údajů, které vypovídají o obsahu, tématu či předmětu dokumentu jako celku, jeho jednotlivých částí nebo dokonce jednotlivých, v dokumentu obsažených informací.[1)]

Cílem věcného pořádání informací je vytvořit takový selekční obraz dokumentu, který by umožnil koncovému uživateli dosáhnout při procesu vyhledávání s podporou systémových nástrojů optimální přesnosti a optimální úplnosti, tedy zpřístupnit požadované relevantní dokumenty efektivním způsobem.

Podmínkou racionálního věcného pořádání informací a jejich adekvátního zpřístupnění je inteligentní porozumění textu, předpokládá se aktivní účast přirozené či umělé inteligence v tomto procesu.

Charakteristika tradičního procesu pořádání informací a jejich zpřístupnění

Pro věcné pořádání informací jsou rozhodující obsahové údaje, které byly zjištěny obsahovou analýzou. Pořádacími znaky přirozeného, formalizovaného nebo umělého informačního (selekčního) jazyka, jak se projevuje na vstupu do informačního systému i na výstupu z něho, mohou být v nejjednodušší podobě slova převzatá z názvu nebo podnázvu dokumentu, případně slova převzatá z jiného místa dokumentu, z redukce textu dokumentu v anotaci či referátu, nebo formulovaná informačním pracovníkem, a to zcela volně, či na podkladě nějakého řízeného slovníku. Na znaky téhož jazyka jsou převáděny také informační dotazy. Znaky, jimiž jsou při věcném pořádání informací označovány obsahy dokumentů, jsou pak směrodatné také pro uložení informací do jakékoli tradiční či netradiční informační paměti, pokud je uspořádána podle obsahových údajů o dokumentech. Věcné pořádání informací úzce souvisí s referováním a anotací. Dříve než informační pracovník zformuluje předmětové heslo, množinu unitermů či deskriptorů, musí nejprve ve svém vědomí provést určitou redukci nebo kondenzaci textu ve smyslu souvislé, syntakticky svázané věty nebo stručnějšího či rozsáhlejšího souboru vět. Na podkladě této redukce se pak provádí další redukce, jejímž výsledkem je předmětové heslo, nebo sestava deskriptorů. Referát a anotace mohou však být vytvářeny i záměrně a písemně fixovány. Jako formy redukce textu dokumentu mohou být podkladem pro věcné pořádání informací, ze kterého informační pracovník (a v posledních letech stále častěji i samočinný počítač podle určitého algoritmu) vybírá takové termíny, které mu umožňují vystihnout obsah dokumentu nebo jednotlivých v něm obsažených informací a které se spolu s ostatními termíny (odjinud z dokumentu) stávají součástí selekčního obrazu dokumentu.[2)]

Charakteristika procesu pořádání informací a jejich zpřístupnění v prostředí internetu

V současném období informační exploze jsou předmětem mnoha diskusí otázky, s jakou pravděpodobností lze v reálném čase, zvláště pak v prostředí internetu, nalézt relevantní informace odpovídající uživatelovu dotazu a jak odlišit informace fiktivní a ničím nepodložené od informací ověřených a exaktních. Od vyhledávacích systémů se očekává, že nabídnou uživateli informace po obsahové stránce již zhodnocené (utříděné, zhuštěné). Koncový uživatel by uvítal redukci objemu textů, které má k dispozici při vyhledávání. Algoritmizací procesu porozumění textu a jeho následnou komprimací či interpretací získáme výsledný text, který představuje referát a anotaci. Při komprimaci textu se používají metody statistické (jednotky s vyšší frekvencí výskytu jsou nejzávažnějšími nositeli obsahu), syntaktické (obsahově nejzávažnější úseky textu se stanoví na základě porovnání syntaktických struktur textu se slovníkem vzorových syntaktických struktur) a metody sémantické (na základě sémantické analýzy textu se stanoví obsahově zatížené prvky: nejjednodušší – slova z názvu, podnázvu, složitější – sémantický rozklad na dílčí významově zatížené úseky promluvy – uplatňuje se lingvistická analýza textu). Některé programy pracující těmito metodami umožňují sestavit smysluplný text i z výrazů, které neobsahoval výchozí text, podstatně krátí výchozí text (až o 80 %), sumarizují text, vynechávají redundantní věty, registrují podstatné numerické údaje. Výsledkem této komprimace je text, který má plně nahradit text výchozí.

Jiné systémy pracují na bázi slovníků a vyhodnocují frekvenci výskytů termínu v analyzovaném dokumentu vzhledem k frekvenci výskytů v celém souboru extrahovaných dokumentů, jsou schopny identifikovat jména osob, organizací, lokalit i ve víceslovných výrazech stejného významu, ale odlišného způsobu zápisu. Identifikují jednotlivé technické termíny, mohou identifikovat zkratky i v jejich rozepsané podobě.

Pomocí některých systémů je možno přiřadit souboru analyzovaných textů předdefinované předmětové katego-

rie (Categorization tool), jiné jsou schopny uspořádat vyhledané texty do shluků podle obsahové podobnosti (Clustering tool).[3)]

Oba odkazované texty nám poslouží jako odrazový můstek k dalším úvahám na téma věcného pořádání informací. První text je datován do poloviny 70. let a odráží první nesmělé krůčky v zavádění informačních technologií do procesu pořádání informací. Ve druhé ukázce je patrný nástup znalostních (pojmově orientovaných) systémů.

Avšak oba texty se v základních rysech shodují. Hovoří se v nich o pořádání informací a jejich přeměně ve znalosti, pracují se shodnými pojmy, jako jsou obsahová/textová analýza, řízené/vzorové slovníky, redukce a komprimace textu (anotace, referát), kategorizace, klasifikace. Oba texty pracují s pojmem inteligentní porozumění textu: v 70. letech převládá podíl lidské inteligence na procesu pořádání informací, v průběhu 90. let se zdá, že tuto oblast zcela opanují systémy umělé inteligence a zásah lidského intelektu při pořádání webovských zdrojů nebude nutný.

Toto očekávání se zcela nenaplnilo. Dodejme, že prozatím. Počítačový systém zatím není vybaven schopností porozumět textu. Algoritmy, které by měly umožnit počítači textu porozumět, jsou natolik složité, že jsou prakticky nepoužitelné.

Proto se zpřístupněním webovských dokumentů zabývají v převážné míře internetové roboty.

Služby typu „search engines", fulltextové prohlížeče jsou založeny na automatizovaném sběru dat. Předmětem jejich činnosti je indexace internetových stránek, které poté v pravidelných intervalech ukládají do svých pamětí. Takto vzniklé databáze používají při procesu vyhledávání. Při formulaci dotazu používají pravostranné krácení, Booleovské operátory, umožňují hledání podle fráze, částečně umožňují stemming. Dosahují však pouze 20% úspěšnosti, protože při vyhodnocování dotazu pracují nejčastěji s frekvenční strukturou textu. Jsou charakteristické množstvím zpracovaných informací, nezaručují však kvalitu těchto informací.[4)]

Další typ těchto služeb představují předmětově orientované vyhledávací nástroje, katalogové prohlížeče, které zpracovávají a zpřístupňují dokumenty prostřednictvím hierarchicky uspořádaných předmětových kategorií. Zpracovávají menší objem informací, tyto informace jsou však vyhodnoceny.

Do této kategorie vyhledávacích služeb můžeme řadit i tzv. informační brány.

