

# Indexace dokumentů

**Předmět: Selekční jazyky**

22. 11. 2013

Přednášející: Mgr. Silvie Kořínková Presová

<http://kisk.phil.muni.cz/mgr-silvie-korinkova-presova-dis>

# Výklad pojmu

**Indexace** (výklad z TDKIV) - ekv heslování, indexování, angl. ekv. indexing

- **Proces vyjádření obsahu dokumentu pomocí prvků selekčního jazyka**, obvykle s cílem umožnit zpětné vyhledávání. Podle použitých metod se rozlišuje pojmová a slovní indexace, podle použitých postupů se rozlišuje intelektuální, automatická a poloautomatická indexace. Z hlediska použitých selekčních jazyků se rozlišuje prekoordinovaná indexace a postkoordinovaná indexace.

**Indexování** (ČSN ISO 5963, 1996, s. 5)

- Pracovní postup popisování nebo identifikace dokumentu ve vztahu na jeho věcný obsah.

# Výklad pojmu

Různá pojetí termínu indexace:

- proces spjatý s předmětovými SJ
- proces zahrnující jak identifikační, tak věcný popis dokumentů
- proces vyjádření obsahu dokumentu

# Účel indexace

- Hlavním smyslem indexace je tvorba reprezentací publikovaných dokumentů ve formě vhodné pro zahrnutí do různých typů databází. (Lancaster, 2003, s. 1)
- Věcné vyhledávání - vyhledávání podle obsahu.

- Vytváření obsahové charakteristiky dokumentů (sémantická redukce dokumentů) není doménou pouze indexace, ale též **referování/abstrahování**.

Při indexaci je obsah popsán znaky SJ či volně pomocí slov přirozeného jazyka.

x Abstrahování je založeno na používání vět přirozeného jazyka k vytváření zhuštěných textů, které charakterizují původní dokument. (Pinkas, 2002, s. 79)

- K referování více Pinkas, 2002, s. 79-81

# Volné indexování

Volné indexování - indexování, které neužívá žádného definovaného selekčního jazyka.

- Pořádací znaky se přiřazují volně a nekontrolovaně.
  - Obvykle se vychází z přirozeného jazyka.
  - Pořádací znaky se čerpají z textu dokumentu nebo z přirozeného slovníku jednotlivce.
- (Pinkas, 2002, s. 85)

# Selekční obraz dokumentu (SOD)

**SOD** (výklad z TDKIV) -  
Množina věcných  
selekčních údajů  
vyjadřující obsah  
dokumentu. Výjimečně se  
termín používá také pro  
identifikační část  
záznamu dokumentu.

Technologizace slova :  
mluvená a psaná řeč /  
Walter J. Ong

psané texty  
mluvený projev  
jazyk a kultura  
sociologie kultury  
studie  
81 - Lingvistika. Jazyky  
81'42  
81:316.7  
316.7  
(048.8)

# Fáze procesu indexace

**Trojfázový proces** (ČSN ISO 5963, 1996):

- analýza dokumentu a určení jeho věcného obsahu (obsahová analýza)
- identifikace pojmů obsažených v předmětu
- vyjádření těchto pojmů znaky SJ

**Dvojfázový proces** (Lancaster, 2003)

- pojmová analýza - obsahová analýza, identifikace pojmů
- vyjádření pojmů znaky SJ



# Obsahová analýza

- Zjištění obsahu - kombinace pečlivé četby a rychlého čtení (zjištění shrnujících částí v textu).
- Pečlivé čtení - titul, abstrakt, shrnutí, závěr
- Za pozornost též stojí - názvy kapitol, popisy u obrázků a tabulek.
- Indexátor - brát v úvahu dokument jako celek.

# Obsahová analýza

Obsahovou analýzu není vhodné opřít pouze o název či referát - viz příklad publikace *Drahokamy / Květoslav Minařík*

Při obsahové analýze vychází indexátor z textu dokumentu, přičemž pozornost věnuje následujícím položkám:

- titul, obsah
- abstrakt, anotace
- autorská klíčová slova
- úvod a závěr, úvodní věty kapitol, odstavců
- zvýrazněné úseky textu
- ilustrace, schémata, tabulky apod.

# Dějiny loutkového divadla v Evropě 1. / Charles Magnin

- loutkové divadlo -- Evropa

Na základě anotace se pokuste určit další  
důležitá témata či téma.

