Praktické aspekty hodnocení kvality a konzistence indexace

Josef Schwarz (schwarz@psp.cz), Parlamentní knihovna

Pozn. Článek je upravenou verzí příspěvku, který byl přednesen na semináři Pořádání znalostí 2000 dne 18.12.2000.

Obsah

Úvod

Používáte ve vaší knihovně nebo informačním středisku pro věcné zpracování dokumentů tezaurus? Chcete vědět, zda vaši indexátoři pracují kvalitně? Jak zajistit, aby indexované dokumenty byly zpracovány kvalitně? Existuje rozdíl mezi kvalitou a konzistencí indexace? Co to vlastně je ta kvalita a konzistence indexace a jak se projevuje při vyhledávání dokumentů?

Pokud jste na první dvě otázky odpověděli kladně a zbylé vás zaujaly, pak je tento článek určen právě vám. Článek má blízko k případovým studiím indexačních experimentů, nicméně se v něm o typické indexační experimenty nejedná, protože zdrojová data jednotlivých analýz byla získána během rutinní práce při indexaci dokumentů a nejedná se tudíž o systemizovaný vzorek, který by byl předem vybrán, zhodnocen a zpracován podle metodických požadavků na tyto typy úloh.[1] Původním účelem sběru a hodnocení dat bylo připravit podklady pro rozhodování o způsobu a metodách indexace v konkrétním informačním systému. Předložené výsledky je tedy nutno interpretovat a zobecňovat se zřetelem na tuto skutečnost.

Text je rozdělen na dvě základní části, teoretická část obsahuje stručný úvod do problematiky kvality a konzistence indexace[2], praktická část shrnuje výsledky jednotlivých analýz, které byly zaměřeny na kontrolu kvality indexace, indexační chyby, hodnocení indexátorů, konzistenci indexátorů a konzistenci dokumentů. Teoretická část není v žádném případě vyčerpávající studií na dané téma, jejím cílem je především objasnit pojmy a přiblížit základní metodická a teoretická východiska, která jsou základem pro praktickou část. Účely analýz byly čistě praktické a nikoliv teoretické, nicméně některé výsledky potvrzují dosavadní výzkum a v některých směrech jej i rozšiřují, zejména v oblasti metodologie a typologie dané problematiky.

Před vlastním vstupem do problematiky je třeba zodpovědět základní otázku: proč vlastně hodnotit a kontrolovat kvalitu a konzistenci indexace? Důvod je jednoduchý: aby byla zajištěna kvalita vyhledávání, resp. aby byly vytvořeny podklady pro její kontrolu. Kvalita a konzistence indexace patří mezi faktory, které podstatně ovlivňují výsledek vyhledávání. Zejména se to týká relevance vyhledaných záznamů měřené pomocí indexů úplnosti (recall) a přesnosti (precision), které bezprostředně s kvalitou a konzistencí indexace souvisejí (jejich vztah je popsán dále v textu).

[1] Požadavky na zdrojová data pro indexační experimenty a jejich obecnou metodiku podrobně rozvádí např. MAIXNER, V. Metodika indexačních experimentů. *Knižnice a vedecké informácie*, 1984, roč. 16, č. 2, s. 62-65.

[2] Teoretická část vychází zejména z publikace LANCASTER, F.W. *Indexing and abstracting in theory and practice*. London : Library Association Press, 1998. Další prameny jsou citovány v jednotlivých poznámkách.

Teoretická část

Co je kvalita indexace

Při hodnocení kvality indexace je třeba vycházet z toho, že kvalita indexace není absolutní, nezávislou hodnotou, ale je třeba ji posuzovat v kontextu celého informačního systému, a to zejména z hlediska jeho základní funkce, kterou je poskytování relevantních informací.

Kvalita indexace je také výrazně určena účelem informačního systému (např. tím, jakým uživatelům slouží) a dostupnými prostředky, které jsou v rámci informačního systému dostupné zejména pro věcné zpracování a vyhledávání informací (indexátoři, selekční jazyk, indexační pravidla, rešeršní systém atd.).

Obecně a velmi zjednodušeně lze říci, že kvalitní indexace je taková indexace, která zajistí, aby dokument byl vyhledán tehdy, když má být vyhledán, a nebyl vyhledán, když vyhledán být nemá. Kvalitní indexace je tedy indexace, která zajistí maximální relevanci výsledků vyhledávání.

Pokud bychom chtěli kvalitu indexace definovat přesněji, lze říci, že je to míra shody obsahu selekčního obrazu[3] s obsahem dokumentu a zároveň s obsahem uživatelského, resp. rešeršního dotazu, přičemž pojem „obsah“ je třeba chápat s ohledem na komprimaci textu, ke které při indexaci a vyhledávání dochází.

Z výše uvedeného vyplývá, že kvalita indexace je relativní charakteristika, jejíž hodnotu nelze obecně určit (tzn. nelze obecně stanovit „správný“ popis konkrétního dokumentu) a jenž se musí řídit potřebami konkrétního informačního systému, v rámci kterého by měla být stanovena pravidla, jak kvalitní indexace dosáhnout (indexační pravidla). Kromě těchto pravidel ovlivňuje kvalitu indexace řada dalších faktorů.

Faktory ovlivňující kvalitu indexace

Indexace je komplexní proces, v rámci kterého indexátor určitým způsobem přiřazuje termíny selekčního jazyka k dokumentu tak, aby vystihl jeho obsah. Všechny aktivní i pasivní složky tohoto procesu, tj. indexátor, selekční jazyk, dokument, pravidla pro zpracování a pracovní podmínky, mohou různým způsobem ovlivnit kvalitu indexace.

Indexátor

Osoba indexátora ovlivňuje kvalitu indexace nejpodstatněji, děje se tak prostřednictvím objektivních i subjektivních charakteristik indexátora. Mezi objektivní předpoklady kvality indexace patří zejména zkušenost s věcným zpracováním dokumentů včetně znalosti konkrétního selekčního jazyka, znalost oboru, schopnost porozumět textu konkrétního dokumentu po obsahové i jazykové stránce, schopnost systematického, logického a analytického myšlení a také schopnost efektivní práce s dokumentem (metody racionálního čtení, využití pomocného aparátu dokumentu apod.). Mezi subjektivní charakteristiky lze zařadit např. schopnost soustředění, pozornost, pečlivost, systematičnost, ale i náladu, únavu, pracovní motivaci apod.

Jednotlivé faktory mají z hlediska kvality indexace jinou váhu, paradoxní např. je prakticky získaný poznatek, že znalost oboru je pro kvalitu indexace méně rozhodující než znalost konkrétního selekčního jazyka.

[3] Selekční obraz je množina termínů selekčního jazyka, která byla přiřazena ke konkrétnímu dokumentu

Selekční jazyk

Faktory související se selekčním jazykem jsou také podstatným činitelem kvality indexace. Rozhoduje zejména rozsah a složení slovníku lexikálních jednotek a jeho struktura (řešení homonymie a synonymie, vztahy hierarchie a asociace)[4]. Velice podstatným aspektem je, aby indexátor dokázal určit správný význam lexikální jednotky, který může být určen jednak strukturou, do které je daná lexikální jednotka zařazena, jednak poznámkou o rozsahu nebo obsahu dané lexikální jednotky. Součástí selekčního jazyka jsou i pravidla pro správný výběr lexikálních jednotek, jejich kombinaci apod., tedy pravidla, která se označují jako gramatika selekčního jazyka a která by měla být součástí indexačních pravidel. Pravidla pro konkrétní lexikální jednotky lze upřesnit formou indexační poznámky uvedenou přímo u dané lexikální jednotky.

Dokument

Kvalita indexace je nesporně ovlivněna i předmětem indexace, tedy dokumentem, a to faktory související s jeho obsahem nebo formou. Z hlediska obsahu je pro kvalitu indexace nejpodstatnější obor, do kterého dokument patří, a tematická struktura dokumentu. Téma dokumentu může být strukturováno na hlavní a dílčí témata, jež mohou být dále členěna s různou složitostí a komplexností. Z formálních náležitostí dokumentu, resp. textu, rozhoduje jeho délka, forma zpracování (použitá stylistika, terminologie atd.), dostupnost pomocného aparátu (obsah, rejstřík apod.) a v neposlední řadě také jazyk dokumentu.