Informační brána je předmětově orientovaný zdroj, který

- umožňuje jednotný přístup k heterogenním zdrojům/objektům v elektronické podobě
- zpřístupňuje databáze sekundárních dokumentů v elektronické podobě
- zpřístupňuje sbírky kvalitních (řízených, kontrolovaných) původních elektronických zdrojů sestavených podle předem stanovených kritérií (obdoba budování klasického fondu)
- nabízí kvalitní popis těchto zdrojů na podkladě inteligentního porozumění textu
- nabízí řízený přístup k těmto zdrojům pomocí intelektuálně definovaných předmětových kategorií tvořených v současnosti většinou na podkladě klasifikačních schémat

S problematikou informačních bran úzce souvisí kategorizace a klasifikace informačních zdrojů, která předpokládá jistou míru integrace tradičních pořádacích systémů do současného procesu organizace informací. Tato integrace představuje na jedné straně velkou výzvu pro „klasické" knihovníky, kteří se považují za osoby povolané převzít zodpovědnost za organizaci poznatků univerza i v tomto prostředí. Na straně druhé tato integrace vzbuzuje nemalé obavy u informačních pracovníků zabývajících se zpřístupněním zdrojů na webu, kteří tyto nástroje považují za zastaralé a neschopné jakékoli transformace.

Jedním z cílů této práce je přispět k integraci tradičních pořádacích systémů (univerzálních i specializovaných) do současného procesu vyhledávání. Předpokladem takovéto integrace je transformace a harmonizace těchto pořádacích systémů.

Mezi tradiční pořádací systémy řadíme systémy předmětových hesel, tematické a polytematické tezaury a klasifikační systémy.

Charakteristika jednotlivých systémů

Jazyk předmětových hesel představuje nástroj, ve kterém je téma dokumentu vyjádřeno sestavou lexikálních jednotek podle předem stanovených syntagmatických a syntaktických pravidel už v průběhu indexování. V lístkových katalozích a u první generace OPAC katalogů tvořila tato sestava nedílnou součást v celém procesu pořádání informací, bylo možné do ní vstupovat pouze prostřednictvím vstupní pozice; později v souvislosti se zdokonalením systémových nástrojů pro vyhledávání, tj. s možností vyhledávat po slovech, bylo možné vstupovat do systému a vyhledávat i prostřednictvím jednotlivých segmentů předmětového hesla (podhesel, zpřesnění).

Tento systém sloužil v prostředí lístkových (manuálních) katalogů. Předmět dokumentu byl vyjádřen několika předmětovými hesly (3-6), přičemž předmětové heslo představovalo strukturu (větu) tvořenou podhesly a uspořádanou podle předem stanovených syntaktických pravidel. Slovní zásobu tohoto jazyka představoval soubor předmětových hesel sestavených z dílčích lexikálních jednotek. Nejčastějším reprezentantem dílčí lexikální jednotky bylo jednoslovné substantivum. Do systému bylo v konkrétně stanovených situacích možno zařadit i víceslovné lexikální jednotky, jejichž typickou vlastností byla inverze (priorita substantiva). Systémy však měly tendenci tyto víceslovné lexikální jednotky rozkládat (*výkon vazby: vazba – výkon*). Soubor předmětových hesel se vytvářel induktivní metodou (tvorba „zdola"), na základě konkrétního materiálu. Mezi lexikálními jednotkami byly definovány dva typy vztahů: vylučovací a přidružovací. V syntaktické

rovině byla dominantním principem prekoordinace, kdy byla definována komplexní pravidla umožňující vystihnout obsah dokumentu detailním způsobem. Standardizace se řešila povýšením řetězce předmětového hesla do autoritního souboru. K nesporným výhodám tohoto systému je nutno zařadit

- detailní vyjádření předmětu dokumentu
- odpovídající specifičnost
- maximální informační hodnotu řetězce předmětového hesla
- efektivní servis pro uživatele v tradičním prostředí

K nevýhodám, které vynikly až v elektronickém prostředí, patří

- rozklad víceslovných lexikálních jednotek
- délka řetězce předmětového hesla – koncové údaje se nezobrazí vůbec, případně na druhé řádce
- princip prekoordinace uplatněný v syntaktické rovině, a z toho pramenící
- komplikovaná pravidla aplikační syntaxe
- redundantnost informací v bibliografickém záznamu
- rozsáhlost autoritního souboru
- komplikovaná údržba

Jazyky deskriptorového typu jsou jazyky postkoordinované. Téma dokumentu je vyjádřeno sestavou izolovaných jednotek, mezi nimiž nejsou syntaktické vztahy explicitně vyjádřeny.

Typickým reprezentantem jsou oborové tezaury, které se plně rozvinuly zvláště v prostředí elektronických katalogů v souvislosti s nástupem vyhledávacích systémů 2. generace, charakterizovaných možností vyhledávat pomocí kombinace slov. Předmět dokumentu bývá vyjádřen sestavou izolovaných lexikálních jednotek – deskriptorů, variantní vstup do systému bývá zajištěn pomocí nedeskriptorů. Typickým reprezentantem lexikální jednotky je jednoslovné substantivum, přičemž je možno zařadit do systému i víceslovné lexikální jednotky v přirozeném slovosledu, avšak velmi často dochází k jejich rozkladu. Typické pro tezaury je důsledné vyjádření vztahů mezi lexikálními jednotkami, a to ekvivalence (synonymie, homonymie), hierarchie (nadřazenost, podřazenost), asociace. Při vyjádření komplexního obrazu dokumentu je důsledně uplatněn princip postkoordinace. Autoritní soubor vzniká deduktivní metodou (postup „shora").
K výhodám tohoto systému patří

- princip postkoordinace a z toho pramenící
- přehlednost selekčního obrazu dokumentu
- snadná tvorba hierarchických struktur
- snadná údržba
- snadná manipulace

K nevýhodám patří

- rozklad víceslovných jednotek – nerespektování kompaktnosti termínu
- informační šum způsobený parazitním (náhodným) spojením deskriptorů a v důsledku toho
- velký ohlas irelevantních dokumentů
- omezení pouze na tematickou část obsahové charakteristiky dokumentu: systém deskriptorů a nedeskriptorů zahrnuje pouze tematické termíny; tento nedostatek se odstraňuje připojením podpůrných souborů identifikátorů (personália, jména korporací, geografické názvy, atd.)

Klasifikační systémy

MDT (Mezinárodní desetinné třídění) je univerzální klasifikační systém, který slouží k indexování a vyhledávání věcných informací o dokumentech, jejich částech, případně k indexování a vyhledávání jednotlivých informací v dokumentech obsažených.

Tento systém bývá charakterizován jako kombinace třídění hierarchického a fasetového typu. Předností tohoto pořádacího systému je univerzálnost a flexibilita.

Univerzálnost je dána

- schopností indexovat veškeré poznatky univerza a uspořádat je v rámci jednotlivých vědních oborů, tj. tříd, podtříd, oddílů atd.
- schopností respektovat specifičnost termínu a adekvátním způsobem ji vyjádřit
- možností volby rozsahu používaných notací; MDT lze používat v úplné, střední i zkrácené verzi, např. jako součást věcných údajů lze v bibliografických záznamech ukládat znak středního rozsahu, v bibliografických soupisech a při organizování knihovního fondu ve volném výběru jeho zkrácenou verzi
- mezinárodní srozumitelností, tedy nezávislostí na přirozeném jazyce, neboť obsahový údaj je vyjádřen soustavou arabských číslic; srovnáme-li tento klasifikační systém s jinými, např. DDC (Dewey Decimal Classification), zjišťujeme, že abecední znaky se v systému MDT uplatňují ve velmi omezené míře.
- bohatostí používaných znaků
- aktuálností slovního vyjádření
- rozšířením tohoto klasifikačního systému; MDT se používá ve 24 zemích světa, především v Evropě, v Latinské Americe a částečně v Africe

Flexibilita systému MDT je podmíněna strukturou používaných notací MDT a jejich zápisem. Většinu prvků složené notace lze považovat za relativně samostatné celky a zapisovat je individuálně do jednotlivých polí/podpolí (podle možností používaného knihovnického systému). Jejich volba a především tzv. citační pořadí, tj. pořadí v rámci celého řetězce notačních znaků MDT, spočívá ve velké míře na rozhodnutí dané katalogizační agentury. Toto přizpůsobení se potřebám jednotlivých institucí představuje zároveň i úskalí pro jednotnou aplikaci MDT, protože institucionální, národní i mezinárodní kooperace předpokládá shodný a na základě konvencí dohodnutý jednotný postup v aplikaci standardů.

