ČSN ISO 5963

Vydal a vytiskl ČESKÝ NORMALIZAČNÍ INSTITUT, Praha
Rok vydání 1995, 12 stran, náklad 1000 výtisků, 6397, 706/95
Distribuce: Český normalizační institut, Hornoměcholupská 40, 102 04 Praha 10
Cenová skupina 110

ICS 01.140.20

ČESKÁ NORMA

Březen 1996

ČSN ISO 5963 (01 0174)

Dokumentace METODY ANALÝZY DOKUMENTŮ, URČOVÁNÍ JEJICH OBSAHU A VÝBĚRU LEXIKÁLNÍCH JEDNOTEK SELEKČNÍHO JAZYKA

ČSN ISO 5963

01 0174

Documentation - Methods for examing documents, determing their subjects, and selecting indexing terms
Documentation - Méthodes pour l'analyse des documents, la détermination de leur contenu et à sélection des termes
Dokumentation - Methoden für Dokumentenanalyse, Inhaltsbezeichnung und Auswahl der Indexierungstermine

Tato norma je identická s ISO 5963:1985

This standard is identical with ISO 5963:1985

Národní předmluva

Citované normy

ISO 2788 zavedena v ČSN 01 0193 Dokumentace. Pokyny pro vypracování a rozvíjení jednojazyčných tezaurů (neq ISO 2788:1986)

ISO 5964 zavedena v ČSN ISO 5964 Pokyny pro vypracování a rozvíjení vícejazyčných tezaurů (01 0172)

Další souvisící normy

ČSN ISO 5127-6 Dokumentace a informace. Slovník. Část 6: Selekční jazyky (01 0163)
ČSN ISO 5964 Pokyny pro vypracování a rozvíjení vícejazyčných tezaurů (01 0172)
ČSN 01 0180 Mezinárodní desetinné třídění (MDT). Výběr nejdůležitějších znaků
ČSN 01 0188 Tvorba předmětových hesel
ČSN 01 0192 Rejstříky publikací
ČSN 97 6030 Abecední řazení

Obdobné mezinárodní, regionální a zahraniční normy

ISO 5963 Documentation - Methods for examing documents, determing their subjects, and selecting indexing terms (Dokumentace. Metody analýzy dokumentů, určování jejich obsahu a výběru lexikálních jednotek selekčního jazyka)

Vypracování normy

Zpracovatel: Národní informační středisko České republiky, IČO 0000 1571, PhDr. Karel Pech
Pracovník českého normalizačního institutu: Jindřiška Bouřilová

MEZINÁRODNÍ NORMA

Dokumentace
Metody analýzy dokumentů, určování jejich obsahu a výběru lexikálních jednotek selekčního jazyka

ISO 5963
První vydání
1985-12-01

MDT 001.815

Deskriptory: documentation, subject indexing

Předmluva

ISO (Mezinárodní organizace pro normalizaci) je celosvětovou federací národních normalizačních organizací (členů ISO). Na mezinárodních normách obvykle pracují technické komise ISO. Každý člen ISO, který se zajímá o předmět, pro který byla vytvořena technická komise, má právo být zastoupen v této technické komisi. Práce se zúčastňují i mezinárodní organizace, vládní i nevládní, s nimiž ISO navázalo pracovní styk. ISO úzce spolupracuje s Mezinárodní elektronickou komisí (IEC) ve všech záležitostech normalizace v elektrotechnice.

Návrhy mezinárodních norem přijaté technickými komisemi se rozesílají členům ISO k hlasování. Vydání mezinárodní normy vyžaduje souhlas alespoň 75% z hlasujících členů.

Mezinárodní norma ISO 5963 byla připravena technickou komisí ISO/TC 46.

Uživatelé se upozorňují, že všechny mezinárodní normy jsou po určité době revidovány a odkazy na další mezinárodní normy obsažené v této normě se vztahující k jejich poslednímu vydání, nebylo-li uvedeno jinak.

