

Datové formáty v knihovnictví a katalogizační pravidla

Jana Kurfürstová
Moravská zemská knihovna v Brně

Bibliografická metadata

- **Co je obsahem bibliografických záznamů?**

Bibliografická metadata

- **Popisná:** info o obsahu, formě a původu zdroje umožní systémům různé způsoby pořádání a prezentace a uživatelům vyhledávání a odhad relevance
- **Strukturální:** od TOC až po hierarchie u seriálů a cyklů
- **Administrativní:** lokační, o právech užití, logování změn, technické parametry (u digitálních objektů)

Bibliografická metadata

- **Co je obsahem bibliografických záznamů?**
- **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.

Bibliografická metadata

- **Co je obsahem bibliografických záznamů?**
- **Čím se vyznačují kvalitní metadata?**
- **Co ještě je užitečné v knihovnictví popisovat kromě dokumentů?**
- **Jaké znáte metadatové formáty v knihovnictví?**

Metadatové formáty

- [Dokumentace k MARC 21](#)
- MARC umožňuje zápis širokého spektra údajů, ale v praxi se setkáte s popisem zhruba následujícího [takového](#) rozsahu (další NKP materiály ke katalogizaci [zde](#) a [zde](#))
- Výhody: mezinárodně rozšířeno (jednoduché přebírání)
- Nevýhody: zastaralost, nečitelnost/nepoužitelnost mimo obor, problematická rozšiřitelnost, nemožnost obohacení o netextová data, podpora hierarchií jen na úrovni pole-podpole, nízká granularita pro některé potřeby (např. 773\$g), nepraktické řešení TOC a odkazů na ostatní objekty a entity, vazba na jazyk

Metadatové formáty

- [MODS](#) a [MADS](#) - "hezčí" XML, používané např. v Krameriu (kde jde o konverzi z MARCu)
- [METS](#) - podpora zápisu složitějších struktur např. seriálů
- PICA - od OCLC, Nizozemsko a některé německojazyčné
- [Dublin Core](#) - lze používat smysluplně (DCMI iniciativa se nyní orientuje na LD), ale láká k nekonzistentnímu a neformalizovanému použití ve své nejjednodušší formě (15 základních elementů s libovolným obsahem)
- [BIBFRAME](#) - aplikace RDF pro sémantickou katalogizaci, používáno LOC, dobré pro nové záznamy, pochybná možnost konverze z MARCu

Bibliografická metadata

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

Katalogizační pravidla

- **Jaký je vztah mezi kat. formáty a pravidly?**
 - Pořadí pramenů popisu, způsob zápisu údajů do polí daného formátu
- **Do 2015 AACR2 (Anglo-American Cataloguing Rules)**
- **Nyní RDA (Resource Description and Access)**
 - Zohlednění FRBR a existence současné techniky :-)
 - Materiály na [webu NKP](#) - jsou jediní v ČR s přístupem k oficiální dokumentaci :-)
- **slidy dr. Vochozkové:**
<http://webserver.ics.muni.cz/hanan/index.htm>

Syntaxe dle ISBD aj.

- **International Standard Bibliographic Description**
 - Pořadí prvků, interpunkce mezi nimi
 - Původně šlo o čitelnost na lístku, resp. čitelnost napříč jazyky
 - Formáty i pravidla tvořeny s ohledem na tento standard
- **Obskurnosti jako různé systémy transkripce pro různé nelatinkové jazyky**
- **[KATDotaz](#) - co se tam zodpoví, stává se pravidlem**

Praktická ukázka

- <https://koha2111-intra.rbit.cz/>
- Zkuste v modulu katalogizace přebrat záznam ze SKC
- Prohlédněte si předvyplněná pole, interpunkci, dropdowny, nápovědy
- To, co vidíte a máte povoleno vyplňovat ve formuláři, můžete jako admin upravovat v šablonách (např. zde skryto pole 001, které se inkrementuje automaticky)

Interoperabilita

Interoperabilita

➤ Z39.50 (od 70. let)

- př. dotazu `f @and @attr 1=1003 knuth @attr 1=4 @attr 5=1 computer`

➤ SRU/SRW (2002, Search/Retrieve via URL/Search/Retrieve Web service)

- nástupci, HTTP protokol, "normální" CQL

➤ V administraci AKS můžete přidávat servery, ze kterých mohou knihovníci stahovat záznamy; na wiki popsán i způsob, jak zprovoznit vlastní server.

