Zpracování a analýza metadat

Jana Kurfürstová
Moravská zemská knihovna v Brně

Obsah bloku

Metadata v knihovnictví

Zpracování metadat v discovery systémech

- ➢ Deduplikace
- ➢ Indexace
- ➢ Relevance

Linked Data

Principy fungování federace eduID

Analýza a úprava metadat s nástrojem OpenRefine

Bibliografická metadata

- Co je obsahem bibliografických záznamů?
- Co ještě je užitečné v knihovnictví popisovat kromě dokumentů?
- Jaké znáte metadatové formáty v knihovnictví?
- Jaký je vztah katalogizačních pravidel a metadatových formátů?
- Čím se vyznačují kvalitní metadata?

Bibliografická metadata

- *Kvalita metadat je určena jejich **syntaktickou a sémantickou správností**.*
- *Kvalitní metadata **odpovídají standardům** pro svůj deklarovaný formát a sdělují **maximum relevantních, strojově rozlišitelných údajů o informačním objektu**.*
- *Pro kvalitní metadata je zvolen vhodný formát, jehož struktura umožňuje efektivní zpracování údajů obsažených v záznamu.*
- *Klíčovou vlastností kvalitních metadat je jejich **konzistentnost**, tj. jednotnost rozsahu zaznamenaných údajů a způsobu zápisu každého údaje ve všech záznamech metadatového zdroje.*

Deduplikace záznamů

Co dělá deduplikace?

- Knihovny.cz
- SK ČR
- Google Scholar
- …

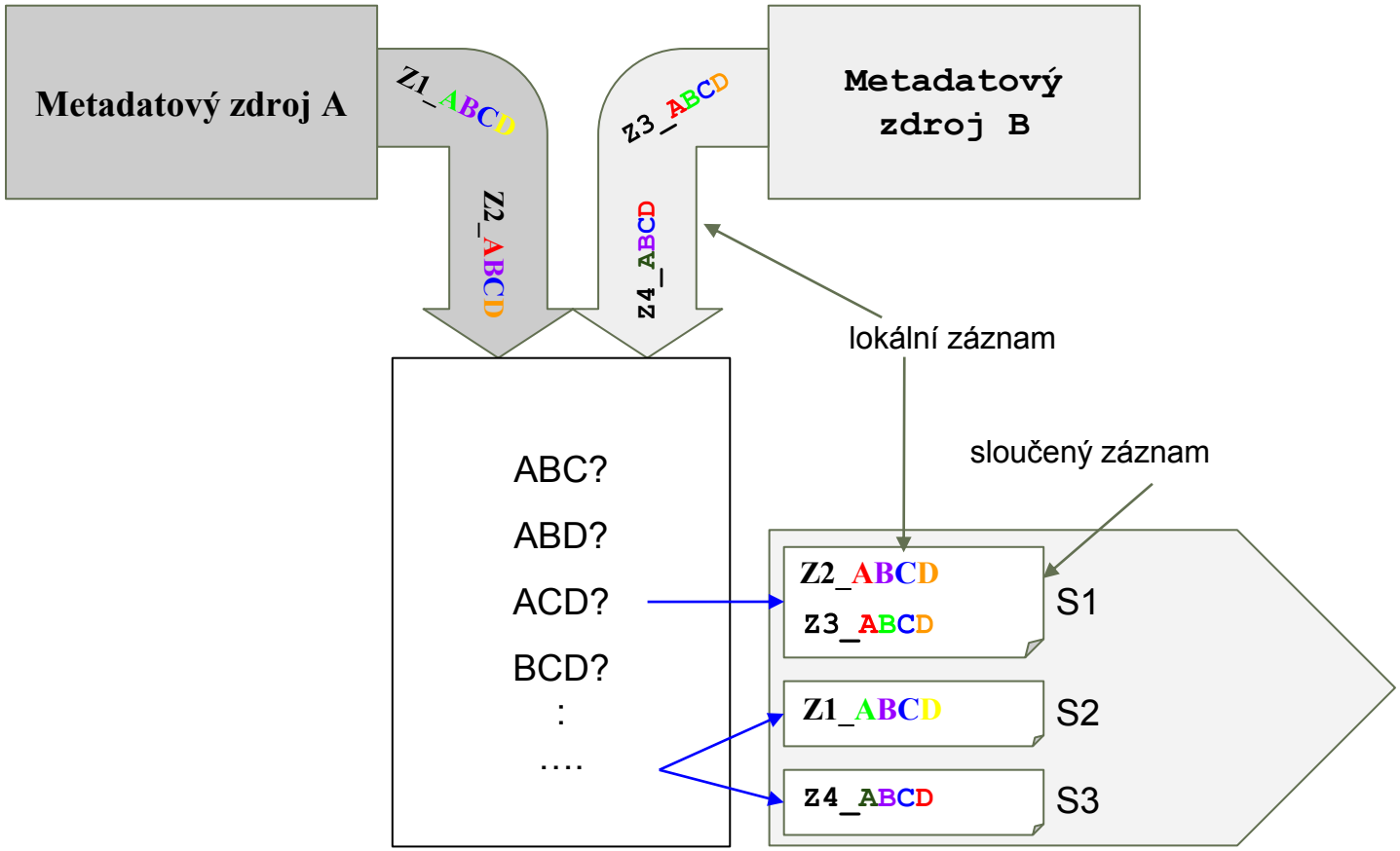
Jak myslíte, že je toho dosaženo?