Obsah

1 Předmět normy a aplikační oblast

1.1 Tato norma popisuje doporučené metody analýzy dokumentů, určování jejich obsahu a výběru vhodných lexikálních jednotek selekčního jazyka. Omezuje se na úvodní fáze indexování a nezabývá se praktickými postupy jednotlivých systémů indexování, ať už prekoordinovaných nebo postkoordinovaných. Popisuje rovněž obecné metody analýzy dokumentů, použitelné kdykoliv při indexování. Tyto metody jsou však zejména určeny pro použití v systémech indexování, ve kterých jsou témata dokumentů vyjádřena ve zhuštěné formě a kde jsou pojmy zaznamenány pomocí řízeného selekčního jazyka. V tomto kontextu je řízený jazyk obvykle chápán jako podmnožina termínů vybraných z přirozeného jazyka, řízená např. tezaurem. Tyto metody by však mohly být aplikovány na systémy, ve kterých jsou pro vyhledávání pojmy zastoupeny znaky vybranými z tabulek klasifikačního systému.

1.2 Metody popsané v této normě lze aplikovat na kterémkoliv pracovišti, kde indexátoři analyzují témata dokumentů a vyjadřují je lexikálními jednotkami selekčního jazyka. Neměly by se aplikovat v systémech používajících metody automatického indexování, kde jsou termíny z textu uspořádány do množin nebo tříd podle počítačem stanovitelných kritérií, jako např. četnost výskytů a/nebo sousedství v textu, ačkoliv cíle těchto systémů jsou totožné.

1.3 Tato norma je v prvé řadě určena indexátorům jako pomůcka při analýze dokumentů a identifikaci pojmů. Může být rovněž pomocí při analýze požadavků uživatelů a jejich překladu do řízených termínů selekčního jazyka při vyhledávání a mohla by sloužit jako směrnice pro zpracování referátů. Je však třeba respektovat, že ačkoliv jsou všechny tyto operace analogické, nejsou identické.

1.4 Tato norma je určena k rozšíření normalizované praxe

a) v rámci pracoviště nebo sítě pracovišť;

b) mezi různými indexačními pracovišti, zejména těmi, která si vyměňují bibliografické záznamy.

2 Citace

ISO 2788, *Dokumentace - Pokyny pro vypracování a rozvíjení jednojazyčných tezaurů.*

ISO 5964, *Dokumentace - Pokyny pro vypracování a rozvíjení vícejazyčných tezaurů.*

3 Termíny a definice

Pro potřeby této normy se používají tyto definice:

3.1 **dokument:** jakýkoliv předmět, který byl zhotoven tiskem nebo jiným způsobem a lze jej katalogizovat nebo indexovat

POZNÁMKA - Tato definice se vztahuje nejen na psané a tištěné dokumenty v papírové nebo mikrografické podobě (např. knihy, časopisy, vyobrazení, mapy), ale také na netištěné dokumenty (např. strojem čitelné záznamy, filmy, zvukové nahrávky), a trojrozměrné předměty nebo reálie používané jako ukázky.

3.2 **pojem:** jednotka myšlení

Sémantický obsah pojmu lze jinak vyjádřit kombinací dalších a odlišných pojmů, které se mohou lišit v jednotlivých jazycích nebo kulturách

3.3 **předmět:** jakýkoliv pojem nebo kombinace pojmů reprezentující téma dokumentu

3.4 **lexikální jednotka selekčního jazyka:** reprezentace pojmu ve formě buď

- termínu odvozeného z přirozeného jazyka, nejlépe podstatného jména nebo spojení podstatných jmen, nebo
- klasifikačního znaku

POZNÁMKA - Lexikální jednotka selekčního jazyka se může skládat z více než jednoho slova, V řízeném selekčním jazyce se rozlišují termíny *preferovaný termín* nebo *nepreferovaný termín*.

3.5 **preferovaný termín:** termín trvale užívaný k reprezentaci daného pojmu při indexování; někdy se nazývá "deskriptor"

3.6 **nepreferovaný termín:** synonymum nebo kvazisynonymum preferovaného termínu

Nepreferovaný termín není přidělován dokumentům, ale používá se jako heslo v rejstříku, kde je uživatel veden pokynem (např. VIZ) k preferovanému termínu; někdy se nazývá "nedeskriptor"

3.7 **rejstřík:** abecední nebo systematický soupis předmětů, které odkazují na umístění jednotlivých předmětů v dokumentu nebo souboru dokumentů

3.8 **indexování:** pracovní postup popisování nebo identifikace dokumentu ve vztahu na jeho věcný obsah.

4 Postup a účel indexování

4.1 Indexování nesouvisí s popisem dokumentu jako hmotného předmětu (např. vyjádření formy, vydavatele, vročení atd.), ačkoliv tyto faktory lze zařadit do předmětového rejstříku, umožní-li taková informace uživateli určit přesněji relevanci dokumentu.