Z hlediska uživatele bývají klasifikační systémy, tedy i systematický selekční jazyk MDT, považovány za uživatelsky nevstřícné, zvláště pak ve srovnání se selekčními systémy založenými na bázi přirozených jazyků. Dále je jim vytýkána nedůslednost v hierarchických strukturách a technické problémy, které notační znaky bohaté na diakritiku způsobují jednotlivým integrovaným systémům.

Z charakteristik jednotlivých systémů vyplývá, že uvedené selekční jazyky mají v sobě zabudovány nástroje pro kategorizaci a klasifikaci elektronických zdrojů.

Požadavky kladené na organizační nástroj moderního typu operující v prostředí internetu můžeme shrnout takto:

Věcný selekční jazyk v elektronickém prostředí by měl umožnit

- získat v reálném čase relevantní dokumenty
- dopracovat uživatelův dotaz
- pomoci definovat účinné rešeršní strategie

K tomu musí splňovat následující podmínky

- jednoduchá struktura umožňující aplikaci v různých formátových prostředích (metadatové formáty, marcové formáty)
- snadná údržba
- univerzálnost tematická i objektová (umožňující postihnout i objekty netextové povahy)
- potřebná míra specifičnosti
- schopnost vyjádřit formální a sémantické vztahy mezi lexikálními jednotkami
- v oblasti lexikálních jednotek umožnit přímou vazbu na plnotextové vyhledávání

Z tohoto pohledu je zřejmé, že transformace věcných selekčních jazyků je nevyhnutelná, a to v oblasti

- jazyka, tj. lexikální jednotky se zvláštním důrazem na její význam
- aplikační syntaxe u prekoordinovaných jazyků
- unifikace/harmonizace věcných selekčních jazyků

Kvalita procesu pořádání informací zvláště v elektronickém prostředí je ovlivněna

- kvalitou selekčního jazyka[5)]
- zvolenou rešeršní strategií a danými systémovými rešeršními nástroji
- hodnotícími algoritmy při stanovení relevance dokumentů
- způsobem prezentace věcných údajů

Selekční jazyk

Kvalita selekčního jazyka je dána mírou jeho schopnosti adekvátním způsobem vyjádřit obsah dokumentů, poskytnout selekční obraz dokumentu. „Je značně závislá na tom, do jaké míry odráží selekční jazyk strukturu textu. Aby bylo možné splnit tyto podmínky, musel by selekční jazyk informačního systému obsahovat alespoň některé ze signifikantních prvků textu."[6)] Je třeba si však uvědomit, že role selekčního jazyka v plnotextových systémech je rozdílná od role selekčního jazyka v databázích sekundárních dokumentů (bibliografických). Koncový uživatel sice vyhledává v obou typech systémů pomocí lexikálních jednotek selekčního jazyka, ale role těchto jednotek je odlišná: v bibliografických databázích jsou lexikální jednotky součástí selekčního obrazu dokumentu; jsou tvořené, případně vybrané indexátorem z předepsané a řízené množiny selekčních prvků (předmětových hesel, deskriptorů), přičemž mohou, ale nemusejí být i součástí dokumentů samotných. Lexikální jednotky zde slouží ke zpřístupnění obsahu dokumentu v procesu indexování (formulace obsahu dokumentu prostřednictvím předepsaných lexikálních jednotek) i k vyhledání (formulace dotazu uživatelem prostřednictvím jeho vlastních slov, případně pomocí předepsaných lexikálních jednotek). V plnotextových systémech lexikální jednotky neslouží k indexování dokumentů, nýbrž jsou vždy součástí dokumentu samotného; dokument je vyhledán na základě shody lexikálních jednotek obsažených v uživatelově dotazu a jednotek obsažených v dokumentu za podpory systémových nástrojů informačních technologií.

Je-li tato shoda stanovena pouze na podkladě frekvenční analýzy bez obsahového/pojmového porozumění textu, výsledná relevance i pertinence vyhledaných dokumentů je nízká. Je nezbytné, aby do tohoto procesu byly zakomponovány moderní lingvistické metody vycházející např. ze synchronního modelu jazyka, který zkoumá schopnost jednotlivých jazykových prostředků ztvárnit/vyjádřit význam. Z hlediska synchronního pohledu rozlišujeme rovinu morfologickou (zkoumá morfém jako nejmenší významovou jednotku), lexikální (zkoumá lexikální význam slova), syntaktickou (zkoumá slovo v kontextu fráze, věty), sémantickou (zabývá se významem slova v kontextu), rovinu promluvy (zkoumá význam věty v kontextu), pragmatickou (zabývá se participací zkušeností/poznatků individua při vnímání obsahu textu).

Pro proces indexování a vyhledávání dokumentů v databázích sekundárních dokumentů je nejdůležitější rovina lexikální, při zamýšlené integraci tradičních pořádacích systémů do současného procesu vyhledávání vystupuje do popředí i rovina sémantická (slovo v kontextu). V pojmově orientovaných plnotextových systémech jsou do procesu porozumění obsahu textu a ztvárnění jeho selekčního obrazu zakomponovány všechny jazykové roviny.

K signifikantním prvkům obsahu textu společným všem selekčním jazykům na bázi přirozených jazyků patří především lexikální jednotky.

Lexikální jednotka/preferovaný termín selekčního jazyka má „obecně vzato, vyjadřovat terminologii, kterou nacházíme v literatuře, a to zcela nezávisle na tom, kolik jednotlivých slov je v daném případě zapotřebí k vyjádření pojmu"[7)].

Tato zásada je při automatickém zpracování textu dodržena: termín (v podstatě každé smysluplné slovo) je podle kritérií vlastních selekčnímu jazyku vybrán a povýšen na termín preferovaný.

Při této příležitosti je nutno zmínit jeden z nejdůležitějších aspektů, a to kompaktnost (celistvost, jednolitost) termínů/lexikálních jednotek.

V pojmově orientovaných systémech je termín „rozkládán" na jednotlivé podpojmy a je především uváděn v kontext (předmětem zájmu je sémantické okolí termínu).

V tradičních pořádacích systémech, které se používají při vyhledávání v bibliografických databázích, dochází často k formálnímu rozkladu termínu.

V procesu organizace informací je žádoucí, aby se obsah lexikálních jednotek (preferovaných termínů) co nejvíce blížil obsahu termínů používaných v odborné literatuře a při odborné komunikaci. Identita formy preferované lexikální jednotky a jejího sémantického obsahu na straně jedné s obsahem termínu používaného v odborné literatuře a obsahem pojmu (denotátem) na straně druhé výraznou měrou ovlivňuje přesnost formulace uživatelova dotazu, přesnost vyhodnocování získaných dokumentů, tudíž i kvalitu procesu pořádání a vyhledávání.

Pojem se vztahuje vždy na určitý objekt a je abstraktní povahy. Od objektu se liší tím, že odráží jenom jeho podstatné znaky. Vědní obory jsou vybudovány pomocí systému pojmů, v němž jednotlivé pojmy mohou vystupovat jako nezávislé nebo závislé. Jsou-li závislé, mohou být rovnocenné nebo nerovnocenné, nerovnocenné jsou ve vztahu nadřazenosti nebo podřazenosti.