Jak poznáme, že X záznamů popisuje stejný dokument?

Příklady: https://github.com/JanaKurfurstova/vyukaKISK

Jak to dělá SK ČR: https://www.caslin.cz/caslin/spoluprace/jak-prispivat-do-sk-cr/dodavani-dat/jak-probiha-davkovy-import/deduplikacni-procedury

Deduplikační klíče a kroky



Deduplikace - diskuze

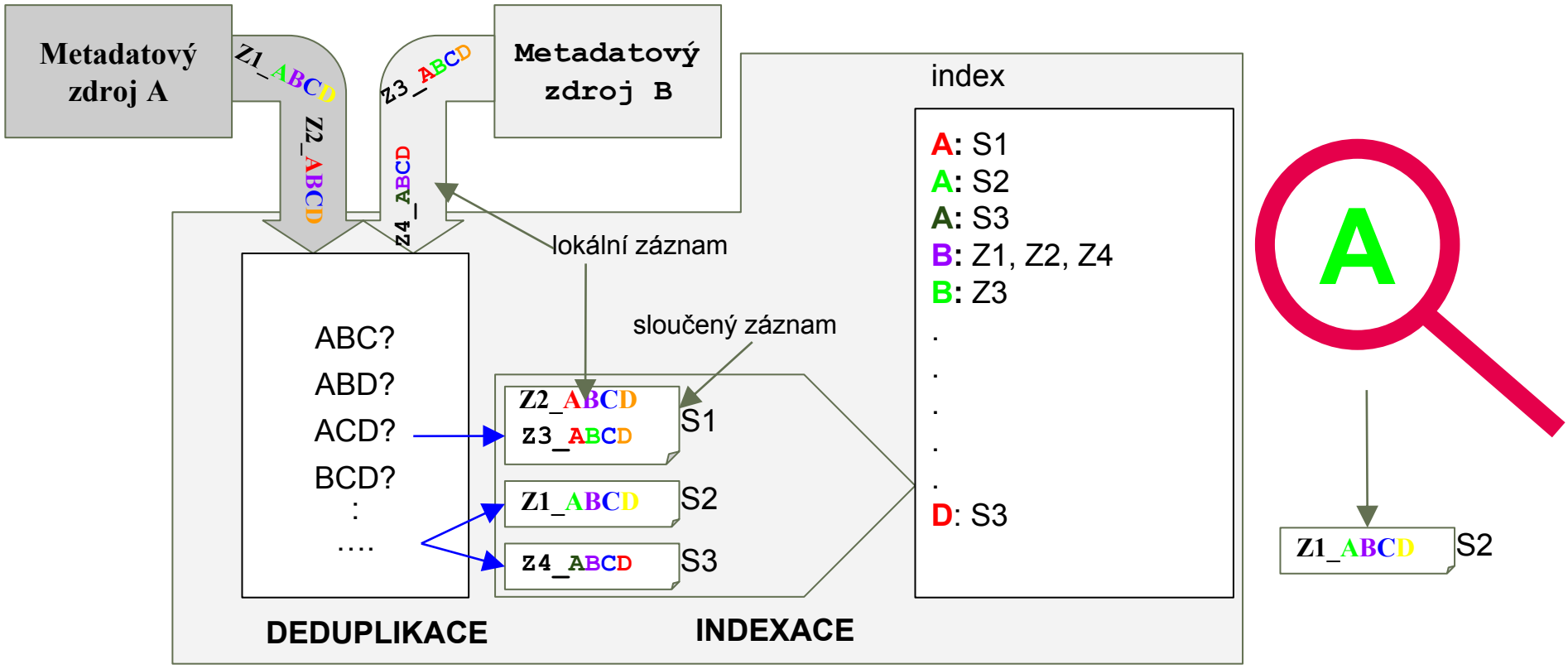
- Odstrašující příklad: https://www.knihovny.cz/Record/mkklat.50380#dedupedrecord
- Neexistuje 100 % správné řešení deduplikace:
  - moc přísná pravidla = nepřehlednost vyhledávání pro velké množství multiplicit NEBO nutnost vyhazovat na vstupu nekvalitní záznamy
  - moc benevolentní pravidla = vysoký podíl ošklivých shluků NEBO použití ve statickém prostředí s možností ručních zásahů
  - moc komplikovaná pravidla (např. podobnostní porovnávání, více kroků s více klíči) = neúnosná výpočetní náročnost NEBO použití na malých datasetech
- Tj. děláme, co se dá, ale část odpovědnosti vždy leží na knihovnách.