Interoperabilita

➤ OAI-PMH

- Velmi jednoduchý protokol (6 instrukcí)
- **HTTP** (GET, POST), **XML** (MarcXML, DC...)
- Každý záznam má **hlavičku** (ID, datum poslední změny, status, příslušnost do setů) a **tělo**
- Knihovní systém má **OAI data provider**
- Agregátor či discovery system (SK ČR, Knihovny.cz, Summon či EBSCO discovery) má **OAI harvester**
- **Někdy to vážne:** data provider neumí všechny instrukce, neaktualizuje datestampy, negeneruje do záznamu potřebná pole, korektně nemaže záznamy

Interoperabilita

- **různá API (application programming interface)**
 - GET request:
`https://url.cz/api?q1=v1&q2=v2&...`
 - POST request: posílá se paklík dat na konkrétní adresu (tzn. lze používat i requesty obsahující citlivější údaje)
 - Koha, Aleph (RESTful API, X-Server),..., Získej, ObálkyKnih.cz, ..., NCIP
 - Jako response může přijít XML, JSON

Deduplikace záznamů

Co dělá deduplikace?

- Knihovny.cz
- [SK ČR](http://SK.ČR)
- Google Scholar
- ...

Jak myslíte, že je toho dosaženo?

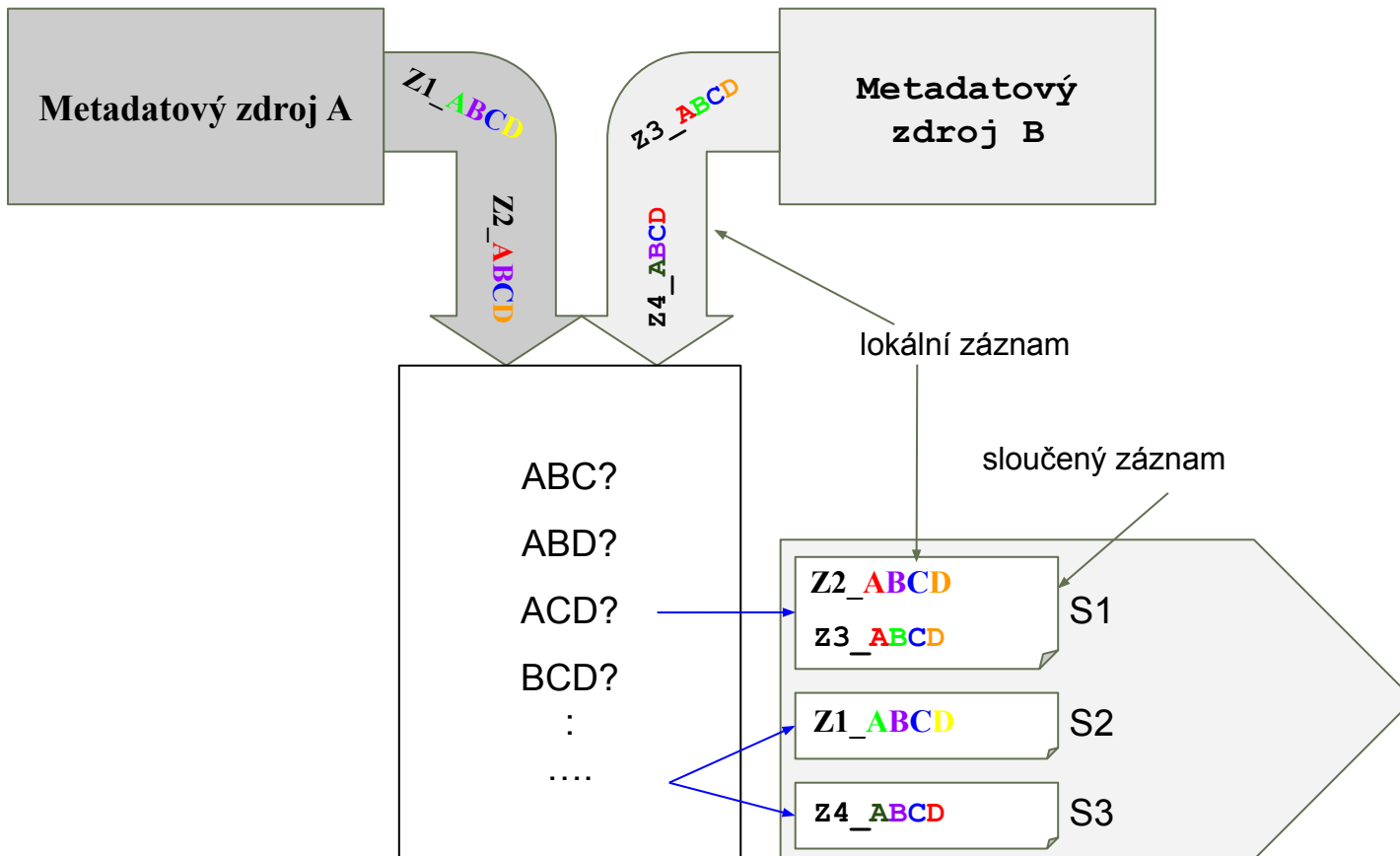
Jak poznáme, že X záznamů popisuje stejný dokument?