4.2 Během indexování jsou intelektuální analýzou z dokumentu vybírány termíny, které se následně transkribují do lexikálních jednotek selekčního jazyka. Analýza i transkripce by se měly provádět s pomocí nástrojů indexování jako jsou tezaurus a klasifikační systémy.

4.3 Indexování se v podstatě skládá ze tří následujících fází, které se však v praxi překrývají:

a) analýza dokumentu a určení jeho věcného obsahu;

b) identifikace hlavních pojmů obsažených v předmětu;

c) vyjádření těchto pojmů lexikálními jednotkami selekčního jazyka.

Všechny tyto etapy spolu s kontrolou kvality jsou pojednány v oddílech 5 až 8.

5 Analýza dokumentu

5.1 Zevrubnost, s níž lze dokument analyzovat, závisí z velké části na jeho fyzické formě. Rozlišují se dva rozdílné případy, tj. tištěné a netištěné dokumenty.

5.2 Tištěné dokumenty jsou typické v případě knihoven a informačních středisek, kde se fond skládá většinou z monografií, časopisů, zpráv, konferenčních materiálů atd. Úplné porozumění těmto dokumentům teoreticky spočívá na pročtení většiny textu. Čtení celého dokumentu je často nepraktické, indexátor by se však měl ujistit, že nepřehlédl žádné důležité informace. Důležité části textu by měly být pečlivě posouzeny, přičemž zvláštní pozornost by měla být věnována následujícím položkám:

a) název;

b) referát, byl-li zpracován;

c) obsah;

d) úvod, úvodní věty kapitol nebo odstavců a závěr;

e) ilustrace, diagramy, tabulky a jejich popisy;

f) slova nebo skupiny slov, které jsou podtrženy nebo tištěny neobvyklým typem písma.

Všechny tyto prvky by měly být během studia dokumentu indexátorem prohlédnuty a posouzeny. Indexování pouze na základě názvu se nedoporučuje a je-li v dokumentu referát, neměl by být považován za dostatečnou náhradu analýzy vlastního textu.Názvy mohou být zavádějící; jak názvy, tak i referáty mohou být neadekvátní; v mnoha případech není ani jeden z nich spolehlivým zdrojem informací potřebných pro indexátora.

5.3 Netištěné dokumenty, jako audiovizuální, vizuální a zvukové dokumenty včetně reálií vyžadují odlišné zpracování. V praxi není vždy možné analyzovat celý dokument (např. promítání filmu). Indexování se pak obvykle provádí na základě názvu nebo synopse, ačkoliv indexátorovi by mělo být umožněno shlédnout nebo vyslechnout dokument, jestliže psaný popis není adekvátní nebo se nejeví přesný.

6 Identifikace pojmů

6.1 Po skončení analýzy dokumentu by měl indexátor systematicky identifikovat ty pojmy, které tvoří podstatu popisu předmětu dokumentu. Indexační pracoviště by měla vytvořit kontrolní soupisy faktorů, které se v oblasti pokryté rejstříkem považují za důležité.

Níže uvedené otázky ilustrují obecné faktory, které by měly kontrolní seznamy definovat:

a) Zabývá se dokument předmětem nějaké činnosti?

b) Obsahuje předmět nějaký činný pojem (např. činnost, operace, proces)?

c) Je předmět činnosti identifikován?

d) Zabývá se dokument nositelem činnosti?

e) Odkazuje se na zvláštní prostředky k provádění činnosti (např. Zvláštní nástroje, techniky nebo metody)?

f) Byly tyto faktory hodnoceny v kontextu konkrétního umístění nebo prostředí?

g) Jsou stanoveny nějaké závislé nebo nezávislé proměnné?

h) Bylo o tématu pojednáno z nějakého speciálního hlediska, které se obvykle nespojuje sincerely oborem výzkumu (např. sociologický výzkum náboženství)?

Uvedené otázky se nabízejí jako příklady obecných faktorů, které by měly být aplikovány na jakýkoliv obor. Speciální disciplíny mohou vyžadovat další otázky.

6.2 Indexátor nemusí nutně vyjadřovat v lexikálních jednotkách selekčního jazyka všechny pojmy, které identifikoval při analýze dokumentu. Volba těch pojmů, které mají být vybrány nebo vyřazeny závisí na účelu, kterému budou lexikální jednotky selekčního jazyka sloužit. Lze rozlišit řadu cílů od přípravy tištěných abecedních rejstříků až ke strojovému ukládání dat pro pozdější počítačové nebo jiné vyhledávání. Identifikace pojmů může být rovněž ovlivněna (jak bylo poznamenáno výše) druhem indexovaného dokumentu. Např. indexování odvozené z textu knih, časopiseckých článků atd. se liší od indexování na základě referátů nebo synopsí. Dvě charakteristické vlastnosti rejstříků, které jsou těmito faktory nejvíce ovlivněny, jsou úplnost a specifičnost.