Vztah pojem – termín lze charakterizovat takto:

Pojem je nositelem kategoriálních vlastností obsahu, termín kategoriálních vlastností formy. Ve vztahu pojem – termín představuje pojem obsah a termín formu. Obsah pojmu se vyvíjí s postupujícím poznáním, forma je stálá, tak dochází k tenzi mezi obsahem pojmu a jeho formou, tj. termínem.

Ve vztahu k obsahu pojmu mohou nastat tyto možnosti:

- termín je indiferentní, neutrální k obsahu – většinou jde o převzaté mezinárodní termíny (*technologie*)
- termín je v souladu s obsahem pojmu, je správně motivovaný (*letadlo*)
- termín odporuje obsahu pojmu – je nutno jej odstranit z terminologické soustavy daného oboru (*dielektrická konstanta* – nahrazeno termínem *permitivita*, neboť se ukázalo, že veličina není fyzikální konstantou)[8)]

Vlastnosti termínu

- jednoznačnost – jeden pojem = jeden termín (jednoznačný termín je srozumitelný i mimo kontext); ne všechny termíny jsou jednoznačné, viz problém synonymie a homonymie v univerzálních pořádacích systémech
- kompaktnost (celistvost, jednolitost)
- systémovost – zařaditelný na jedno místo v systematice oboru
- motivovanost – termín má vyjadřovat některý z podstatných znaků pojmu, což má praktický dopad na výběr preferovaného termínu: při existenci několika synonym má být vybrán termín s lepší motivovaností
- přesnost – co největší míra shody mezi obsahem pojmu a obsahem termínu
- spisovnost
- ustálenost
- krátkost
- přeložitelnost
- všeobecná přijatelnost a srozumitelnost
- osvojitelnost

Kvalitu použitých jazykových prostředků selekčního systému, kvalitu indexování, a tedy i kvalitu selekčního systému, výrazně ovlivňuje soulad, tj. identičnost obsahu pojmu, obsahu termínu a obsahu lexikální jednotky (u lexikální jednotky k tomu navíc přispívá i soulad mezi obsahem a formou). Zásada, „že deskriptory by měly být minimálními lexikálními jednotkami jak po stránce významové, tak i výrazové, a volit tedy jako základní prvky selekčního jazyka (tezauru) převážně jednoslovné pojmy, které se pak mohou kombinovat do libovolně početně stanovených řetězců schopných vystihnout obsah polytematických dokumentů"[9)], vede k formálnímu rozkladu termínu a jeho náhradě několika lexikálními jednotkami z principu obecnějšího významu. Proces vyhledávání je tak posunut do obecnější roviny, což v plnotextových bázích má za následek větší ohlas irelevantních dokumentů. Proto je nutné posoudit oprávněnost existence víceslovných spojení a stanovit kritéria pro jejich používání. Do systému selekčního jazyka je tedy nutné zahrnout i víceslovné lexikální jednotky, které jsou pro uchopení obsahu dokumentu (textu) podstatné. Předpokladem jejich plné integrace do systémů všech typů je jejich jednoznačná identifikace, v opačném případě jejich přítomnost „prudce snižuje schopnost popisu obsahu textu i vyhledávací schopnost selekčního jazyka".[10)]

K této jednoznačné identifikaci víceslovných lexikálních jednotek zahrnutých do selekčních jazyků mohou výraznou měrou přispět tradiční pořádací systémy, zvláště pak tezaury polytematického a univerzálního charakteru, které představují soubor přesně vymezených termínů, propojených přesně definovanými vztahy.

Víceslovné lexikální jednotky se v univerzálních pořádacích systémech vyskytují častěji než v oborových selekčních systémech. Důvodem je univerzálnost a rozsáhlost tematiky, frekvence výrazů obecného charakteru/významu, u společenskovědních oborů frekvence adjektiv hodnotící povahy jako součásti víceslovných spojení.

Jako preferované výrazy univerzálního selekčního jazyka mohou být řazena víceslovná spojení, jde-li o

- lexikalizovaná ustálená slovní spojení, která plní terminologickou funkci – termíny
- ustálená slovní spojení typu *černá skříňka, černá díra* – celkový význam tohoto víceslovného spojení se neodvíjí od významu jeho jednotlivých složek
- víceslovnou lexikální jednotku zařazenou na vyšším než posledním hierarchickém stupni (je-li možno utvořit termíny NT – narrower term)

  obrábění kovů NT *frézování*
  fyziologie rostlin NT *fotosyntéza*
  peněžní reforma NT *revalvace*, *devalvace*

- jednoznačnost významu daného slovního spojení, srozumitelnost (*filozofie – dějiny = filozofie dějin* nebo *dějiny filozofie*)
- přesnost znění (*právo průmyslového vlastnictví*)

Při věcném pořádání v jednotlivých oborech, případně ve skupině příbuzných oborů, lze vyváženého vztahu mezi denotátem a lexikální jednotkou dosáhnout snadněji, zatímco v univerzálních pořádacích systémech se velmi často potýkáme s problémem nejednoznačnosti daného termínu. Termín, který je příznačný pro daný obor a je v jeho pojmové soustavě systémově zakotven, může totiž být součástí pojmové/terminologické soustavy dalšího/dalších oborů. V univerzálním pořádacím systému se pak slévá terminologie různých oborů a z dříve jednoznačných termínů systémově zakotvených v jednotlivých oborech se stávají homonyma *(operace, morfologie*).

Proces výběru lexikálních jednotek pro účely zařazení do lexika univerzálního selekčního jazyka je tedy složitý. Je nutné

- posoudit jednoznačnost a systémovost termínu v dominantním oboru
- posoudit jeho případný výskyt v jiných oborech
- porovnat definice v jednotlivých oborech
- posoudit jeho zastupitelnost v nedominantních oborech

V případě neúspěchu je nutno pojednat termín jako homonymum, tj. navrhnout specifikaci pomocí adjektiva, předložkové vazby, případně pomocí rozlišujícího kvalifikátoru (relátoru); např. v případě zmíněné *morfologie* (= nauka o tvarech) slouží tento termín ke zpřístupnění v oboru biologických věd a pro oblast lingvistiky se používá český synonymní výraz *tvarosloví*; mnohovýznamovost termínu *operace* (= provádění, provedení, výkon, úkon) řešíme specifikací, např. *binární operace* (matematika), *bankovní operace* (finančnictví), *operace srdce* (lékařství), *bojové operace* (vojenství) atd.

Harmonizace věcných selekčních jazyků jako předpoklad univerzálního integrovaného nástroje pro pořádání informací

Rozvoj poznání, nutnost pojmenovat složitá interdisciplinární témata a uživatelská nevstřícnost tradičních pořádacích nástrojů vedla ke vzniku a rozvoji specializovaných selekčních jazyků na bázi přirozeného jazyka – oborových tezaurů. Jejich vázanost na národní jazyky, tematické omezení na jednotlivé obory, případně příbuzné skupiny oborů, se v souvislosti s postupnou internacionalizací a globalizací informační scény brzy pociťují jako výrazný limitující faktor. Ztěžují pozici koncového uživatele, protože dochází k atomizaci a dezintegraci věcných selekčních údajů v bibliografických záznamech, je ohrožena i plnohodnotná směnitelnost záznamů na národní a mezinárodní úrovni. Od počátku 70. let jsme svědky renesance univerzálních pořádacích nástrojů, která vyústila v současné aktivity usilující o základní transformaci stávajících věcných pořádacích systémů – předmětových hesel, MDT, DDC, LCC (Library of Congress Classification) v rešeršní nástroje plně odpovídající požadavkům kladeným na organizační nástroje univerzálního typu v současném informačním prostředí (navigace a filtrování, referenční jazyk sloužící jako propojující element při tvorbě vícejazyčného pořádacího systému). Na rozdíl od předchozího pojetí se nyní zkoumá možnost a účelnost propojení obou typů selekčních jazyků (verbálního a systematického) se záměrem získat univerzální systém založený z hlediska uživatele na bázi přirozeného jazyka, který by v sobě spojoval pozitivní vlastnosti obou systémů a omezil ty negativní (omezenost, zastaralost a neřízenost slovního aparátu u systematických selekčních jazyků, neúplnost hierarchických struktur u verbálních selekčních jazyků).