- Anotace je dostupná z  
<http://www.kosmas.cz/knihy/128653/dejiny-lout>

# Dějiny loutkového divadla v Evropě 1. / Charles Magnin

- loutkové divadlo -- Evropa
- loutkové divadlo -- dějiny
- loutky -- Evropa
- loutky -- dějiny
- přehledy
- loutkové divadlo
- loutky
- dějiny
- Evropa
- přehledy

# Obsahová analýza

**Netištěné dokumenty** (audiovizuální, vizuální a zvukové dok., digitální dok.)

- Indexace na základě jejich textové složky.
- Nestačí-li textová složka, je třeba se seznámit s úplným dokumentem.

# Identifikace pojmů

Indexátor by měl identifikovat ty pojmy, které tvoří podstatu popisu předmětu dokumentu. (ČSN ISO 5963)

- Co ovlivňuje identifikaci pojmů vhodných pro reprezentaci předmětu dokumentu?
  - **Konkrétní cíl indexace** - např. tvorba SOD či rejstříku ve smyslu informačního aparátu primárního dokumentu
  - **Druh indexovaného dokumentu**
  - **Funkce IS a potřeby uživatelů**

# Identifikace pojmů

Během pojmové analýzy si musí indexátor klást následující otázky týkající se dokumentu (Lancaster, 2003, s. 9):

1. O čem dokument je?

2. Proč má být dok. přidán do fondu?

3. Jaké aspekty budou zajímat naše uživatele?

→ Efektivní indexace zahrnuje nejenom rozhodnutí o tom, o čem dokument je, ale také z jakého důvodu bude zajímat určitou skupinu uživatelů.

# Identifikace pojmů - vliv uživatelů

- Z hlediska indexace zaměřené na uživatele -  
**dokument** → **různé IS** → **různé SOD**
- **Zachycení témat, které mají** pro uživatele IS  
potencionální hodnotu.
- Čím je specializovanější pracoviště a jeho uživatelé, tím je pravděpodobnější, že indexace může být a měla by být přizpůsobena přesným zájmům uživatelů.
- **User-centered indexing** (Fidel, 1994) -  
Odkazuje k indexaci na základě požadavků očekávaných od určité skupiny uživatelů.



# Identifikace pojmů

→ **Pojmová analýza by neměla být ovlivněna charakterem SJ**, který bude použit při převodu do PZ. Indexátor musí nejprve rozhodnout jaká témata je potřeba reprezentovat. (Lancaster, 2003, s. 26)

X

→ „Analýza i transkripce by se měly provádět s pomocí nástrojů indexování jako jsou tezaurus a klasifikační systémy.“ (ČSN ISO 5963, 1996, s. 5)

# Vyjádření pojmů znaky SJ

„Překlad“ identifikovaných pojmů do pořádacích znaků SJ - např. deskriptory, předmětová hesla, klasifikační znaky, volně tvořená klíčová slova.

- Nutné zohlednit syntaxi (indexační pravidla) jednotlivých systémů, např. tvorba předmětových hesel NK ČR, klasifikačních znaků MDT apod.

# Vyjádření pojmů pomocí deskriptorů tezauru Eurovoc

- ženská práva
- domácí násilí
- feminismus
- gender
- postavení žen
- ženská hnutí

# Vyjádření pojmů pomocí deskriptorů tezauru Eurovoc

- ženská práva ----- práva žen
- domácí násilí ----- domácí násilí
- feminismus ----- ženské hnutí
- gender ----- sociologie
- postavení žen ----- ženská otázka
- ženská hnutí ----- ženské hnutí

# Vyjádření pojmů znaky SJ

Pojmy, které nejsou obsaženy ve slovníku SJ se vyjádří:

- **bud' novými znaky, které se současně zařadí do selekčního jazyka,**
- **nebo významově širšími znaky selekčního jazyka a nové pojmy lze zařadit mezi kandidáty.**

# Vyjádření pojmů znaky SJ - specifičnost slovníku

- Druh SJ (PSJ, SSJ) není nejdůležitějším faktorem ovlivňujícím fázi překladu.
- **Důležitější je šíře tematického pokrytí a specifičnost slovníku.**
- Různé SJ mohou specifičnosti dosáhnout různě, např. kombinací PZ

např. vyjádření složeného tématu **dějiny hygieny**

- SJ 1 **hygiena - dějiny**      předmetové heslo
- SJ 2 **613/614(091)**      klasifikační znak MDT
- SJ 3 **hygiena**      → deskriptory  
**dějiny**      →

# Vyjádření pojmů znaky SJ

Ve fázi vyjádření pojmů znaky SJ lze rozlišit dva typy indexace (Lancaster, 2003, s. 18):

- **extrakce** (slovní indexace) - Slova nebo slovní spojení pro reprezentaci předmětu dokumentu jsou vybrána přímo z dokumentu.
- **přiřazování** (pojmová indexace) - Obsah dokumentu je reprezentován pomocí slov, slovních spojení či znaků umělého jazyka, která jsou vybrána z jiného zdroje, než je samotný dokument.