6.3 Úplnost se vztahuje k počtu faktorů (souvisejících s otázkami v 6.1), které jsou reprezentovaný termíny přidělenými dokumentu indexátorem.

6.3.1 Indexátor, který se řídí postupy formulovanými výše, by měl být schopen identifikovat všechny pojmy v dokumentu, které mají pro uživatele informačního systému potencionální hodnotu. V některých případech se v tomtéž dokumentu vyskytují dvě nebo více témat v rámci oblasti pokrývané rejstříkem. Mohou být zpracovány odděleně a dokonce, je-li to nutné, několika odborníky.

6.3.2 Záběr daného rejstříku by neměl být interpretován příliš úzce. S růstem informačních sítí je třeba počítat s tím, že selekční údaje bezprostředně vytvořené pro jednu skupinu uživatelů (např. vědce nebo techniky) by mohly být s výhodou využity jinými skupinami uživatelů (např. ekonomy).S ohledem na toto potencionální využití se doporučuje, aby například indexátoři vědecké a technické literatury nepřehlédli další fasety předmětu, např. jeho společenské a ekonomické aspekty.

6.3.3 Hlavním kritériem při výběru pojmů by měla být vždy potenciální hodnota pojmu jakožto elementu při vyjádření předmětu dokumentu a při jeho vyhledávání. Při volbě pojmů by měl indexátor uvažovat dotazy do té míry, do jaké je s nimi seznámen, které mohou být adresovány informačnímu systému. V podstatě toto kritérium znovu potvrzuje hlavní funkci indexování. V této souvislosti by indexátor měl:

a) vybírat pojmy, které by daná skupina uživatelů považovala s ohledem na účel rejstříku za nejpřesnější;

b) je-li to nutné, modifikovat jak indexační nástroje, tak i techniky na základě zpětné vazby z dotazů. Taková modifikace by neměla být prováděna tam, kde by to ohrozilo strukturu nebo logiku selekčního jazyka.

6.3.4 Počet termínů nebo deskriptorů, které lze přidělit dokumentu, by se neměl omezovat. Jejich počet by se měl řídit pouze množstvím informací obsažených v dokumentu v souvislosti s očekávanými potřebami uživatelů rejstříku. Stanoví-li se omezení, může to vést ke ztrátě objektivity indexování a ke zkreslení informací, které by mohly být hodnotné pro vyhledávání. Je-li nutné na daném pracovišti počet termínů omezit, výběr pojmů by se měl řídit posouzením indexátora s ohledem na význam každého pojmu pro vyjádření celkového tématu dokumentu.

6.4 Specifičnost se vztahuje k míře, do jaké je konkrétní pojem, vyskytující se v dokumentu, přesně specifikován selekčním jazykem. Tato specifičnost se ztrácí, jestliže je druhý pojem reprezentován termínem s obecnějším významem.

Pojmy by měly být určeny co nejspecifičtěji. Obecnějším pojmům lze dát přednost za určitých podmínek, v závislosti na následujících faktorech:

a) indexátorem předpokládaná míra, do jaké by přílišná specifičnost mohla nepříznivě ovlivnit efektivnost systému indexování. Indexátor může např. rozhodnout, že dílčí modely vybavení by měly být reprezentovány obecnějšími termíny, jako je jméno výrobce a případně jméno skupiny modelů, zvláště, objevují-li se tyto pojmy pouze v dílčích oblastech tematické oblasti pokryté rejstříkem;

b) váha přisuzovaná pojmu autorem. Domnívá-li se indexátor, že myšlenka není plně rozvinuta nebo je autorem zmiňována pouze příležitostně, může to opravňovat k indexování na obecnější úrovni.

7 Výběr lexikálních jednotek selekčního jazyka

7.1 Při překladu pojmů do lexikálních jednotek selekčního jazyka by měl indexátor používat tyto metody (viz také dodatek):

a) pojmy, které jsou již zastoupeny v selekčním jazyce, by se měly překládat do příslušných doporučených termínů;

b) termíny, které představují nové pojmy, by měly být prověřeny z hlediska přesnosti a přijatelnosti pomocí referenčních nástrojů jako jsou:

- v jednotlivých oborech uznávané slovníky a encyklopedie;
- tezaury, především vytvořené v souladu s ISO 2788 nebo ISO 5964;
- klasifikační systémy.