Faktory ovlivňující proces unifikace věcných selekčních jazyků

- vzrůstající počet informačních zdrojů v digitální podobě
- vzrůstající počet klasických dokumentů a jejich selekčních obrazů v databázích sekundárních dokumentů (trend, který nebyl očekáván)
- vzrůstající počet bibliografických záznamů cizí provenience, které mají být v rámci projektu sdílené katalogizace integrovány do databází (pokud možno bez úprav)
- rozvoj virtuálních elektronických knihoven, online dialogových služeb, ve kterých se koncoví uživatelé stále častěji potýkají s problémem, jak vyhledat, tj. identifikovat a lokalizovat informační zdroje ve vícejazyčném prostředí, protože stále větší množství informačních institucí zpřístupňuje jejich prostřednictvím kombinaci sekundárních dokumentů (bibliografických záznamů) a plných textů psaných v různých jazycích; ve snaze umožnit koncovému uživateli i v tomto prostředí vyhledávat v jeho rodném jazyce usilují informační pracovníci o vznik vícejazyčného pořádacího nástroje

Současnou babylonskou změt' věcných selekčních jazyků lze zdánlivě snadno řešit „uzákoněním" používání jednoho přirozeného jazyka (národního) v celém procesu pořádání a vyhledávání informací: preference angličtiny je zřejmá, nereálnost tohoto návrhu je však evidentní.

Mezi dalšími návrhy je nutno zmínit zavedení druhého (paralelně používaného) selekčního jazyka, a to systematického, který tím, že není vázán na verbální národní jazyk, je mezinárodně srozumitelný. Horkým kandidátem je DDC, které se postupně velmi razantně začíná šířit i v evropském prostředí.

Tuto variantu lze poněkud upravit zavedením dvou systematických jazyků – MDT a DDC – a vypracováním bilaterálních konkordančních tabulek. Toto řešení je však do jisté míry nereálné, neboť většina specializovaných institucí nepoužívá univerzální systematický selekční jazyk a dává přednost oborovým tezaurům, příp. specializovaným tříděním.

Obě uvedené varianty se prozatím jeví jako neschůdné; musíme tedy přistoupit k harmonizaci věcných selekčních jazyků.

Harmonizovat lze jazyky, které vykazují společné rysy. Věcné selekční jazyky se však navzájem liší

- oborovým zaměřením, tematickým rozsahem
- specifičností používaných lexikálních jednotek
- strukturou vztahů mezi lexikálními jednotkami
- způsobem organizace lexikálních jednotek v selekčním obraze dokumentu

Existuje několik způsobů harmonizace věcných selekčních jazyků

- konvergence
- kompatibilita
- konvertibilita
- konkordance

Při konkrétní harmonizaci příbuzných selekčních jazyků se uvedené postupy prolínají.

Konvergence selekčních jazyků spočívá v postupném sbližování, ve vyrovnávání rozdílů. Jako projev konvergence lze označit změnu pojetí lexikální jednotky v předmětových a deskriptorových jazycích určených pro vyhledávání v bibliografických bázích. Pod vlivem celostního pojetí termínu v pojmově orientovaných systémech opouštějí tyto jazyky metodu formálního rozkladu termínu a častěji zařazují do svého lexika víceslovné lexikální jednotky.

Plně kompatibilní jsou jazyky, jejichž lexikální jednotky (sémantika i syntax) jsou převoditelné: pro každou lexikální jednotku jednoho jazyka existuje ekvivalentní termín v jazyce druhém. Úplná kompatibilita informačních jazyků na bázi přirozeného (národního) jazyka není možná, protože informační jazyky založené na bázi přirozeného jazyka (uživatelsky vstřícné) jsou závislé na přirozených jazycích, neustále se měnících a vzájemně odlišných systémech. Markantně se tato proměna projevuje právě v lexikální rovině, především v terminologii, která je pro věcné pořádací systémy dominantní. U prekoordinovaných jazyků brání plné kompatibilitě i rozdíly v rovině syntagmatické (provázanost použitých lexikálních jednotek již v procesu indexování).

Velkou míru kompatibility vykazují jazyky typologicky příbuzné, např. prekoordinovaný jazyk předmětových hesel a systematický selekční jazyk MDT. Jde o kompatibilitu bilaterální, která je podmínkou při harmonizaci věcných selekčních jazyků pomocí referenčního jazyka.

Konvertibilita je vlastnost informačních systémů projevující se jako schopnost jednoho informačního systému zpracovat informační výstupy tak, aby byly použitelné jako vstupy v jiném informačním systému.[11)] Je třeba si však uvědomit, že reciproční automatizovaná konverze je možná jen mezi selekčními jazyky stejného typu (postkoordinovaný vs. postkoordinovaný, prekoordinovaný vs. prekoordinovaný), jinak je možná pouze jednosměrná konverze prekoordinovaný jazyk versus postkoordinovaný.

Konkordance spočívá v přiřazení významově shodných lexikálních jednotek porovnávaných selekčních jazyků.

Lexikální harmonizace založená na obsahu pojmu

Předpokladem lexikální harmonizace založené na obsahu pojmu je pojmová kompatibilita lexikálních jednotek selekčních jazyků tematicky oborově zaměřených. Zvolí se pilotní jazyk, tj. jazyk s nejbohatším lexikem, v němž jsou silně zastoupeny prekoordinované lexikální jednotky a který obsahuje přiměřený výčet termínů se značnou mírou specifičnosti a má v žádoucí míře propracovanou strukturu vztahů mezi lexikálními jednotkami. Metoda spočívá ve výběru vhodného preferovaného termínu, v porovnání synonym a kvazisynonym.

Vytváření nové hierarchické struktury je méně náročné, vezmeme-li za základ hierarchickou strukturu jednoho z původních jazyků a dopracujeme ji definováním dalších nutných vztahů.

Přesvědčivým reprezentantem úspěšné lexikální harmonizace založené na obsahu pojmu (integrace více než 60 slovníků, heslářů, klasifikačních systémů) je UMLS (Unified Medical Language System) Metathesaurus.[12)] Lexikální jednotky metatezauru jsou organizovány podle významu: každému pojmu je přiřazen identifikátor pojmu, jsou k němu vztaženy všechny termíny v angličtině opatřené identifikátorem, a dále jsou k tomuto pojmu vztaženy všechny varianty použité v původních zdrojích, z nichž každá je také opatřena specifickým identifikátorem. V rámci metatezauru jsou lexikálním jednotkám přiřazovány základní informace o jejich významu z hlediska integrace a jsou uváděny ve vztah synonymie, hierarchie a jiné vztahy nutné z hlediska makrostruktury (mezi lexikálními jednotkami metatezauru je definováno devět typů vztahů). UMLS Metathesaurus je určen pro vyhledávání informací v bibliografických a v jiných dokumentografických databázích i v plnotextových/faktografických databázích. Ve faktografických bázích je účinnost tohoto nástroje umocněna zapojením pojmově orientovaného lexikálního programu – sémantické sítě (SPECIALIST lexicon). Do lexikonu metatezauru mohou být zahrnuty i lexikální jednotky jazyků, do kterých byl přeložen původní pilotní jazyk tohoto systému – MeSH.