Indexace

K čemu je indexace?

- https://www.knihovny.cz/Search/Advanced
- https://www.cochranelibrary.com/advanced-search/search-manager
- https://isdv.upv.cz/webapp/!resdb.pta.frm

Indexace nad zdeduplikovanými záznamy



Indexace - příklady

- Indexace autorů:
  - jmenovci 1
  - jmenovci 2
  - nesmysl
- Indexace titulů:
  - co byste zaindexovali zde?
  - nesmysly: příklad
- Různé:
  - vadný rok
  - 653 smetí: příklad 1, příklad 2, příklad 3, příklad 4

Indexace - diskuze

- Obohacování
- Potenciál formátu vs. jeho skutečná využívanost
- Maximalismus vs. minimalismus a jejich důsledky
- Konsolidace hodnot ano/ne?
- Obsahy a fulltexty
- A co dál? Folksonomie, AI…

Relevance

Nastavování relevance ve vyhledávačích

- Co ovlivňuje pořadí výsledků třeba v Googlu?
- Jak byste řekli, že to dělají Knihovny.cz?
  - Vyhledejte: Harry Potter, Návrat krále, Alois Jirásek
  - Vyhledejte: librarian, librarianship, plášť, plast
- Různé typy knihoven mají odlišné potřeby:
  - základní vs. velké vědecké vs. úzce specializované knihovny
  - různé potřeby nemusí jít naplnit v různém kontextu

Linked data

Semantic web

- *I have a dream for the Web in which computers become capable of analyzing all the data on the Web – the content, links, and transactions between people and computers. A "Semantic Web", which makes this possible, has yet to emerge, but when it does, the day-to-day mechanisms of trade, bureaucracy and our daily lives will be handled by machines talking to machines. The "intelligent agents" people have touted for ages will finally materialize.*

Tim Berners-Lee (1999)

- Předpokladem pro sémantický web jsou Linked Data
  - Značkovací jazyky
  - Kontrolované slovníky
  - Ontologie

RDF a Bibframe

- RDF (W3C) - model pro zpracování metadat v podobě grafu (tj. uzly a hrany)
  - URI: povinné pro vše kromě "volných hodnot" v objektu
  - Formáty: Turtle (přímo pro tento účel), XML, JSON, cokoliv…

- strukturální model – trojice: *zdroj*, *vlastnost*, *hodnota*
- *„Angličan Shakespeare je autorem hry Hamlet, kterou napsal v roce 1599“*