Pravidla pro zpracování

Pravidla pro věcné zpracování dokumentů jsou určujícím faktorem kvality. Pod pravidly pro zpracování rozumíme především způsob, resp. typ věcného zpracování dokumentů, tedy to, zda se jedná o indexaci nebo klasifikaci, zda se používá spíše postupů prekoordinace nebo postkoordinace apod. Konkrétní pravidla by měla být explicitně vymezena a zachycena v indexačních pravidlech nebo jiném obdobném materiálu. Praktické zkušenosti ukazují, že vágní a pouze „ústní tradicí“ předávaná pravidla podstatně snižují kvalitu a konzistenci indexace.

Pracovní podmínky

Indexace vyžaduje vhodné pracovní podmínky a pracovní prostředí. Na kvalitu práce indexátora může mít podle jeho osobních dispozic vliv např. pracovní doba, kdy indexace probíhá (ráno, poledne, večer), požadovaná produktivita práce (při tlaku na velkou produktivitu práce – velký počet indexovaných záznamů za hodinu – kvalita indexace klesá), ale i pracovní prostředí s danou úrovní světla, hluku apod. Nezanedbatelné jsou také technické prostředky, které má indexátor pro indexaci dokumentů k dispozici.

[4] Ke kvalitě selekčního jazyka a jeho analýze stručně a podnětně JONÁK, Z. Omezení a možnosti zvýšení selekčních schopností internetových robotů. *Daidalos* [online], 2001, č. 1. Dostupné z <http://daidalos.ff.cuni.cz/2001/leden/zj_sj.php>.

Hodnocení kvality indexace

Jakým způsobem lze kvalitu indexace hodnotit? Než odpovíme na uvedenou otázku, je třeba poukázat na to, že se zde setkáváme s určitými metodologickými problémy, které spočívají především v obtížné kvantifikovatelnosti kvality indexace a potažmo i jejího hodnocení. Před hodnocením indexace je proto vhodné stanovit kvalitativní kritéria, podle kterých se budou indexované záznamy posuzovat, a ta se posléze pokusit kvantifikovat.

Pro hodnocení kvality indexace lze použít následující metody:

Kontrola indexace

Jedná se o metodu hodnocení kvality indexace, která spočívá v přímé kontrole obsahové i formální správnosti selekčních obrazů dokumentů. Rozhodnutí o správnosti nebo nesprávnosti selekčního obrazu se provádí na základě definovaných indexačních pravidel, míru a charakter správnosti selekčního obrazu lze vyjádřit prostřednictvím indexačních chyb. Kontrola indexace může probíhat bezprostředně po provedení indexace nebo jako součást hodnocení indexace jinými metodami. V každém případě je třeba uvést, že kontrola indexace je časově a personálně značně náročná, protože předpokládá opakování nebo znovuprovedení celého procesu indexace (obsahová analýza, identifikace pojmů, výběr lexikálních jednotek selekčního jazyka) s tím, že kontrolu indexace musí samozřejmě provádět nezávislý indexátor nebo skupina indexátorů. Pokud je však kontrola indexace prováděna bezprostředně po indexaci, resp. před definitivním zařazením indexovaného záznamu do databáze, je pro kvalitu indexace a vyhledávání nesporným, výrazným a okamžitým přínosem.[5]

V samých základech této metody je ukryta závažná námitka: ten, kdo provádí kontrolu indexace, je také „pouze“ indexátor stejně jako ten, kdo dokument původně indexoval, a tudíž se stejně jako on může dopouštět indexačních chyb. Tento argument je jistě oprávněný, je však třeba podotknout, že kontrola indexace by měla být svěřena zkušenému indexátorovi, který je pokročilý v ovládání selekčního jazyka i obsahové analýze, podrobně zná indexační pravidla a je tudíž schopen chybovost indexace minimalizovat. V rámci možností je pak zřejmě ideálním řešením, pokud kontrolu indexace provádí skupina zkušených indexátorů, kteří se ve výsledku shodnou na nejadekvátnějším selekčním obrazu dokumentu, jenž pak může být indexačním vzorem i pro další obdobné dokumenty.[6]

Kromě kontroly jednotlivých indexovaných záznamů je také možná komplexní kontrola indexace, která je založena na analýze množin dokumentů, které jsou popsány stejným nebo podobným selekčním obrazem. Nejprve se vyhodnotí význam selekčního obrazu a pak se stanovuje, nakolik mu jednotlivé dokumenty skutečně odpovídají.

Problematika kontroly indexace včetně indexačních chyb je podrobněji rozpracována v praktické části na základě konkrétních dat a výsledků.

Hodnocení relevance

Při hodnocení kvality indexace je možné použít metodu, která vychází ze základního účelu indexace: vyhledávání dokumentů. Tato metoda je založena na vztahu mezi kvalitou indexace a relevancí výsledku vyhledávání. Předpokládá se, že existuje závislost mezi charakteristikami

[5] Dokladem toho jsou i snahy o automatizaci těchto procesů pomocí expertního systému, viz TODESCHINI, C., FARRELL, M.P. An expert system for quality control in bibliographic databases. *Journal of the American Society for Information Science*, 1989, roč. 40, č. 1, s. 1-11.

[6] Pozoruhodný experiment byl proveden na záznamech databáze Chemical Abstracts, kdy kvalitu indexace záznamů posuzovali autoři dokumentů s tím, že 80% záznamů bylo z jejich hlediska indexováno kvalitně. Blíže viz BRAAM, R.R., BRUIL, J. Quality of indexing information : authors' views on indexing of their articles in Chemical Abstracts online CA-file. *Journal of Information Science,* 1992, roč. 18, č. 5, s. 399-408.

jednotlivých veličin, tzn. mezi úplností (exhaustivity) a specifičností (specifity) indexace na jedné straně a úplností (recall) a přesností (precision) výsledku vyhledávání na straně druhé.[7]

Úplnost indexace je míra, která vyjadřuje, nakolik je obsah dokumentu popsán termíny selekčního jazyka, jinak řečeno, nakolik jsou pro indexaci zachycena hlavní a dílčí témata popř. pojmy obsažené v dokumentu. Úplnost indexace je malá, pokud jsou indexována pouze vybraná témata, a naopak je maximální, pokud jsou indexována všechna témata (popř. pojmy) identifikovatelná v dokumentu.

Specifičnost indexace souvisí s obecným pravidlem, které říká, že pro indexaci určitého pojmu je třeba použít co nejspecifičtější termín ze slovníku selekčního jazyka. Specifičnost indexace je tedy příliš nízká, pokud byl vzhledem k obsahu dokumentu vybrán příliš obecný termín, a příliš vysoká, pokud byl vzhledem k obsahu dokumentu vybrán příliš specifický termín.

Pro úplnost připomeňme, že úplnost výsledku vyhledávání (recall) je poměr mezi nalezenými relevantními dokumenty a všemi relevantními dokumenty v databázi a přesnost výsledku vyhledávání (precision) je poměr mezi nalezenými relevantními dokumenty a mezi všemi nalezenými dokumenty.

Úplnost a přesnost indexace musí obecně stanovovat indexační pravidla, v každém případě nejsou žádoucí extrémní hodnoty obou charakteristik, které negativně ovlivňují přesnost i úplnost při vyhledávání. V případě, že byl dokument indexován s přílišnou úplností, je při vyhledávání negativně ovlivněna přesnost. Dokument indexovaný s příliš malou úplností snižuje úplnost při vyhledávání. Nízká nebo naopak příliš vysoká specifičnost může negativně ovlivnit jak úplnost, tak i přesnost vyhledávání.

Tyto závislosti jsou vcelku zřejmé, hůře se to má s jejich kvantifikací. Zatímco indexy úplnosti a přesnosti se vypočtou podle standardních vzorců, pro hodnoty úplnosti a specifičnosti indexace neexistuje obecně uznávaná kvantifikace. Další problém je v tom, že úplnost a přesnost se vztahuje k množině vyhledaných dokumentů, kdežto úplnost a specifičnost indexace k jednotlivých indexovaným záznamům. Proto ani není možné matematicky definovat vztah mezi indexy pro kvalitu indexace a indexy pro relevanci výsledku vyhledávání.