Lexikální harmonizace založená na konkordanci

Předpokladem lexikální harmonizace založené na konkordanci je usouvztažnění ucelených hierarchických struktur (deskriptorových odstavců). Jako příklad nám může sloužit projekt MACS (Multilingual Access to Subject), který se pokouší o propojení tří verbálních autoritních systémů: LCSH (Library of Congress Subject Headings), RAMEAU (Répertoire d´autorité-matière encyclopédique et alphabétique unifié) a SWD (Schagwortnormdatei) definováním linek mezi autoritními termíny. Pilotním jazykem je systém Library of Congress, tj. prekoordinovaný jazyk s bohatým lexikem a nejpropracovanějšími hierarchickými vazbami. Velký problém působí různá míra specifičnosti termínu a rozdílná struktura lexikálních jednotek/deskriptorů těchto tří systémů, jak vyplývá z následující ukázky:

LCSH	SWD	RAMEAU
Cycling accidents	*Radsport + Sportveletzung*	*Cyclistes – Lésions et blessures*
Cycling for women	*Radsport + Frauensport*	*Cyclisme Féminin*

Podmínkou realizace projektu je existence bilaterálních konkordančních tabulek všech zúčastněných subjektů, což je systémově i finančně značně náročné: porovnávací tabulky musejí být obousměrné u všech tří systémů, je tedy nutné vytvořit šest konkordančních tabulek:

- LCSH – RAMEAU : RAMEAU – LCSH
- LCSH – SWD : SWD – LCSH
- RAMEAU – SWD : SWD – RAMEAU

Jak vyplývá z uvedeného, ústřední problém spočívá ve volbě pilotního jazyka, je-li jím selekční systém založený na přirozeném jazyku. Celý proces se zjednoduší, zvolíme-li za pilotní systematický selekční jazyk, který se stane referenčním jazykem. Jednotky přirozeného jazyka se převádějí do znaků systematického jazyka, samotné vyhledávání pak probíhá v systematickém jazyce. Podmínkou je vytvoření konkordančních tabulek přiřazujících lexikální jednotky v přirozeném jazyce znakům systematického selekčního jazyka. Tím se sníží počet potřebných konkordančních tabulek na polovinu. Systém je otevřený, neboť je založen na referenčním jazyce systematickém, tj. nezávislém na přirozených jazycích. Je však třeba poznamenat, že celý proces může být nepříznivě ovlivněn právě probíhajícími změnami v systému MDT, tj. posilování fasetového principu. Řešením je přijetí jednotných mezinárodně platných pravidel pro tvorbu notačních znaků.

Soubor věcných autorit – nástroj integrace na národní úrovni

Soubor věcných autorit (také „SVA") se řadí k věcným selekčním systémům zabývajícím se obsahovou charakteristikou dokumentu. Představuje standard pro věcný popis (národní soubor věcných autorit) tvořený v návaznosti na mezinárodní standardy pro věcné pořádání a vyhledávání informací a aplikovatelný v databázích s rozsáhlým univerzálním fondem. Při obsahové charakteristice dokumentu se klade největší důraz na zpřístupnění tematických informací v dokumentech obsažených, její součástí v širším kontextu jsou však i údaje formální, geografické a chronologické.

Proto se soubor věcných autorit skládá z několika dílčích souborů

- soubor tematických selekčních prvků – je dominantní, základním prvkem jsou tematické lexikální jednotky, mezi kterými jsou vyjádřeny sémantické hierarchické vztahy
- soubor formálních selekčních prvků
- soubor geografických selekčních prvků
- soubor chronologických selekčních prvků

Personální a korporativní jména, v odborné literatuře považovaná za identifikátory, do souboru věcných autorit neřadíme. Jsou součástí souboru autorit jmenných.

Pro zjednodušení problematiky se dále budeme zabývat především souborem tematických lexikálních jednotek.

Soubor věcných autorit je řízený a měnitelný abecedně uspořádaný soubor věcných selekčních údajů – lexikálních jednotek, mezi nimiž jsou vyjádřeny základní sémantické vztahy (ekvivalence, hierarchie, asociace), který je určen ke zpracování a vyhledávání dokumentů a informací v nich obsažených. Součástí autoritního záznamu je notační znak systematického selekčního jazyka související s autoritním záhlavím a nezbytný poznámkový aparát. K základním charakteristikám tohoto souboru patří slovní zásoba, struktura sémantických vztahů a aplikační syntax.

Slovní zásoba (lexikum)

Základním prvkem lexika souboru věcných autorit je lexikální jednotka, která bývá v odborné literatuře definována jako „slovní vyjádření určitého pojmu, pokud možno ve formě substantiva nebo substantivního spojení".[13)]

V souboru věcných autorit existují dva typy lexikálních jednotek

- preferovaný termín – „lexikální jednotka užívaná závazně při indexování k vyjádření určitého pojmu"[14)]
- nepreferovaný termín – „ekvivalent nebo kvaziekvivalent preferovaného termínu; nepreferovaný termín není dokumentu přiřazován, ale slouží jako uživatelský vstup do abecedního rejstříku; uživatel je odkázán příslušným pokynem (např. viz) k ekvivalentnímu preferovanému termínu"[15)]

Tvar lexikálních jednotek

Při výběru lexikálních jednotek se řídíme těmito zásadami

- volí se substantivní tvar, kterým může být
  - jednoslovné substantivum
  - víceslovná lexikální jednotka, jejíž základ tvoří substantivum, a to

  - adjektivní spojení
  - předložková vazba
  - komplexní termín – dvě souřadně spojená substantiva označující komplexní pojem, tj. vztah daných pojmů, např. *rodina a škola*; rozklad tohoto komplexního pojmu na jednotlivé dílčí pojmy by vedl ke vzniku nežádoucího ohlasu irelevantních dokumentů
- dodržuje se přirozený slovosled
- pravopisná forma se řídí platnou normou, v případě existence pravopisných dublet se preferuje progresivní podoba
- transliterace se řídí platnými normami
- singulár a plurál
  - počitatelná substantiva se uvádějí v plurálu
  - abstrakta, počitatelná substantiva použitá jako abstrakta, názvy vědních oborů se uvádějí v singuláru
- cizojazyčné výrazy je možné použít
  - neexistuje-li adekvátní překladový výraz
  - je-li termín běžně používán v daném vědním oboru
- zkratky – preferuje se rozepsaná podoba, zkrácená podoba se odkáže
- synonymie lexikálních jednotek se řeší odkazovým aparátem (odkaz viz)
- homonymie, polysémie (polyvalence lexikálních jednotek)
  - různé významy homonym se důsledně rozlišují
    - specifikací termínu
    - uvedením kvalifikátoru (relátoru) v závorce, závorkové doplnění je součástí deskriptoru

Definice, vymezení rozsahu preferovaného termínu je důležitou součástí záznamu věcné autority univerzálního systému, protože v tomto systému se často vyskytují výrazy, které se běžně používají ve více oborech.

Při tvorbě univerzálního autoritního souboru musí být zvláštní pozornost věnována obsahové stránce tematických lexikálních jednotek a integraci víceslovných spojení do souboru lexikálních jednotek (k této problematice viz výše).