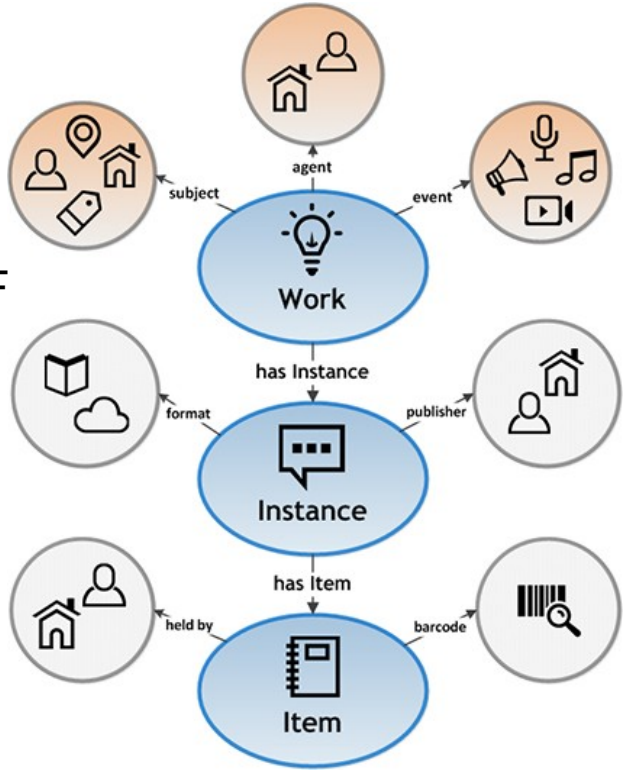
RDF:

zdroj		*vlastnost*		*hodnota*
Hamlet	-->	AUTOR	-->	Shakespeare
Hamlet	-->	TYP	-->	hra
Hamlet	-->	DATUM	-->	1599
Shakespeare	-->	NÁRODNOST	-->	anglická

- Popis v RDF: soubor tvrzení (v podobě trojic) umožňující
  - propojit *vlastnost* s konkrétním metadatovým schématem (DC, MODS, MARC (definovat její význam)
  - kombinovat prvky různých metadatových schémat (interoperabilita)
  - hodnotou zdroje může být jiný zdroj (hierarchický popis)
  - vlastnost může bý také zdrojem (složité vlastnosti)
  - … takže i pomocí jednoduchého mechanismu (trojic) můžu vytvářet libovolně složité a košaté popisy!

RDF a Bibframe

- BIBFRAME (LOC)
  - Aplikace RDF pro bibliografická metadata
  - příklad
- Problematika konverzí:
  - Jen mapovaní MARC, MODS, MADS, DC na RDF
  - Generování pořád dalších URI, které nebudou dále nikde použity
  - Upřednostňování volně tvořených hodnot místo využívání již existujících zdrojů
  - S chudými daty se nedá čarovat
  - Tj. výsledek by byl tak trochu fake, který splňuje formální požadavky, ale míjí se s hlavní myšlenkou.
  - Více o problematice např. zde či zde



SSO a eduID.cz

Single Sign-On

- Jedno přihlašování - více služeb
  - Znáte "přihlásit pomocí Google/Facebook účtu", bankovní identita…
  - Akademické prostředí - eduID.cz:
    - Přístup k licencovaným zdrojům nasmlouvaným na školu, knihovnu, ústav…
    - Nejde jen o "kdo to je?", ale i "odkud je?" a "opravdu tam pořád ještě je?"
- SAML (Security Assertion Markup Language)
  - Implementováno v nástrojích Shibboleth, SimpleSAMLphp…

eduID.cz role

poskytovatel služby
zpřístupňuje obsah / SW
uživatelům na základě smlouvy
s jejich institucí
SP = Service Provider