Přestože vztahy uvedených charakteristik nelze kvantifikovat, lze je využít např. pro analýzu indexace dokumentů, které byly v rámci konkrétní rešerše vyhledány chybně, tj. jsou z hlediska daného informačního dotazu nerelevantní. Uvedené charakteristiky byly také použity pro kategorizaci indexačních chyb, která je popsána dále v textu v praktické části. Vlastní příklad analýzy indexace na základě hodnocení relevance vyhledaných dokumentů však v praktické části chybí. Tato analýza vychází z výsledků rešeršních experimentů a je náročná zejména z toho důvodu, že je třeba určitými postupy v databázi dohledat dokumenty, které jsou relevantní, ale které nebyly vyhledány. Teprve poté je lze analyzovat z hlediska kvality a konzistence indexace. Analýza dokumentů, které byly nalezeny, ale nejsou relevantní, je podstatně jednodušší.

Třetí metodou pro hodnocení kvality indexace je analýza konzistence, která je podrobněji pojednána v následující podkapitole.

[7] Jedná se o zjednodušený model, který lze rozpracovat podstatně podrobněji. Viz Soergel, D. Indexing and retrieval performace : the logical evidence. *Journal of the American Society for Information Science*, 1994, roč. 45, č. 8, s. 589-599.

Konzistence indexace

Co je konzistence indexace

Konzistence indexace je míra shody dvou nebo více selekčních obrazů dokumentů. Číselná hodnota konzistence dvou selekčních obrazů se vypočítá jako poměr počtu souhlasných termínů k celkovému počtu termínů obsažených v obou selekčních obrazech:

$$C=ab/(a+b)$$

kde
C=index konzistence indexace
ab=počet souhlasných termínů v selekčních obrazech A a B
a=počet jedinečných termínů v selekčním obraze A
b=počet jedinečných termínů v selekčním obraze B

Index konzistence může nabývat maximální hodnoty 1 a minimální hodnoty 0, všechny ostatní hodnoty se pohybují v intervalu (0;1). Index konzistence o hodnotě 1 znamená, že dva selekční obrazy jsou totožné, index konzistence o hodnotě 0 znamená, že selekční obrazy nemají žádný společný termín. První případ označíme jako absolutní konzistenci, v druhém případě budeme mluvit o nulové konzistenci. Případy z intervalu (0;1) budeme označovat jako částečnou konzistenci.

Praktický význam konzistence je nasnadě: v databázi je třeba zajistit, aby obsahově totožné nebo podobné dokumenty byly indexovány stejně nebo přibližně stejně. Pokud tomu tak není, dochází při vyhledávání dokumentů ke snížení celkové relevance výsledků vyhledávání.

Typy konzistence indexace

Konzistenci indexace lze rozdělit do dvou typů: konzistence indexátorů a konzistence dokumentů. Všeobecně se v literatuře uvádí pouze konzistence indexátorů, protože konzistence dokumentů pramení z konzistence indexátorů, nicméně zde je konzistence dokumentů popsána z důvodu dalšího rozboru v praktické části.

Konzistence indexátorů[8] se nejčastěji měří mezi dvěma indexátory, kteří indexovali totožný dokument (tzv. konzistence mezi indexátory – interindexer consistency), i když je možné stanovit i konzistenci jednoho indexátora, který indexoval daný dokument v určitém časovém odstupu (tzv. konzistence indexátora – intraindexer consistency). Přestože se teoretické výzkumy zaměřují především na konzistenci mezi indexátory, v praxi má častokrát podstatně větší význam konzistence indexátora z důvodu omezeného počtu indexátorů.

Konzistenci dokumentů lze opět rozdělit na konzistenci mezi dokumenty (srovnání selekčních obrazů dokumentů, které pojednávají o stejném tématu) a konzistenci dokumentu (srovnání selekčních obrazů, které se vztahují k jednomu dokumentu, např. v případě různých vydání nebo jazykových mutací jednoho díla). Na rozdíl od obou typů konzistence indexátorů (konzistence mezi indexátory a konzistence indexátora), které charakterizují odlišné veličiny, se v případě konzistence dokumentu jedná o speciální případ konzistence mezi dokumenty.

Z hlediska správy databáze je primární dosáhnout konzistence dokumentů, která může být mj. zajištěna konzistencí indexátorů.

[8] Konzistence indexátorů je častým předmětem indexačních experimentů. Zahraniční studie shrnuje LANCASTER, op.cit., s. 86-89, z českých zdrojů uveďme např. HRAŠOVÁ, J., JANÁKOVÁ, I. Faktory ovlivňující shodu indexátorů (při indexování dokumentů klíčovými slovy). Československá informatika, 1986, roč. 28, č. 9, s. 250-254 nebo SMETÁČEK, V., KÖNIGOVÁ, M., SODOMKOVÁ, J. Pilotážní průzkum indexátorů. *Československá informatika,* 1974, č. 5, s. 268-277.

Konzistence versus kvalita indexace

Jak souvisí konzistence a kvalita indexace? Není konzistence vlastně také kvalita indexace?

Na druhou otázku je třeba jednoznačně odpovědět, že nikoliv. Množina obsahově podobných dokumentů může být indexována konzistentně, tzn. že všechny dokumenty budou indexovány stejně, ale to ještě neznamená, že jsou tyto dokumenty indexovány z hlediska obsahu správně, tj. kvalitně.

Vztah konzistence a kvality indexace není přímočarý[9], nicméně lze říct, že konzistence indexace pozitivně ovlivňuje kvalitu indexace v tom smyslu, že zlepšuje efektivitu vyhledávání. Z hlediska správy databáze je totiž konzistence nespornou kvalitou: podstatně snáze lze opravit nesprávně, ale konzistentně indexované záznamy, než záznamy indexované nesprávně a zároveň inkonzistentně.

Příklady dvou analýz konzistence indexace se nacházejí v praktické části.

[9] Praktickou metodologii hodnocení kvality indexace založenou na konzistenci indexátorů kombinovanou s auditem kvality rozvíjí STUBBS, E.A., MANGIATERRA, N. E., MARTÍNEZ, A.M. Internal quality audit of indexing : a new application of interindexer consistency. *Cataloging & Classification Quarterly*, 1999, roč. 28, č. 4, s. 53-69.

Praktická část

Praktická část je rozdělena na tři podkapitoly, které obsahují analýzy konkrétních dat. Data pro tyto analýzy pocházejí z knihovny Evangelické teologické fakulty Univerzity Karlovy (UK-ETF – první a třetí analýza) a z Parlamentní knihovny (PK – druhá analýza). O kvalitě a původu dat jsem se již zmínil stručně v úvodu, podrobněji jsou data popsána v jednotlivých podkapitolách.

Kontrola indexace, indexační chyby a hodnocení indexátorů (UK-ETF)

Účel a metodika kontroly indexace

Knihovna UK-ETF od r. 1996 používá pro indexaci dokumentů Český teologický tezaurus, jehož první verze vznikla v r. 1995. V r. 1996 probíhala zkušební indexace na omezeném vzorku dokumentů, následující rok byla zahájena indexace v poloprovozu a od r. 1998 se dokumenty indexují v rutinním provozu.

Kontrola indexace byla systematicky prováděna v letech 1997-1999 a jejím hlavním smyslem bylo zajistit kvalitu a konzistenci indexace. Kontrola indexace byla prováděna následně po indexaci dokumentu indexátorem, prováděl ji tým, který tezaurus vytvořil a implementoval do informačního systému knihovny UK-ETF. Hlavními důvody, proč byla kontrola indexace prováděna, byly zejména tyto skutečnosti:

- indexace byla prováděna pomocí nového selekčního jazyka, se kterým neměli indexátoři zpočátku žádné zkušenosti
- věcné zpracování dokumentů bylo v knihovně UK-ETF novinkou, protože před zavedením tezauru se dokumenty v knihovně systematicky věcně nezpracovávaly
- indexátoři nebyli kmenoví zaměstnanci knihovny, ale externí spolupracovníci, kteří pocházeli zejména z řad studentů UK-ETF
- v počátečních obdobích docházelo k relativně značné fluktuaci indexátorů, která by mohla mít nepříznivý vliv na kvalitu indexace
- v relativně krátkém období byl indexován velký objem dokumentů, u nichž bylo potřeba zajistit kvalitní indexaci, protože pozdější opravy nebo změny by byly komplikované
- bylo potřeba získat zpětnou vazbu nejen pro indexátory, ale i pro další rozvoj tezauru a indexačních pravidel.