Struktura sémantických vztahů

Mezi lexikálními jednotkami je možno definovat tyto sémantické vztahy

- vztah ekvivalence, který je jedním ze základních předpokladů řízeného slovníku. Do vztahu ekvivalence jsou uváděny synonymní lexikální jednotky, tj. termíny, které se liší formou, ale jejichž obsah je identický (označují stejný denotát). Jeden termín je vybrán jako preferovaný, ostatní se považují za nepreferované termíny a jsou odkázány na termíny preferované;
- vztah hierarchie, který nastává mezi lexikálními jednotkami téhož sémantického okruhu a vyjadřuje poměr nadřazenosti a podřazenosti. Pojem je podřazený druhému pojmu tehdy, jestliže k jeho identifikaci musíme použít všechny znaky nutné k identifikaci nadřazeného pojmu, přičemž podřazený pojem má minimálně o jeden znak, kterým se od sebe liší, méně;
  BT (broader term) *vodní nádrž*
  NT *vodárenská nádrž*
  NT *rekreační nádrž*[16)]
- vztah asociace, který nastává mezi lexikálními jednotkami, které spolu významově souvisejí, avšak jejich vzájemný vztah není možno považovat za hierarchický.

Aplikační syntax

Pravidla pro tvorbu SVA nabízejí dvě možnosti postupu

1. autorizovat řetězce předmětových hesel;
   výhodou postupu je možnost vyjádřit komplexní téma dokumentu výstižně už v prvním plánu, což ve svém důsledku znamená
   - zachovat informační hodnotu řetězce předmětových hesel a
   - umožnit snadnější orientaci jedné skupiny uživatelů

   nevýhodou tohoto postupu je
   - komplikovanost struktury řetězce předmětového hesla, která má za následek
     - nesnadnou orientaci uživatele v případech, kdy se bohatě větvená struktura předmětového hesla nezobrazí (mizí za obrazovkou), případně se zobrazí na jiném řádku
     - nedostupnost těchto složitě strukturovaných záznamů pro většinu vyhledávacích služeb
   - komplikovanost pravidel pro tvorbu předmětového hesla
   - nesnadná a nákladná údržba takto koncipovaného autoritního souboru
   - rozdílnost při formulaci dotazu při vyhledávání v dokumentografických a plnotextových databázích

2. autorizovat jednotlivé segmenty předmětových hesel; výhodou postupu je
   - přehlednost záznamů
   - snadná orientace uživatele
   - jednoduchá pravidla pro aplikaci
   - snadnější budování hierarchických struktur
   - snadná údržba
   - homogenní prostředí pro vyhledávání

Výzkum realizovaný Národní knihovnou v roce 1997 prokázal, že v oblasti věcných selekčních jazyků naprosto převládá používání izolovaných lexikálních jednotek, tj. klíčových slov, nejčastěji tzv. volně tvořených, tedy ve své podstatě neřízených předmětových termínů, a dále deskriptorů oborových, případně polytematických tezaurů a heslářů. Prekoordinovaná předmětová hesla s rozvinutou strukturou jsou v naprosté menšině a navíc se nejčastěji používají pouze jako alternativní selekční jazyk.[17)] Znamená to tedy, že při věcném pořádání je dominantní princip postkoordinace. Za této situace jsme dospěli k závěru,

že z výše uvedených variant postupu v oblasti aplikační syntaxe autoritních prvků bude více odpovídat našemu prostředí varianta druhá, tedy autorizace jednotlivých prvků předmětového hesla.

Při konkrétní aplikaci souboru věcných autorit, tj. při ukládání věcných údajů v bibliografickém záznamu ve výměnném formátu UNIMARC, volíme kombinovaný postup (princip prekoordinace a postkoordinace se kříží):

- tematické selekční prvky se ukládají ve vstupních pozicích – princip postkoordinace
- geografické a chronologické okolnosti tématu, pokud nejsou předmětem dokumentu, se zapisují v příslušném podpoli – princip prekoordinace

Při tvorbě souboru věcných autorit respektujeme tyto standardy a mezinárodní doporučení

- doporučení IFLA[18] – „To consider possible relationships between subject authority records and classification."
- UNIMARC/AUTORITY – formát, který definuje strukturu zápisu záznamu věcné autority:
  - blok 2-- autoritní záhlaví/unifikované záhlaví
  - blok 3-- poznámky
  - blok 4-- variantní záhlaví – odkaz typu viz
  - blok 5-- příbuzné/související záhlaví – odkazy viz též: termíny BT, NT, RT
  - blok 6-- klasifikační znak
  - blok 7-- propojovací záhlaví
  - blok 8-- poznámky o konzultovaných zdrojích

Funkce notačního znaku MDT v záznamu věcné autority

Základní funkcí notačního symbolu v univerzálním systému, kterou můžeme pracovním názvem označit jako fixaci termínu, je určení dominantního oboru, posouzení jeho pozice v univerzu lidského poznání, posouzení jednoznačnosti a systémovosti daného termínu v pojmové struktuře oboru/oborů. Neméně důležitou funkcí je podpora při tvorbě hierarchické struktury univerzálního souboru, podíl na mezinárodní srozumitelnosti selekčního jazyka. Jako referenční jazyk napomáhá při tvorbě vícejazyčného nástroje. V neposlední řadě, především ve fázi vzniku autoritního souboru, slouží jako kontrola kvality přidělovaných údajů.

Funkce souboru věcných autorit

Soubor věcných autorit představuje národní standard pro věcné pořádání a vyhledávání informací, který respektuje mezinárodní standardy a je určen zejména pro databáze s rozsáhlým univerzálním fondem. Soubor věcných autorit by měl sloužit zejména jako

- nástroj standardizace a unifikace věcných selekčních údajů na národní úrovni
- nástroj pro sdílenou katalogizaci
- nástroj integrace a unifikace věcných selekčních jazyků různých typů
- podpora vzniku vícejazyčného pořádacího systému
- uživatelsky vstřícný rešeršní nástroj pokrývající heterogenní prostředí

Předpokládáme, že vybudování souboru věcných autorit přispěje k

- unifikaci prostředí tolik potřebné pro budování předmětově orientovaných zdrojů
- znásobení rešeršní možnosti koncového uživatele: zpřesněním a dopracováním dotazu se zvýší přesnost a úplnost vyhledaných dokumentů

Rešeršní strategie a nástroje

Rešeršní strategie je postup, jak efektivním způsobem získat relevantní dokumenty v reálném čase. Předpokládá přesnou formulaci dotazu, analýzu tématu a znalost vyhledávacích služeb. Mezi rešeršní nástroje řadíme věcné selekční jazyky a systémové nástroje podmíněné informačními technologiemi.

Podle stupně dokonalosti těchto rešeršních prostředků se systémy dělí do tří kategorií, tzv. generací. Pro 1. generaci je charakteristické jednoduché vyhledávání slov, jednoduché maskování, pravostranné rozšíření. 2. generace je charakteristická používáním booleovských a proximitních operátorů, maskováním, pravo-levostranným rozšiřováním slov, vyhledáváním podle pole, ostrým vyhodnocováním dotazů. 3. generace je zaměřena na pojmově orientované vyhledávání, rozklad pojmu na podpojmy, vážení pojmů, neostré vyhodnocování dotazů.[19] Tyto systémové rešeršní nástroje, tedy prostředky informačních technologií, ovlivňovaly rešeršní strategie, které měly zásadní vliv na vývoj věcných selekčních jazyků (viz dramatický odklon od jazyků prekoordinovaného typu v souvislosti s nástupem systémů druhé generace).

Lze předpokládat, že plný rozvoj systémů třetí generace (pojmově orientovaných) a nástup systémů umělé inteligence, případně systémů založených na porozumění přirozenému jazyku, vyvolá neméně dramatickou, ne-li ještě dramatičtější proměnu této oblasti.

Kvalita věcného pořádání informací a jejich zpřístupnění je ovlivněna hodnotícími algoritmy při stanovení relevance dokumentů.[20]

Současné vyhledávací služby pracují na podkladě statisticko-pravděpodobnostních metod. Bylo již řečeno, že tyto vyhledávací nástroje nejsou schopny porozumět významu slov, nejsou schopny posoudit význam slova v kontextu. Nejsou také schopny uspokojivě řešit problémy vyvolané synonymií, homonymií a polysémií termínů vyskytujících se v textu.