# Úplnost indexace

Úplnost se vztahuje k tomu, nakolik jsou v SOD zachycena hlavní témata, popř. dílčí témata a klíčové pojmy, které mají pro uživatele IS potencionální hodnotu.

Vztahuje se k počtu pořadacích znaků.

- **úplná indexace** - Předmět dokumentu je pořadacími znaky SJ pokryt kompletně - je užito dostatečný počet PZ.

X

- **výběrová/selektivní indexace** - Jsou zachycena pouze hlavní témata, mnohem méně PZ než u úplné indexace.



# Úplnost indexace

Úplnost indexace ovlivňuje pozitivně úplnost vyhledávání (umožňuje vyčerpávající vyhledávání), zapříčiňuje ale nižší míru přesnosti vyhledávání.

- **míra úplnosti rešerše** - Jak velká část relevantních dok. byla vyhledána? Poměr počtu nalezených relevantních záznamů k počtu všech relevantních záznamů v databázi.
- **míra přesnosti rešerše** - Jak velká část nalezených dok. je relevantní? Poměr počtu nalezených relevantních záznamů k celkovému počtu záznamů v rešerši.

Dotaz (X Y Z)

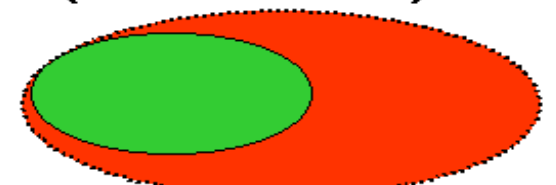
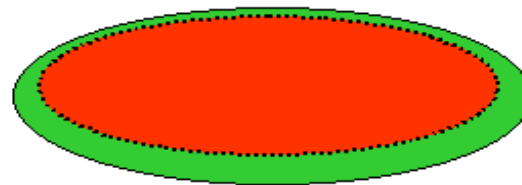
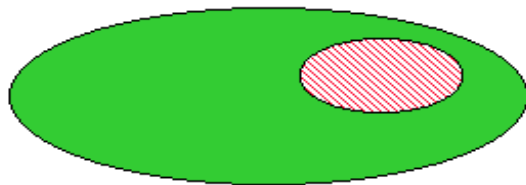
CHCI NAJÍT

NAJDU

(X and Y and Z)

Accrue (X, Y, Z)

(X or Y or Z)



**Ztráta úplnosti**

Klasická technologie

**Optimální výsledek**

technologie **Verity**

**Ztráta přesnosti**

Klasická technologie

Vejlupek, 2001

# Úplnost indexace – Věčný přepych/ Gilles Lipovetsky

## SVK Kladno

- luxusní zboží
  - luxus -- sociologické aspekty
  - eseje
- Anotace je dostupná z <http://www.kosmas.cz/k>

# Úplnost indexace – Věčný přepych/ Gilles Lipovetsky

## PH V FF MU

- luxus -- dějiny
- luxus -- filozofické aspekty
- luxus -- sociologické aspekty
- konzumní společnost
- eseje

## SVK Kladno

- luxusní zboží
- luxus -- sociologické aspekty
- eseje

# Úplnost indexace

## Kdy neindexovat dílčí témata?

- Pokud se dokument zabývá obecnější problematikou a dílčí témata jsou z hlediska celkového obsahu dokumentu nepodstatná.
- Pokud se dokument zabývá obecnější problematikou, v rámci které je systematicky rozpracována většina dílčích témat, která do dané obecné problematiky náleží.  
obecně platí - pokud ze stejné hierarchie 3 termíny  
- indexovat nadřazeným termínem

# Harmonizace věcné katalogizace v Česku: sen či realita? – M. Balíková

## KAPITOLY

- **Partneři procesu harmonizace věcné katalogizace**
- **Role Národní knihovny ČR**
- **Předpoklady harmonizace věcné katalogizace**
- **Charakteristika věcné katalogizace, mezinárodní pravidla a doporučení, principy**
- **Obsahová analýza**
- **Metody věcné katalogizace**
- **Nástroje věcné katalogizace**
  - Klasifikační systémy - systematické selekční jazyky**
  - Integrované systémy**
  - Soubor věcných autorit**
  - Funkce souboru věcných autorit**
- **Hloubka indexace – varianty věcného zpřístupnění**