Kontrola indexace prakticky probíhala tak, že indexátorem indexované dokumenty byly indexátorem-kontrolorem znovu nezávisle věcně zpracovány a případné rozdíly v selekčních obrazech dokumentů byly ze strany indexátora považovány za chyby. Tyto chyby byly v záznamech opraveny a zapsány do souboru chyb, který sloužil jako podklad pro zpětnou vazbu k indexátorovi a pro další zpracování, jehož výsledkem bylo komplexní hodnocení indexátorů (viz níže).

Indexační chyby

Podrobná typologie indexačních chyb byla předpokladem pro kontrolu indexace. Indexační chyby byly zpočátku formulovány pragmaticky a až později byla provedena jejich kategorizace a analýza. První výsledky byly publikovány již v r. 1998[10], níže uvedené analýzy jsou provedeny na podstatně větším objemu dat z let 1998-1999.

Přehled č. 1. **Typologie indexačních chyb - popis jednotlivých indexačních chyb**

A1a Nesprávná obsahová analýza - chybné pochopení obsahu dokumentu, které má za následek přiřazení deskriptorů, které nevyjadřují obsah dokumentu a nepřiřazení deskriptorů, které obsah dokumentu vyjadřují

A1b Povrchní indexace - použití malého počtu deskriptorů s obecným významem namísto několika specifických deskriptorů, neproběhla identifikace všech indexačních hledisek

A1c Přílišná hloubka indexace - použití velkého množství specifických deskriptorů tam, kde lze použít několik deskriptorů s obecnějším významem

A1d Dokument se neindexuje - indexace dokumentu, který je na základě indexačních pravidel z indexace vyloučen

A2a Opominutí hledisek - zanedbání hledisek indexace, které je třeba určit při obsahové analýze a na základě kterých je třeba vybrat příslušné deskriptory

A2b Indexace okrajových hledisek - indexace hledisek, které nejsou z hlediska vyhledávání a uživatele podstatné a lze je při indexaci zanedbat

A2c Nesprávné uplatnění hledisek - indexace hledisek, která byla z hlediska obsahu dokumentu chybně identifikována

A2d Záměna oboru a jeho předmětu - použití deskriptoru označujícího vědní obor nebo odvětví namísto specifického deskriptoru z dané vědní oblasti nebo naopak

A2e Chybná postkoordinace - syntaktický nebo sémantický rozklad termínu, který je v tezauru vyjádřen jediným deskriptorem

A2f Chybná prekoordinace - označení dvou sémanticky oddělených pojmů jedním deskriptorem

A2g Chybný překlad z cizího jazyka - nesprávný překlad cizojazyčného termínu do češtiny a v důsledku toho výběr chybného deskriptoru z tezauru

A2h Indexace okrajových témat - indexace témat, které nejsou z hlediska vyhledávání a uživatele podstatné a lze je při indexaci zanedbat

A3a Opominutí typu dokumentu - zanedbání hlediska typu dokumentu nebo textu, které je třeba určit při obsahové analýze a na základě kterého je třeba vybrat příslušné deskriptory

A3b Chybné použití typu dokumentu - indexace hlediska typu dokumentu nebo textu, které bylo z hlediska obsahu dokumentu chybně identifikováno

A3c Indexace typu dokumentu jako jediného hlediska indexace - indexace hlediska typu dokumentu nebo textu s tím, že pro indexaci nejsou použity žádné další deskriptory

A3d Záměna obsahu a typu dokumentu - záměna věcného obsahu a formálního typu dokumentu

A4a Nesprávné použití širšího termínu pro užší pojem - indexace deskriptorem se širším významem v případě, kdy je vhodné nebo nutné použít specifičtější deskriptor

A4b Nesprávné stanovení rozsahu nebo významu deskriptoru - chybné určení významu nebo rozsahu deskriptoru

A4c Specifický identifikátor jako jediný indexační výraz - použití vlastního jména (osoba, geopolitický celek) jako jediného hlediska indexace

A4d Nesprávné použití užšího termínu pro širší pojem indexace deskriptorem s užším významem v případě, kdy je vhodné nebo nutné použít obecnější deskriptor

A4e Nadbytečné termíny - pohyb po jedné hierarchické větvi - použití nadbytečných deskriptorů z různých úrovní jedné hierarchické struktury, aniž by k tomu opravňoval specifický obsah dokumentu

[10] Viz SCHWARZ, J. Český teologický tezaurus : nové věcné třídění pro teologii, filozofii a religionistiku. *Národní knihovna*, 1998, roč. 9, č. 3, s. 114-115.

A4f Nadbytečné jednotlivé deskriptory - indexace jednotlivých pojmů, které nejsou z hlediska vyhledávání a uživatele podstatné a lze je při indexaci zanedbat

A4g Chybějící jednotlivé deskriptory - opominutí jednotlivých pojmů při obsahové analýze, které jsou podstatné z hlediska obsahu dokumentu nebo z uživatelského hlediska

B1 Záměna deskriptoru za nedeskriptor - formální chyba: použití nedeskriptoru místo deskriptoru při indexaci

B2 Duplikace stávajících deskriptorů - formální chyba: vložení deskriptoru do tezauru, který již existuje

Tabulka č. 1 **Podíl jednotlivých chyb na celkové chybovosti**

Kód	Chyba	počet	podíl [%]
A2a	Opominutí hledisek	441	18,6
A4b	Nesprávné stanovení rozsahu nebo významu deskriptoru	291	12,3
A4g	Chybějící jednotlivé deskriptory	272	11,5
A3a	Opominutí typu dokumentu	251	10,6
A1a	Nesprávná obsahová analýza	166	7
A3b	Chybné použití typu dokumentu	147	6,2
A4f	Nadbytečné jednotlivé deskriptory	131	5,5
A4a	Nesprávné použití širšího termínu pro užší pojem	114	4,8
A1b	Povrchní indexace	98	4,1
A1d	Dokument se neindexuje	60	2,5
A2b	Indexace okrajových hledisek	56	2,4
A2c	Nesprávné uplatnění hledisek	58	2,4
A1c	Přílišná hloubka indexace	55	2,3
A2d	Záměna oboru a jeho předmětu	50	2,1
A4e	Nadbytečné termíny - pohyb po jedné hierarchické větvi	32	1,4
A2e	Chybná postkoordinace	31	1,3
A3d	Záměna obsahu a typu dokumentu	25	1,1
A2h	Indexace okrajových témat	18	0,8
A4d	Nesprávné použití užšího termínu pro širší pojem	17	0,7
A4c	Specifický identifikátor jako jediný indexační výraz	15	0,6
A2g	Chybný překlad z cizího jazyka	11	0,5
B1	Záměna deskriptoru za nedeskriptor	12	0,5
B2	Duplikace stávajících deskriptorů	9	0,4
A2f	Chybná prekoordinace	8	0,3
A3c	Indexace typu dokumentu jako jediného hlediska indexace	1	0
	CELKEM	2369	100,0

Uvedený přehled a tabulka jsou poněkud nepřehledné, proto je třeba chyby dále kategorizovat, nicméně již nyní je nutno poukázat na jednu důležitou informaci: první tři nejčastější chyby (tj. A2a, A4b a A4g) jsou nejen shodné s nejfrekventovanějšími chybami v citované analýze z r. 1998, ale dokonce se s nimi téměř shodují i v podílu na celkové chybovosti, který u těchto tří chyb tvoří nyní společně 42,4%, v r. 1998 to bylo cca 42,7% všech chyb.