Při stanovení relevance dokumentů používají algoritmy založené

- na výskytu slov; vyhledané dokumenty jsou uspořádány podle počtu shodných slov vyskytujících se v uživatelově dotazu a ve vyhledaném dokumentu, přičemž výše jsou hodnoceny dokumenty, které obsahují větší počet výskytu jednotlivých slov dotazu
- na počtu odkazů (hyperlinků); dokumenty jsou hodnoceny podle počtu směrovaných odkazů; dokumenty s větším počtem odkazů jsou hodnoceny výše

Kvalitu věcného pořádání informací a jejich zpřístupnění ovlivňuje i způsob prezentace věcných údajů.

Při posuzování tohoto jevu vycházíme z předpokladu, že

- „koncový uživatel, člověk v komunikaci s informačním systémem či informační službou přímou či zprostředkovanou informační institucí nebo informačním specialistou, je středem celého problému, kolem kterého se odehrávají procesy spojené s vyhledáváním informací v informačních a počítačových systémech“ [21)]
- čtení z obrazovky počítačů je o 27 % pomalejší než z papíru
- koncový uživatel je schopen vnímat a identifikovat údaje v rejstřících nacházejících se na prvních 2-3 pozicích[22)]
- údaje prezentované v dalších pozicích jsou pro mnohé elektronické katalogy nezobrazitelné („mizí za obrazovkou”), případně se zobrazují na druhé řádce

K těmto zjištěním musíme přihlížet při rozhodování o zásadách aplikační syntaxe, podle kterých jsou ukládány věcné údaje v bibliografických záznamech a v návaznosti na to v přístupových rejstřících. V online prostředí se jeví jako užitečné původní bohatě strukturované řetězce prekoordinovaných předmětových hesel výrazně zkrátit a ukládat selekční prvky do vstupních pozic.

V prostředí tištěných výstupů se zdůrazňovala maximální informační hodnota řetězce předmětového hesla. Vycházelo se tím vstříc aktuálním potřebám koncového uživatele, který pomocí jednoho vstupu do systému mohl snadno a přesně identifikovat i lokalizovat hledanou informaci. Informační hodnota řetězce byla považována za uživatelsky vstřícnou.

V online prostředí je nutno tuto zásadu přehodnotit: koncový uživatel je stále více konfrontován s existencí hypertextových odkazů i v situacích, kdy je neočekává a které kladou větší nároky na jeho pozornost: prohlížení, listování (browsing). Dlouhá, bohatě strukturovaná informace v této situaci ještě ztěžuje vnímání, a je proto považována za uživatelsky nevstřícnou.

Závěr

Současná problematika věcného zpracování je úzce svázána s problematikou řízeného zpřístupňování webovských informačních zdrojů, zhodnocení a proměny informací ve znalosti, tj. informace přinášející užitek. Z hlediska koncového uživatele je neméně důležitá i vhodná prezentace vyhledaných informací.

Ukazuje se, že řízený přístup k informacím lze vybudovat vhodnou integrací tradičních pořádacích systémů, které také mohou napomoci koncovému uživateli při definování adekvátních rešeršních postupů. Transformace těchto tradičních nástrojů v naznačených oblastech je však nevyhnutelná. Vytvořením jednotného přístupu k informacím v současném heterogenním prostředí, které je charakteristické prolínáním různých typů informačních zdrojů (vyhledávání v databázích primárních i sekundárních dokumentů), přispějeme k uživatelsky vstřícnému prostředí.

Poznámky:

1) KOVÁŘ, Blahoslav. *Problémy teorie procesu věcného pořádání informací a selekčních jazyků*. Praha : Univerzita Karlova, 1976. 165 s.
2) Tamtéž, s. 18
3) JONÁK, Zdeněk. Inteligentní nástroje pro práci s texty na internetu. *Ikaros* [online]. 1998, č. 9. Dostupný z: <http://ikaros.ff.cuni.cz/1998/c09/nastroje.htm>
4) JONÁK, Zdeněk. Omezení a možnosti zvýšení selekčních schopností internetových robotů. *Daidalos* [online]. 2001, č. 1/2. Dostupný z: <http://daidalos.ff.cuni.cz/2001/leden/zj_sj.php>
5) KOVÁŘ, Blahoslav. *Problémy teorie procesu věcného pořádání informací a selekčních jazyků*. Praha : Univerzita Karlova, 1976, s. 39.
6) JONÁK, Z. Omezení a možnosti zvýšení ...
7) VORÁČEK, Josef. *Tvorba tezaurů v českém jazyce.* Praha : ÚVTEI – STK, 1974, s. 25.
8) STOFFA, Ján. *Terminológia v technickej výchove.* Olomouc : UP-PedF, 2000, s. 19.
9) SLAVÍČKOVÁ, Eleonora. Problematika víceslovných spojení v tezaurech. In *Lingvistické metody a automatizované informační systémy*. Praha : Dům techniky ČSVTS, 1988, s. 12-20.
10) JONÁK, Z. Omezení a možnosti zvýšení ...
11) SCHWARZ, Josef. *Vývoj teorie a praxe tezaurů v České republice*. Praha : FFUK-ÚISK, 1999, s. 12.
12) Viz např.: UMLS National Library of Medicine. (February 2000). Fact Sheet: UMLS (r) Metathesaurus (r). Dostupný z: <http://www.nlm.nih.gov./pubs/factsheets/umlsmeta.html>
13) *ČSN 01 0193 Dokumentace. Pokyny pro vypracování a rozvíjení jednojazyčných tezaurů* (1996). Účin. leden 1996. 49 s.; citace převzata z: SCHWARZ, Josef. *Vývoj teorie a praxe tezaurů v České republice.* Praha : FFUK-ÚISK, 1999, s. 15.
14) Tamtéž

15) Tamtéž
16) BAKO, Michal. *Informačné selekčné jazyky/III.* Bratislava : Slovenské pedagogické nakladatelstvo, 1984, s. 110.
17) MDT 64,3 %, klíčová slova aj. 79,1 %, tezaurus 14,5 %, předmětová hesla 24,4 %.
18) *Guidelines for Subject Authority and Reference Entries*. München : K. G. Saur, 1993.
19) ČERVENÝ, Vlastimil. Vyhledávání v databázích plných textů. *Národní knihovna*. 1999, roč. 10, č. 1, s. 6-12.
20) JONÁK, Z. Omezení a možnosti zvýšení …
21) PAPÍK, Richard. Vyhledávání informací II : uživatelské rozhraní a vlivy oboru „human-computer interaction". *Národní knihovna*, 2001, roč. 12, č. 2, s. 81.
22) DRABENSTOTT, Karen Markey and VIZINE-GOETZ, Diane. *Using Subject Headings for Online Retrieval : Theory, Practice, and Potential*. San Diego : Academic Press, 1994. 365 p.

Další literatura:

Subject indexing : principles and practices in the 90´s : Proceedings of the IFLA Satellite Meeting held in Lisbon, Portugal, 17-18 Aug. 1993. München : Saur, 1995. 302 p.

Vizualizing Subject Access for 21st Century Information Resources. Urbana-Champaign : University of Illinois, 1998. 176 p.

LANDRY, Patrice. *The MACS Project : Multilingual Access to Subjects (LCSH, RAMEAU, SWD)*. Classification and Indexing Workshop, 66th IFLA Council and General Conference, Meeting No 181. Dostupný z: <http://www.ifla.org/IV/ifla66/papers/165-181e.pdf>

PhDr. Marie Balíková je vedoucí oddělení věcného zpracování odboru zpracování fondů Národní knihovny ČR.