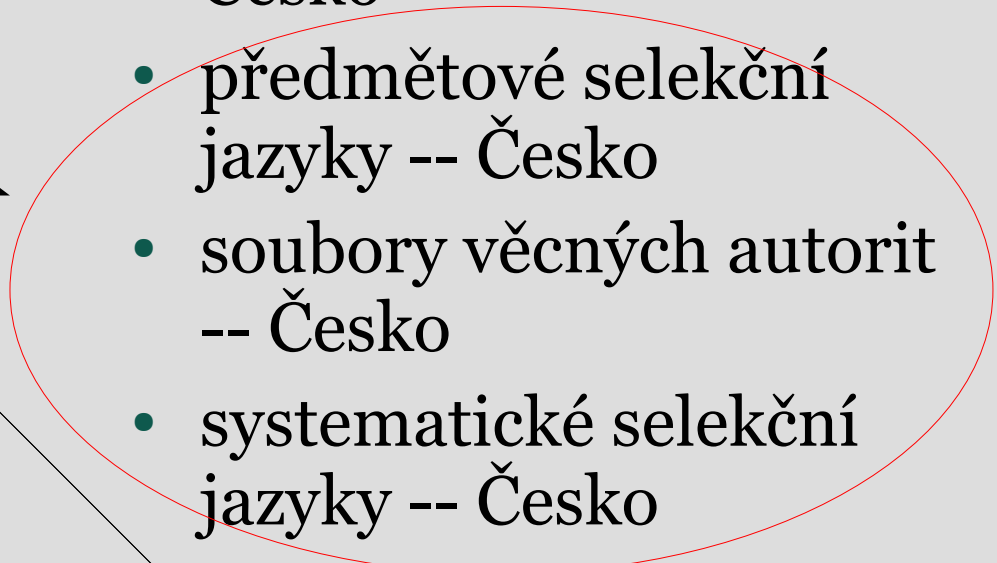
# Harmonizace věcné katalogizace v Česku: sen či realita? – M. Balíková

## SOD 1

- věcná katalogizace -- Česko
- věcné selekční jazyky -- Česko

## SOD 2

- věcná katalogizace -- Česko
- předmětové selekční jazyky -- Česko
- soubory věcných autorit -- Česko
- systematické selekční jazyky -- Česko
- obsahová analýza dokumentů



# Specifičnost indexace

Vztahuje se k míře, do jaké je konkrétní pojem, vyjadřující téma dokumentu, přesně specifikován selekčním jazykem.

Specifičnost indexace souvisí se schopností selekčního jazyka vyjádřit téma dokumentu co nejpřesněji ve vztahu ke specifičnosti obsahu.



# Vyjádření pojmů pomocí deskriptorů tezauru Eurovoc

- ženská práva ----- práva žen
- domácí násilí ----- domácí násilí
- feminismus ----- ženské hnutí
- gender ----- sociologie
- postavení žen ----- ženská otázka
- ženská hnutí ----- ženské hnutí

# Specifičnost indexace – Věčný přepych

## PH V FF MU

- luxus -- dějiny
- luxus -- filozofické aspekty
- luxus -- sociologické aspekty
- konzumní společnost
- eseje

## deskriptory ETF

(Evangelická teologická fakulta) UK v Praze

- bohatství
- postmodernizmus
- sociologie
- studie

# Kvalita indexace

- Indexace, která zajistí maximální relevanci výsledků vyhledávání.
- Míra shody obsahu SOD s obsahem dokumentu a zároveň s obsahem uživatelského/rešeršního dotazu.
- Jde o relativní hodnotu
  - účel a zaměření informačního systému
  - potřeby a požadavky uživatelů
- Nelze hodnotit kvantitativními metodami

# Kvalita indexace - faktory vlivu

- indexátor
- použitý selekční jazyk
- indexovaný dokument
- indexační pravidla
- pracovní podmínky

# Metody hodnocení kvality indexace

- Přímá kontrola obsahové i formální správnosti SOD
- Hodnocení relevance
- Konzistence indexace

# Konzistence indexace

- **Míra shody dvou nebo více SOD**

Typy konzistence

## 1. konzistence indexátorů

a) mezi indexátory (interindexer consistency) - Shoda indexace totožného dokumentu mezi dvěma nebo více indexátory.

b) indexátora (intraindexer consistency) - Konzistence indexace jednoho indexátora.

## 2. konzistence dokumentů

a) mezi dokumenty - Srovnání SOD pojednávajících o stejném tématu.

b) konzistenci dokumentu - Srovnání SOD vztahující se k jednomu dílu.