Lze tedy konstatovat, že pro indexátory byla nejproblematičtějším okruhem jejich práce analýza hledisek dokumentu, stanovení významu jednotlivých deskriptorů a systematický výběr deskriptorů z tezauru. Zda má toto tvrzení obecnou platnost by bylo třeba ověřit provedením podobné analýzy i na jiných informačních pracovištích.

Pro smysluplnou interpretaci indexačních chyb a jejich využití pro hodnocení kvality indexace je třeba provést kategorizaci chyb, a to jednak podle jejich závažnosti, jednak podle fáze jejich vzniku. Rozdělení chyb do jednotlivých kategorií může být diskutabilní vzhledem k tomu, že nebylo provedeno na základě kvantifikovatelných a zcela objektivních rysů, nicméně pro základní kategorizaci chyb a jejich interpretaci bude tento přístup dostačovat.

Kritéria závažnosti chyby byla stanovena dvě:

1) míra, ve které chyba ovlivňuje selekční obraz dokumentu, který jí může být negativně poznamenán částečně nebo komplexně (dílčí nebo komplexní chyby)
2) skutečnost, zda chyba při vyhledávání ovlivňuje index úplnosti a/nebo přesnosti.

Komplexní chyba je závažnější než chyba dílčí a chyba, která negativně ovlivňuje oba indexy, je závažnější než chyba snižující nebo ovlivňující pouze jeden index. Na základě kombinace těchto dvou kritérií bylo stanoveno 8 kategorií chyb podle závažnosti.

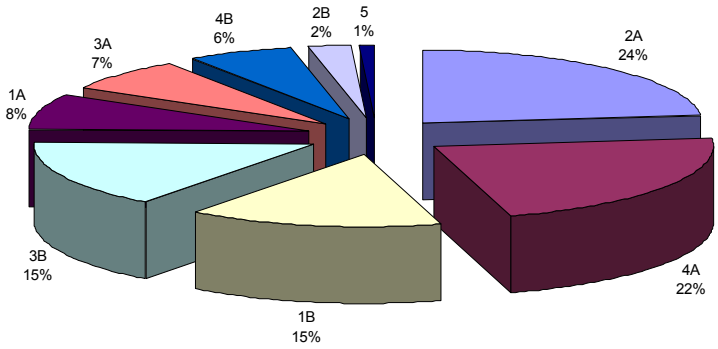
Přehled č. 2 **Kategorie chyb podle závažnosti**

Pozn.: Chyby jsou uspořádány podle závažnosti, nejzávažnější jsou komplexní chyby snižující přesnost i úplnost, nejméně závažné jsou dílčí chyby snižující přesnost nebo úplnost. Chyby, které úplnost a přesnost *ovlivňují*, mají specifický charakter. Formální chyby jsou zcela samostatnou kategorií chyb.

1A Komplexní chyby snižující přesnost i úplnost: A1a, A3d
1B Dílčí chyby snižující přesnost i úplnost:A2d, A2g, A4b
2A Komplexní chyby snižující úplnost:A2a, A1b, A3c, A4c
2B Komplexní chyby snižující přesnost:A2c
3A Komplexní chyby ovlivňující přesnost i úplnost:A1c, A1d, A2b
3B Dílčí chyby ovlivňující přesnost i úplnost:A2h, A4e, A4f, A4d, A2e, A2f, A4a
4A Dílčí chyby snižující úplnost:A3a, A4g
4B Dílčí chyby snižující přesnost:A3b
5 Formální chyby:B1, B2

Tabulka č. 2 **Podíl jednotlivých kategorií chyb podle závažnosti na celkové chybovosti**

Kód	Kategorie chyby	Počet	Podíl [%]
1A	Komplexní chyby snižující přesnost i úplnost	191	8,1
1B	Dílčí chyby snižující přesnost i úplnost	352	14,9
2A	Komplexní chyby snižující úplnost	555	23,4
2B	Komplexní chyby snižující přesnost	58	2,4
3A	Komplexní chyby ovlivňující přesnost i úplnost	171	7,2
3B	Dílčí chyby ovlivňující přesnost i úplnost	351	14,8
4A	Dílčí chyby snižující úplnost	523	22,1
4B	Dílčí chyby snižující přesnost	147	6,2
5	Formální chyby	21	0,9
	CELKEM	2369	100



Graf č. 1 **Podíl jednotlivých kategorií chyb podle závažnosti na celkové chybovosti**

Tabulka č. 2 je uspořádána podle závažnosti jednotlivých kategorií chyb, graf č. 1 podle podílu na celkové chybovosti. Z uvedeného vyplývá, že nejzávažnější kategorie chyb představuje pouze cca 8 % všech chyb, na druhou stranu, další dvě nejzávažnější kategorie chyb tvoří společně téměř 40% všech chyb.

Druhá kategorizace chyb vychází z toho, v jaké fázi indexace chyby vznikají, což umožňuje sledovat nejen to, v jakých indexačních krocích vzniká nejvíce chyb, ale i to, odkud potenciálně pramení.

Přehled č. 3 **Kategorie chyb podle fáze indexace**

I Obsahová analýza (potenciální zdroje chyb: indexátor)
Ia Analýza obsahu dokumentu: A1a,A1b, A1c, A3d, A1d, A2d
Ib Analýza typu dokumentu a textu: A3c, A3a, A3b

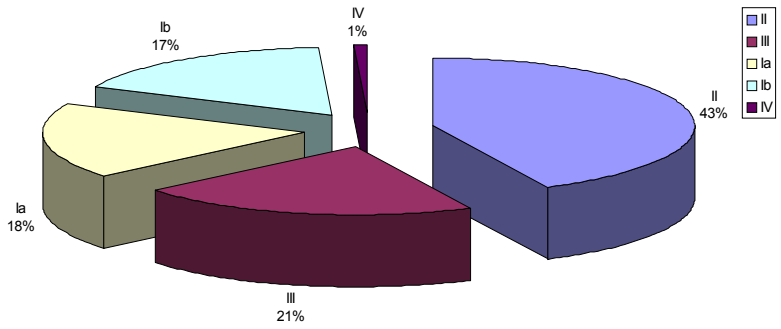
II Identifikace pojmů (potenciální zdroje chyb: indexátor, indexační pravidla): A2b, A2a, A4c, *A2c,* A2g, A2h, A4f, A4g

III Výběr deskriptorů tezauru (potenciální zdroje chyb: indexátor, indexační pravidla, selekční jazyk):A4b, A4e, A4d, A2e, A2f, A4a

Tabulka č. 3 **Podíl jednotlivých kategorií chyb podle fáze indexace na celkové chybovosti**

Kód	Kategorie chyby alfa	Počet	Podíl [%]
Ia	Obsahová analýza - analýza obsahu dokumentu	429	18,1
Ib	Obsahová analýza - analýza typu dokumentu a textu	399	16,8
II	Identifikace pojmů	1002	42,3
III	Výběr deskriptorů tezauru	493	20,8
IV	Formální chyby	21	0,9
	CELKEM	2369	100

Graf č. 2 **Podíl jednotlivých kategorií chyb podle fáze indexace na celkové chybovosti**



Tabulka č. 3 je uspořádána podle indexačního postupu, graf č. 2 podle podílu na celkové chybovosti. Základní krok indexace, obsahová analýza, se na celkové chybovosti podílí necelými 20% (kategorie Ib je specifická kategorie, která je nutno uvažovat odděleně). Největší problémy (41% všech chyb) činila indexátorům identifikace pojmů, která sice částečně patří do obsahové analýzy, ale indexační pravidla ji ovlivňují do větší míry než obsahovou analýzu, proto je vedena odděleně. Je ovšem nutno podotknout, že největší podíl na této kategorii má chyba opominutí hledisek (44%), která se jako jedna z nejvýznamnějších objevila již při prostém frekvenčním seřazení chyb. Tento výsledek potvrzuje to, že indexátorům činilo potíže vícehlediskové uvažování.