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-da-vkovy-import/deduplikacni-procedury>

Deduplikační klíče a kroky



Deduplikace - diskuze

- Odstrašující příklady:
 - <https://www.knihovny.cz/Record/mvk.bf731271-1302-4c4d-88e2-b07b621c2861#dedupedrecord>
 - <https://www.knihovny.cz/Record/mkricany.211198#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.

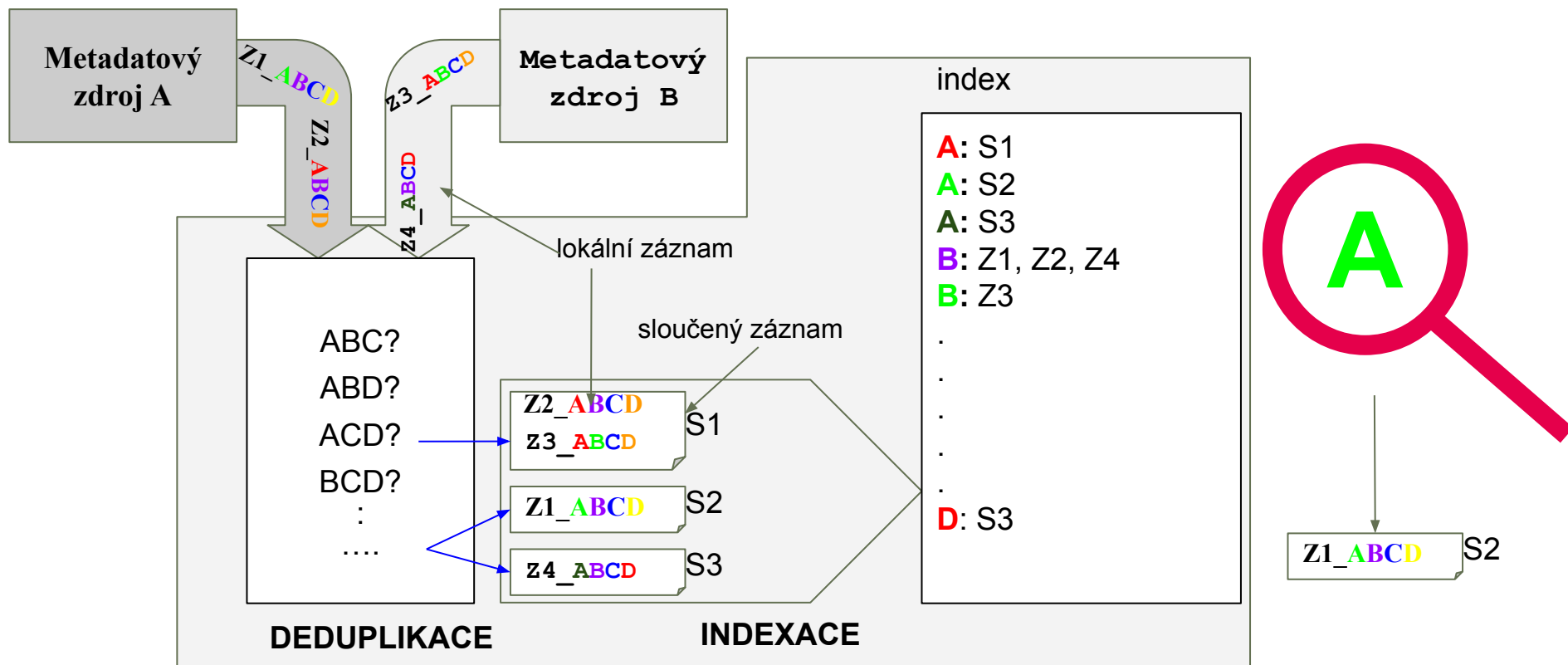
Indexace

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](#)

➤ Indexace titulů:

- co byste zaindexovali [zde](#)?
- nesmysl: [příklad 1](#)

➤ Různé:

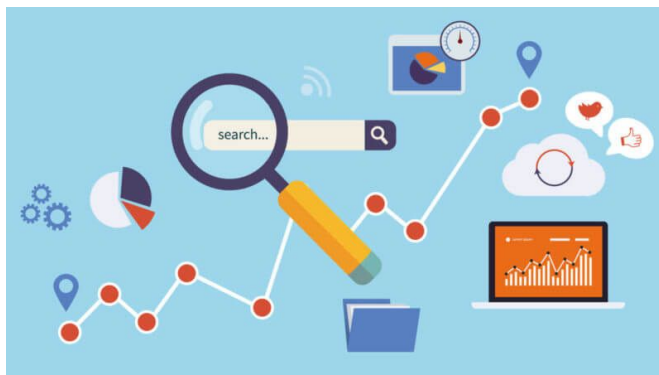
- [vadný rok](#)
- 653 smetí: [příklad 1](#), [příklad 2](#), [příklad 3](#), [příklad 4](#)

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

Prostor na dotazy

DĚKUJI ZA POZORNOST

Jana Kurfürstová
kurfurstova@mzk.cz