Lze se rovněž poradit s odborníky v daném oboru, zejména mají-li základní znalosti indexování nebo dokumentace.

7.2 Indexátor by měl tyto nástroje znát, včetně metod a pravidel používání. Zejména by si měl uvědomovat, že tyto nástroje mohou znamenat určitá omezení, např. předem vytvořený soupis předmětových hesel nebo tabulky klasifikačního systému nemusejí umožnit přesné vyjádření pojmu, který je součástí dokumentu. Jsou-li pojmy vyjadřovány klasifikačními znaky, nesmí se zapomínat, že tato označení obvykle vyjadřují širší nebo užší kontext (tzn. hlavní třídu, která nemusí být zcela vhodná pro zpracovávaný dokument).

7.3 Je-li součástí selekčního jazyka tezaurus, počet termínů přidělených dokumentu a četnost hesel lze omezit beze ztrát, neboť generické a další apriorní vztahy lze odvodit přímo ze samotného tezauru. Při užívání tezauru by se měl pro vyjádření daného pojmu vybírat nejspecifičtější termín.

7.4 Některé selekční systémy používají role, spoje, váhy atd. Indexátor by měl znát všechna specifická pravidla spojená s těmito mechanismy.

7.5 V praxi bude indexátor často konfrontován s pojmy, které se nevyskytují v daném tezauru nebo klasifikačním systému. Podle používaného systému lze tyto pojmy zpracovat různými způsoby, např.:

a) vyjádřit termíny nebo deskriptory, které jsou okamžitě převzaty do selekčního jazyka;

b) vyjádřit dočasně obecnějšími termíny a nové pojmy zařadit mezi kandidáty na pozdější doplnění.

8 Kontrola kvality

8.1 Kvalita a konzistence indexování závisí na faktorech jako jsou

a) kvalifikace a odborná úroveň indexátora;

b) kvalita nástrojů indexování.

V ideálním případě by měly být lexikální jednotky selekčního jazyka přiděleny dokumentu a úroveň úplnosti dosažená při indexování stále stejná bez ohledu na zpracovatele. Tyto faktory by měly navíc zůstat relativně stabilní v průběhu provozu daného systému indexování. Této úrovně konzistence není vždy možné v praxi dosáhnout, ale cílená stejnorodost a z toho plynoucí vysoká předvídatelnost jsou důležitým faktorem výkonnosti selekčního systému, zejména vyměňují-li se informace mezi různými pracovišti v síti.

8.2 Naprostá osobní nezúčastněnost indexátora je nutným faktorem dosažení konzistence indexování. Subjektivní soudy při indexování pojmů a výběru selekčních termínů nepříznivě ovlivňují výkonnost selekčního sytému. Je mnohem těžší docílit konzistenci ve velkém indexačním týmu nebo tam, kde se na indexaci podílejí týmy indexátorů, pracující na různých místech jako v decentralizovaném systému. V těchto případech se doporučuje vytvořit centralizovanou kontrolu se zpětnou vazbou k indexátorovi.

8.3 Indexátor by měl mít odpovídající znalosti z indexovaného oboru. Měl by rozumět v dokumentech se vyskytujícím termínům, stejně jako pravidlům a metodám daného selekčního jazyka.

Pracoviště zabývající se zpracováním dokumentů v cizích jazycích by měla spolupracovat s jazykovými odborníky.

8.4 Kvalitního indexování lze docílit efektivněji, mají-li indexátoři přímý kontakt s uživateli. Mohou pak např. určit, zda se dané termíny nebo deskriptory užívají v chybných kombinacích, což vede k irelevantním výstupům.

8.5 Kvalita indexování rovněž závisí na otevřenosti používaného selekčního jazyka. Měl by bez omezení připouštět nové termíny nebo změny v terminologii a rovněž reagovat na nové potřeby svých uživatelů. Časté aktualizace se považují za nezbytné.

8.6 Kde je to možné, měla by být kvalita indexování testována rozborem rešeršních výstupů, např. statistikou úplnosti a mírou přesnosti.

Příloha

Vývojový diagram indexování pomocí tezauru

(Tato příloha není součástí normy.)