Chyby, které vznikly při výběru deskriptorů z tezauru, představují pouze 20% všech chyb. Po podrobnější analýze, která však nebyla provedena, by bylo možné doložit, které chyby způsobil indexátor tím, že se nedostatečně orientoval ve struktuře a lexiku tezauru, a které chyby byly způsobeny nedostatečností tezauru.

Hodnocení indexátorů

Jak již bylo uvedeno výše, prostřednictvím indexačních chyb byla průběžně hodnocena práce indexátorů. Níže jsou uvedeny výsledky komplexního hodnocení indexátorů sledovaných v období od dubna 1998 do července 1999.

Tabulka č. 4 podává souhrnný statistický přehled práce indexátorů se zaměřením na kvalitu jejich indexace. Pro jednotlivé indexátory lze z tabulky zjistit:

- počet celkem zpracovaných dokumentů a podíl chybně indexovaných dokumentů (sloupce C-E tabulky)
- počet dokumentů zpracovaných na začátku činnosti každého indexátora a z nich opět podíl chybně indexovaných dokumentů (sloupce F-H tabulky)
- všeobecnou úroveň úplnosti jejich indexace (sloupec I tabulky)
- celkový počet chyb (sloupce J a K tabulky)
- relativní kvalitu indexace (sloupec L tabulky)

Podrobný komentář významu jednotlivých údajů obsažených v tabulce následuje pod tabulkou.

Tabulka č.4 **Hodnocení indexátorů**

A	B	C	D	E	F	G	H	I	J	K	L
(a)	4.4.-13.7.1998	280	1809	15,5	76	99	76,8	1,4	327	13,8	0,45
(b)	3.4.1998-27.5.1999	189	620	30,5	83	123	67,5	3,22	219	9,2	0,89
(c)	5.10.1998-13.7.1999	687	1835	37,4	38	78	48,7	1,19	823	34,7	1,13
(d)	6.4.-9.9.1998	13	33	39,4	13	33	39,4	1,5	16	0,7	1,22
(e)	19.1.-1.7.1999	146	337	43,3	42	103	40,8	1,7	181	7,6	1,35
(f)	5.4.-14.9.1998	222	478	46,4	42	100	42	0,87	273	11,5	1,43
(g)	2.5.1999-14.6.1999	33	73	45,2	33	73	45,2	1,5	44	1,9	1,51
(h)	6.4.1998-13.7.1999	329	677	48,6	56	97	57,7	0,66	455	19,2	1,69
(i)	3.2.-11.5.1999	29	86	33,7	29	86	33,7	2	31	1,3	2,68
Celk.	3.4.1998-13.7.1999	1928	5948	32,4	412	792	52	1,56	2369	100,0	---

Přehled č. 4 **Komentář k tabulce hodnocení indexátorů**

A – Jméno indexátora (nahrazeno písmenem abecedy)
B – Období, kdy byla prováděna kontrola indexace
C – Celkový počet chybně indexovaných záznamů
D – Celkový počet indexovaných záznamů
E - Podíl chybně indexovaných záznamů [%]
F - Počet chybně indexovaných záznamů zpracovaných na počátku činnosti indexátora (pro prvních cca 100 záznamů)
G - Počet záznamů zpracovaných na počátku činnosti indexátora (pro prvních cca 100 záznamů)
H - Podíl chybně indexovaných záznamů zpracovaných na počátku činnosti indexátora (pro prvních cca 100 záznamů) [%] (srovnáním H a E lze získat údaj, o kolik procent se indexátor zlepšil od začátku své činnosti do konce sledovaného období; tento údaj není vypovídající v případech, kdy bylo indexováno pouze malé množství záznamů (indexátoři d, g a i). U některých indexátorů (e a f) vykazuje počáteční indexace menší chybovost než celkový soubor sledovaných dokumentů, tito indexátoři z hlediska kvality indexace patřili skutečně k těm problémovějším.
I - Index změny selekčního obrazu - je vypočten z podílu celkového počtu deskriptorů přidaných (p) a odebraných (o) při kontrole indexace (I=p/o). Index nevyjadřuje žádnou absolutní hodnotu, vypovídá pouze o tom, zda se indexátor dopouštěl spíše chyb, které vedly k přílišné úplnosti indexace (index nižší než 1) nebo spíše chyb, které vedly k nedostatečné úplnosti indexace (index vyšší než 1). Většina

indexátorů spíše měla tendenci k nedostatečné úplnosti indexace, pouze dva z nich (f a h) přistupovali k úplnosti indexace opačně

J – Celkový počet chyb (nekoresponduje s počtem chybně indexovaných záznamů (C) vzhledem k tomu, že v jednom chybně indexovaném záznamu mohlo být identifikováno více typů chyb)

K – Podíl chyb na jednoho indexátora [%]

L- Index chybovosti indexátora - proporcionální podíl chyb na jednoho indexátora (L=(J*Dcelkem)/(Jcelkem*D); čím vyšší hodnota, tím vyšší měl indexátor proporcionální podíl na vzniku chyb vzhledem k celkovému počtu indexovaných záznamů v databázi a celkovému počtu chyb ve sledovaném vzorku. Číslo tak vyjadřuje relativní kvalitu indexace jednotlivých indexátorů – čím vyšší hodnota, tím byla práce indexátora méně kvalitní.

Kromě hodnot pro jednotlivé indexátory lze z tabulky vyčíst i celkovou chybovost sledovaného souboru, která činí téměř 33%, tj. cca 1/3 záznamů. Tento podíl potvrzuje původní předpoklad potřebnosti kontroly indexace z důvodů, které byly uvedeny výše. Od sběru dat uběhlo již více než 1,5 roku, v jehož průběhu již kontrola indexace nebyla z důvodů časových a personálních prováděna systematicky, data proto nemají adekvátní vypovídací hodnotu a nejsou zde proto již zpracována. Protože v tomto období došlo zejména ke stabilizaci indexátorů, kteří nyní patří mezi kmenové zaměstnance knihovny, a plnohodnotnému zavedení a standardnímu používání tezauru pro indexaci a vyhledávání dokumentů, lze důvodně předpokládat, že celková kvalita indexace je v současnosti podstatně vyšší i přesto, že neprobíhá systematická kontrola indexovaných záznamů.

V současnosti probíhá revize Českého teologického tezauru, jejímž výsledkem bude nová verze tezauru obsahující řadu změn v oblasti struktury i lexika, rovněž budou modifikována i indexační pravidla. Po implementaci nové verze tezauru a indexačních pravidel bude žádoucí opět na přechodné období provádět systematickou kontrolu indexace, aby byla zajištěna kvalita indexace.

Poté, co se indexátoři adaptují na změněné pracovní nástroje a kvalita indexace bude dosahovat žádoucí úrovně, lze od systematické kontroly upustit a provádět ji pouze namátkově.

Konzistence indexátorů (PK)

Účel a metodika analýzy konzistence indexátorů

Parlamentní knihovna používá pro indexaci dokumentů již od r. 1993 tezaurus Eurovoc. Až do roku 1999 se používal výhradně pro indexaci dokumentů obsažených v knihovním katalogu, tzn. především knižních dokumentů. V r. 1999 se začaly pomocí tezauru Eurovoc indexovat i petice zaslané do Poslanecké sněmovny občany, občanskými sdruženími a dalšími subjekty a od r. 2000 probíhá testovací indexace sněmovních tisků, mezi které patří především návrhy zákonů a novel zákonů, písemné poslanecké interpelace, návrhy státních rozpočtů, výroční zprávy či účetní uzávěrky organizací, které ze zákona tyto dokumenty předkládají sněmovně ke schválení, bilaterální a multilaterální mezinárodní smlouvy, s nimiž vyslovuje sněmovna souhlas, a další typy dokumentů.

V souvislosti s touto rozšířenou implementací Eurovocu v informačním systému kanceláře Poslanecké sněmovny Parlamentu ČR je třeba zabezpečit, aby:

- při indexaci jednotlivých typů dokumentů byla zohledňována jejich specifika
- všichni indexátoři používali stejná indexační pravidla
- dokumenty byly indexovány stejně kvalitně nezávisle na jejich typu.