federace
autorita evidující IdP / SP
eduID.cz

federace federací
umožňuje fungování systému na
mezinárodní úrovni
eduGAIN

instituce
škola / knihovna / ústav
eviduje své uživatele
IdP = Identity Provider

uživatel
čtenář / student / zaměstnanec
instituce

eduID.cz workflow

- https://forms.office.com/r/ErdnQgN4Zc
- Jaké jsou výhody federací typu eduID.cz?
  - Jedny přihlašovací údaje pro více služeb.
  - Pouze mateřská instituce zná přihlašovací údaje.
  - Mateřská instituce nesleduje uživatele v externích službách.
  - Transparentnost - veřejně vystavený rozsah údajů, které budou komunikovány
  - Bezpečnost - do federace nelze jednoduše podvrhnout metadata o IdP nebo SP (leda by byl administrativní kontakt padouch), členství je podmíněno schopností bezpečné komunikace

OpenRefine

OpenRefine
švýcarský nůž na práci se strukturovanými daty

- Ke stažení: https://openrefine.org/download.html
- Návody: https://docs.openrefine.org/manual/grelfunctions
- Stáhněte si soubor, se kterým budeme pracovat: https://github.com/JanaKurfurstova/vyukaKISK/blob/main/cvicnaData.mrk
- Předpokládaná reálná situace:
  - Z knihovního systému vyexportujete MRC
  - V MarcEditu zkonvertujete MRC do MRK
  - S MRKem uděláte potřebné manipulace v OpenRefinu
  - Výsledný MRK zkonvertujete MarcEditem opět do MRC
  - MRC nahrajete do knihovního systému
  - Případně se přizpůsobíte situaci v jiném prostředí…

OpenRefine - možnosti

- **MarcEdit:** ok na kontrolu syntaxe, validaci ISBN, hromadné úpravy bez návaznosti na zbytek záznamu
- **OpenRefine:** umožní zkoumat širší souvislosti:
  - Pole X je syntakticky v pořádku, ale je v něm nesmysl.
  - Pole X je v rozporu s tím, co říká pole Y.
  - Záznamy s nějakým společným jmenovatelem mají chybu v poli X.
  - Vytvoření nového pole namapováním údajů z jiného souboru.
  - OpenRefinem můžete analyzovat a čistit i jiná data (exporty z Google Analytics, systémové logy, datasety vzniklé při výzkumu atd.)
- Dokumentace pro MARC 21 a RDA:
  - https://www.loc.gov/marc/bibliographic/
  - materiály na webu Národní knihovny, https://katdotaz.nkp.cz/
  - slidy dr. Vochozkové: http://webserver.ics.muni.cz/hanan/index.htm
  - kolega katalogizátor v knihovně, kde budete

Úskalí hromadných úprav

- OPATRNĚ s hromadnými úpravami!
  - To, že něco jde, neznamená, že je to dobrý nápad.
  - Nevidíte pramen popisu (nemáte knihu v ruce). Co když je to tam opravdu špatně? Co když si nakladatel říká pokaždé trochu jinak?
  - Chybná ISBN a další identifikátory nejsou vzácnost.
  - Autority jsou skvělá věc, ale automatické doplňování je hazard.
- Co je bezpečné?
  - Nalezení podezřelých či chybějících údajů a oprava v katalogizaci s knihou v ruce.
  - Oprava syntaktických chyb: posuvy a chyby na konkrétních znakových pozicích, interpunkce, špatně použitá podpole.
  - Doplňování opravdu jistých RDA polí.
  - Sjednocování či doplňování volně tvořených klíčových slov.

OpenRefine - příklady

- Různé způsoby otevření MRK, kontrola v řádkové podobě
- Příprava na překlopení, překlopení, seřazení sloupců
- Nalezení záznamu s opakovaným polem 245.
- Nalezení a oprava s posunutým jazykem.
- Nalezení autorů bez autority. Nalezení nesedících roků.
- Doplnění interpunkce do polí 26X.
- Nalezení hudebnin zapsaných jako textový dokument.
- Nalezení knih vydaných po r. 1989 bez ISBN.
- Nalezení RDA záznamů bez RDA polí. Doplnění chybějících 336-338 dle typu dokumentu.
- Cluster & merge volně tvořených klíčových slov.

Prostor na dotazy

DĚKUJI ZA POZORNOST

Jana Kurfürstová
kurfurstova@mzk.cz