# Výpočet konzistence indexace

- Poměr počtu souhlasných pořadacích znaků k celkovému počtu jedinečných pořadacích znaků obsažených v obou SOD.

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

C=index konzistence indexace

ab=počet souhlasných PZ v selekčních obrazech

A a B, tj. shodně zvolených indexátory

a=počet jedinečných PZ v selekčním obraze A

b=počet jedinečných PZ v selekčním obraze B

# Výpočet konzistence indexace - příklad

## indexátor A

- luxus -- dějiny
- luxus -- filozofické aspekty
- luxus -- sociologické aspekty
- konzumní společnost
- eseje

## indexátor B

- luxusní zboží
- luxus -- sociologické aspekty
- eseje

$$ab = 2$$

$$a+b = 6$$

$$C = ab / (a+b) = 2/6 = 0,33 = 33\%$$

tj. částečná konzistence



# Výpočet konzistence indexace - příklad

## indexátor A

- luxus -- dějiny
- luxus -- filozofické aspekty
- luxus -- sociologické aspekty
- konzumní společnost
- eseje

## indexátor B

- luxusní zboží
- společnost
- sociologie
- studie

$$ab = 0$$

$$a+b = 9$$

$$C = ab / (a+b) = 0 / 9 = 0$$

tj. nulová konzistence

# Výpočet konzistence indexace

- více než dva indexátoři
- $C$  = průměr konzistenčních párů

# Výpočet konzistence indexace – více než dva indexátoři

## Indexátor 1

- námořníci
- ztroskotání
- ostrovy
- Robinson Crusoe
- dobrodružné romány

## Indexátor 2

- trosečníci
- ostrovy
- dobrodružné romány
- anglická literatura
- 18. století

## Indexátor 3

- námořníci
- ztroskotání
- ostrovy
- Robinson Crusoe
- dobrodružné romány
- Pátek
- anglická literatura

C = průměr konzistenčních párů

$$C_{1,2} = 2 / 8 = 0,25 = 25 \%$$

$$C_{1,3} = 5 / 7 = 0,71 = 71 \%$$

$$C_{2,3} = 3 / 9 = 0,33 = 33 \%$$

$$C = (0,25+0,71+0,33)/3 = 0,43 = 43 \%$$

# Vztah kvality indexace a konzistence

- Konzistentní indexace se nerovná kvalitní indexace.
- Konzistence indexace zlepšuje efektivitu vyhledávání a tím pozitivně ovlivňuje kvalitu indexace.

# Povinná a použitá literatura

- *ČSN ISO 5963. Dokumentace. Metody analýzy dokumentů, určování jejich obsahu a výběru lexikálních jednotek selekčního jazyka.* Praha : Český normalizační institut, 1995. 10 s.  
dostupné v Ústřední knihovně FF MU – registrační pult
- Pinkas, O. 2002. *Zpracování informačních fondů.* Vyd. 1. V Praze : Vysoká škola ekonomická, 2002. **Kap. 6 Referování a indexování**, s. 79-88.
- Schwarz, J. **Praktické aspekty hodnocení kvality a konzistence indexace.** *Ikaros* [online]. 2001, roč. 5, č. 2. [cit. 2001-02-01]. Dostupné na WWW: <http://www.ikaros.cz/node/3986> ISSN 1212-5075.

# Doporučená a použitá literatura

- Fidel, R. 1994. *User-Centered Indexing*. *Journal of the American Society for Information Science and Technology*. 1994, roč. 45, č. 8, s. 572-576.
- *KTD : Česká terminologická databáze knihovnictví a informační vědy (TDKIV)* [online]. Praha : Národní knihovna České republiky, 2003. Dostupné z WWW: <http://sigma.nkp.cz/cze/ktd>
- Lancaster, F. W. 2003. *Indexing and abstracting in theory and practice*. London : Facet Publishing, 2003. 451 s. ISBN 1856044823.
- Schwarz, J. *Selekční jazyky 2 : Úvod do problematiky : Sémantická redukce dokumentů* [ppt]. Přednáška č. 1 (kombinované studium). 29. 2. 2008.
- Schwarz, J. *Selekční jazyky 2 : Úvod do problematiky : Kvalita a konzistence indexace* [ppt]. Přednáška č. 2 (kombinované studium). 21. 3. 2008.

Vejlupek, T. Firemní zpravodajský informační systém. INFORUM 2001: 7. ročník konference o profesionálních informačních zdrojích [online], [cit. 2013-11-26]. ISSN 1801-2213. Dostupné na WWW: <http://www.inforum.cz/archiv/inforum2001/pris>