Pro uskutečnění těchto cílů zde existovala jedna zásadní překážka: přestože se Eurovoc používal pro indexaci již od r. 1993, nebyla formulována indexační pravidla, která byla doposud předávána výše zmíněnou „ústní tradicí“. Je třeba podotknout, že vzhledem k omezenému počtu

indexátorů a jejich relativní stabilitě by bylo možné předpokládat, že tento stav nemá podstatnější vliv na kvalitu a konzistenci indexace, nicméně vzhledem k závažnosti situace a případným negativním důsledkům při rozšířené implementaci Eurovocu bylo rozhodnuto tuto hypotézu ověřit.

V rámci testovací indexace sněmovních tisků byl tedy vybrán vzorek 51 dokumentů, který byl pro počáteční indikaci míry konzistence a kvality indexace shledán jako dostačující a který byl předložen k nezávislé indexaci třem indexátorům. Indexace jednotlivých dokumentů byla rozdělena na majoritní a minoritní indexaci, resp. majoritní a minoritní deskriptory, které měly vyjadřovat hlavní a vedlejší témata dokumentu. Na základě tohoto rozdělení bylo možné sledovat primární a sekundární konzistenci indexátorů, nicméně analýza výsledků prokázala, že celková konzistence (všechny deskriptory) a primární konzistence (majoritní deskriptory) jsou téměř totožné. Proto bylo u výsledků analýz uvedených níže od tohoto rozdělení upuštěno a uvádějí se data pro celkovou indexaci. Již v průběhu indexace bylo také zjištěno, že rozdělení indexace na majoritní a minoritní deskriptory je vzhledem k obsahu dokumentu značně problematické, proto se s využitím této techniky pro další testovací indexaci a později pro rutinní indexaci nepočítá.

Pro vyhodnocení shody indexátorů byla použita standardní metodika zjišťování konzistence mezi indexátory (viz teoretická část), výsledky analýz jsou uvedeny v následující části.

Výsledek analýzy konzistence indexátorů

Indexátoři A, B a C použili pro indexaci následující celkové počty deskriptorů (v závorce uveden průměrný počet deskriptorů na jeden záznam):

A: 202 (4,0)
B: 122 (2,4)
C: 168 (3,3)

Nejvyšší úplnost indexace měl tedy indexátor A, nejnižší průměrné úplnosti indexace dosahoval indexátor B.

Přestože byl průměrný počet deskriptorů na jeden záznam u jednotlivých indexátorů poměrně nízký (a bylo by tedy možné předpokládat vysokou nebo dokonce absolutní konzistenci), počet dokumentů indexovaných zcela shodně všemi indexátory je minimální. Na druhou stranu, počet dokumentů, které byly indexovány všemi indexátory zcela rozdílně (tj. indexátoři nevybrali ani jeden společný deskriptor), je také nízký. Konkrétní údaje jsou uvedeny níže s tím, že se nejedná o standardní míry konzistence indexace, které se vztahují vždy pouze ke dvěma indexátorům a které následují dále v textu.

Celková konzistence vzorku záznamů (všichni indexátoři, v závorce uvedeny procentní podíly z celkového počtu dokumentů):

Celkový počet dokumentů:	51 (100)
Počet dokumentů s absolutní konzistencí:	1 (2,0)
Počet dokumentů s částečnou konzistencí:	46 (90,2)
Počet dokumentů s nulovou konzistencí:	4 (7,8)

Průměrná konzistence indexace jednotlivých dvojic indexátorů byla následující:

konzistence indexátorů A a B: 27,6 %
konzistence indexátorů A a C: 32,6 %
konzistence indexátorů B a C: 25,0 %

Ve všech třech případech tedy nedosáhla celková konzistence záznamů ani jedné třetiny, přičemž indexátoři A a C se shodovali nejvíce a indexátoři B a C nejméně. Pokud celkovou

konzistenci rozdělíme podle konzistence jednotlivých záznamů a rozškálujeme (viz tab. č. 5), zjistíme, že vyšší než dvoutřetinovou konzistenci mělo jen minimum záznamů, podobně jako absolutní konzistenci, a že větší část dokumentů má ve dvou případech menší než třetinovou konzistenci a v jednom případě patří do této kategorie přibližně polovina dokumentů.

Tabulka č. 5 **Škály konzistence jednotlivých záznamů**

	počet záznamů			podíl záznamů [%]		
indexátoři/škála	A-B	A-C	B-C	A-B	A-C	B-C
0	8	4	6	15,7	7,8	11,8
(0-33>	23	20	24	45,1	39,2	47,1
(33-66>	17	20	15	33,3	39,2	29,4
(66-100>	2	4	2	3,9	7,8	3,9
100	1	1	4	2	2	7,8

Jak vyplývá z výše uvedených analýz, konzistence indexátorů byla ve zkoumaném vzorku dokumentů značně nízká. Částečně to mohlo být způsobeno nedostatečností vzorku z hlediska kvantitativního nebo kvalitativního, nicméně z větší části výsledky analýz indikují značné nedostatky nejen v konzistenci, ale i v kvalitě indexace, která byla sledována během zpracování záznamů. Zde nejsou k dispozici číselné údaje, nicméně všeobecně lze říci, že nejnižší kvalitu indexace měl vzorek indexátora B, který vykazoval nejnižší úplnost indexace (nejnižší průměrný počet deskriptorů na jeden záznam), a jehož index konzistence byl se zbylými dvěma indexátory A i C vždy nižší než index konzistence těchto dvou indexátorů. V tomto konkrétním případě byl tedy doložen pozitivní vztah mezi konzistencí a kvalitou indexace, který však není platný obecně (viz teoretická část).

Praktické důsledky popsaného experimentu jsou nasnadě. Byla potvrzena nezbytnost explicitně formulovaných indexačních pravidel (v současnosti – leden 2001 – je již připravena jejich pracovní verze) a zároveň potřebnost systematické kontroly indexace, která by byla prováděna po nezbytně nutnou dobu na všech pracovištích, která budou indexovat dokumenty pomocí deskriptorů tezauru Eurovoc. Výsledkem analýzy byla i jednotlivá dílčí pravidla pro indexaci, která byla formulována při zpracovávání dokumentů zařazených do testovacího vzorku. Na základě nových indexačních pravidel bude třeba také zkontrolovat a případně opravit (reindexovat) doposud indexované dokumenty, jejichž selekční obraz nebude indexačním pravidlům odpovídat nebo nebude konzistentní vůči ostatním podobným dokumentům.

Konzistence dokumentů (UK-ETF)

Účel a metodika analýzy konzistence dokumentů

V souvislosti s ukončením systematické kontroly indexace v knihovně UK-ETF, která byla popsána v jedné z předcházejících podkapitol, bylo žádoucí ověřit, jak se kontrola indexace a aplikace indexačních pravidel projevuje na konzistenci selekčních obrazů dokumentů. Teoreticky by měly být všechny obsahově podobné dokumenty indexovány stejně nebo podobně, ale prakticky tomu tak být nemuselo, protože:

- kontrolu indexace prováděl tým kontrolorů, kteří se také navzájem kontrovali a konzultovali dílčí problémy, nicméně nebylo možné zajistit stoprocentní konzistenci jejich práce
- systematická kontrola indexace byla prováděna v období 1997 až červenec 1999, před a po tomto období prováděna nebyla nebo byla prováděna pouze namátková kontrola indexace

- v průběhu uvedeného období vznikla nová verze tezauru a indexačních pravidel (duben 1998).

To byly tedy důvody pro provedení analýzy konzistence indexovaných dokumentů, pro kterou byl vybrán vzorek dokumentů na základě různých vydání stejných dokumentů. Tato množina dokumentů byla zvolena vzhledem ke snadnosti výběru z databáze; pro analýzu připadaly do úvahy ještě množiny dokumentů vybrané např. na základě různých jazykových mutací stejných dokumentů nebo na základě shlukování do tématických skupin, které však bylo možné vybrat z databáze podstatně obtížněji a od jejich výběru bylo proto upuštěno. Vzorek obsahoval celkem 397 párů dokumentů.

Abychom se vyvarovali dezinterpretace výsledků analýz, je třeba blíže popsat datovou základnu vzorku a její strukturu. Z hlediska toho, jak byly záznamy dokumentů indexovány, obsahoval vzorek dva typy záznamů:

a) záznamy, které byly indexovány nezávisle na záznamech jiného vydání téhož dokumentu (tj. dva indexátoři zpracovali nezávisle na sobě dvě různá vydání téhož díla nebo jeden indexátor zpracoval dvě různá vydání téhož díla, aniž by navzájem konzultoval jejich selekční obrazy)
b) záznamy, které byly z hlediska indexace replikovány na základě záznamů jiného vydání téhož dokumentu (k tomu docházelo např. při údržbě databáze, kdy byly dva různé selekční obrazy dvou různých vydání téhož dokumentu upraveny do totožné podoby nebo při indexaci, kdy byl selekční obraz záznamu vytvořen na základě záznamu předcházejícího vydání dokumentu).

Tuto skutečnost je třeba mít na mysli zejména při interpretaci podílu celkového počtu inkonzistentních záznamů ve vzorku, který je ovlivněn tím, že selekční obrazy některých záznamů byly před výběrem vzorku sjednoceny. Záznamy výše uvedených dvou typů bohužel nebylo možné při výběru odlišit, lze pouze na základě znalosti indexačních praktik kvalifikovaně odhadnout, že podíl inkonzistentních záznamů by byl vyšší zhruba o 10%.

Z hlediska typů záznamů, které byly do vzorku zařazeny, je třeba upozornit na velkou podmnožinu záznamů (cca 1/3 všech záznamů), které byly indexovány specifickým způsobem a mohly by tudíž také ovlivnit výslednou konzistenci. Vzorek byl proto rozdělen na dvě samostatné části, z nichž jedna obsahovala všechny záznamy a druhá pouze zmíněné specifické záznamy. V průběhu analýzy se však ukázalo, že rozdíly ve výsledcích mezi oběma vzorky nejsou natolik podstatné, aby je bylo nutno dále zpracovávat a prezentovat odděleně.

Při analýze vzorku byla dále brána do úvahy důležitost jednotlivých deskriptorů pro selekční obraz jako celek, aby bylo možno posoudit závažnost inkonzistence. Při hodnocení konzistence proto bylo provedeno rozdělení na majoritní a minoritní deskriptory, od nichž se pak odvíjela primární a sekundární konzistence, přičemž primární konzistence je závažnější než sekundární konzistence a záznamy s nulovou konzistencí se řadí mezi záznamy primárně inkonzistentní.

Výsledek analýzy konzistence dokumentů

Celková konzistence vzorku záznamů (páry dokumentů, v závorce uvedeny procentní podíly z celkového počtu párů dokumentů):

Celkový počet párů dokumentů:	397 (100)
Počet párů s absolutní konzistencí:	251 (63,2)
Počet párů s částečnou konzistencí:	135 (34,0)
Počet párů s nulovou konzistencí:	11 (2,8)

Průměrná konzistence vzorku, tzn. součet konzistencí jednotlivých párů dělený počtem záznamů, je 78 %.

Páry dokumentů s částečnou nebo nulovou konzistencí budeme dále analyzovat z hlediska primární a sekundární konzistence a pro jednoduchost je budeme dále označovat jako inkonzistentní páry.

Tabulka č. 6 **Konzistence párů dokumentů podle závažnosti**

Sloupec A obsahuje podíl z celkového počtu inkonzistentních párů dokumentů
Sloupec B obsahuje podíl z celkového počtu všech párů dokumentů

	Počet	A	B
Inkonzistentní páry celkem	146	100,0	36,8
Páry s primární inkonzistencí	48	32,9	12,1
Páry se sekundární inkonzistencí	98	67,1	24,7
Průměrná konzistence inkonzistentních párů celkem [%]	40,0	---	---

Pokud shrneme výsledky analýzy konzistence indexace, zjistíme, že průměrná konzistence vzorku je téměř 80%, přičemž z celkového počtu zkoumaných párů je inkonzistentně indexováno cca 37%, ale pouze u 9% z celkového počtu párů se jedná o primární, tj. závažnou inkonzistenci a 3% párů jsou indexovány s nulovou konzistencí.

Tento výsledek zajisté není z hlediska konzistence databáze a vyhledávání záznamů ideální, nicméně může sloužit jako podklad pro další práci. Jednak lze zjistit, jaké záznamy a proč jsou inkonzistentní, jednak lze upravit konkrétní záznamy tak, aby bylo dosaženo plné konzistence. Kontrolu konzistence záznamů lze také provést na celé databázi nebo na vybraných částech databáze, a získané poznatky mj. použít pro případnou úpravu selekčního jazyka a/nebo indexačních pravidel.

Pokud srovnáme výsledky analýzy konzistence dokumentů s výsledky analýzy konzistence indexátorů (v předchozí podkapitole), zjistíme značný nepoměr (cca 30 ku 80 % konzistenci). Tento rozdíl je samozřejmě dán především nesouměřitelností obou analýz z hlediska počtu dokumentů, podmínek jejich výběru apod., nicméně je třeba zdůraznit, že se v prvním případě (indexátoři) nepoužívala indexační pravidla (celý experiment byl prováděn mj. z důvodu jejich formulace). Můžeme se tedy důvodně domnívat, že použití indexačních pravidel by míru konzistence alespoň částečně zvýšilo.

Závěr - praktické využití analýz indexace

Není třeba zvlášť zdůrazňovat, že všechny analýzy související s kvalitou a konzistencí indexace jsou časově, personálně, metodologicky i technicky náročné, nicméně jejich význam pro praktickou činnost indexátora, správce tezauru nebo správce obsahu databáze je nesporný. Komplexní analýzu indexace je třeba považovat za natolik závažnou, že se ji nemusíme obávat označit termínem „indexační audit“.[11]

Konkrétní analýzy popsané v praktické části využití indexačního auditu naznačily nebo příkladmo uvedly, v následujícím textu si jeho význam shrňme. Výsledky indexačního auditu lze použít pro optimalizaci charakteristik většiny faktorů, které ovlivňují kvalitu a konzistenci:

- INDEXÁTOR. Indexační audit poskytuje podklady, které jsou využitelné jako zpětná vazba pro indexátory, jež vede ke zkvalitňování jejich práce. Také je lze využít pro komplexní hodnocení práce indexátora, které pak může sloužit jako nástroj pro personální nebo mzdovou politiku.
- SELEKČNÍ JAZYK. Analýzy indexace mohou prokázat, že k inkonzistenci nebo nízké kvalitě indexace přispívá samotný nástroj indexace – selekční jazyk. V tom případě by mělo být adekvátně upraveno lexikum, struktura nebo gramatika selekčního jazyka tak, aby vyhovovala nárokům na kvalitní indexaci. Uveďme příklad. Jednou ze tří nejčastějších chyb indikovanou při kontrole indexace v knihovně UK-ETF (viz praktická část) bylo nesprávné stanovení rozsahu nebo významu deskriptoru. Správce tezauru by měl v takovém případě analyzovat, zda tato chyba vznikala spíše přispěním indexátora nebo byla způsobena nedostatkem tezauru např. v oblasti poznámky o rozsahu nebo indexační poznámky u konkrétního deskriptoru. V prvním případě je třeba sdělit indexátorovi, jakou chybu dělá, v druhém případě je třeba upravit daný deskriptor v oblasti poznámky.
- DOKUMENT. Výsledky indexačního auditu také mohou přispět k přímé úpravě (reindexaci) záznamů, které byly analyzovány, a případně i dalších záznamů, které budou zpracovány obdobným způsobem za použití informací získaných z analýz. Reindexací záznamů se přímo přispívá k jejich kvalitě a konzistenci.
- PRAVIDLA PRO ZPRACOVÁNÍ. Prostřednictvím informačního auditu mohou být také získány podklady pro formulaci nebo úpravu indexačních pravidel. V tomto případě (platí to i v případě selekčního jazyka) je třeba počítat s reindexací stávajících záznamů podle nových pravidel, jinak bude databáze obsahovat inkonzistence.

[11] Termín byl inspirován pramenem uvedeným v pozn. č. 9