Metodika pro zapojení fulltextových zdrojů do Centrálního portálu knihoven

Kurfürstová, Jana; Žabičková, Petra; Žabička, Petr
2018

Metodika pro zapojení fulltextových zdrojů do Centrálního portálu knihoven

Bc. Jana Kurfürstová

Mgr. Petra Žabičková

Ing. Petr Žabička

Realizační výstup programu DG – Program na podporu aplikovaného výzkumu a experimentálního vývoje národní a kulturní identity na léta 2016 až 2022 (NAKI II)

financovaného MK ČR

v rámci projektu pod identifikačním kódem **DG16P02R006**
„CPK – Využití sémantických technologií pro zpřístupnění kulturního dědictví prostřednictvím Centrálního portálu knihoven“

Brno: Moravská zemská knihovna, 2018

Oponenti:

1. **Ing. Jan Kaňka**

   Krajská knihovna Františka Bartoše ve Zlíně, příspěvková organizace

2. **Mgr. Zuzana Kvašová**

   Národní knihovna České republiky

Obsah

Cíl metodiky

Tato metodika předkládá postup pro zapojení fulltextových zdrojů do Centrálního portálu knihoven Knihovny.cz[1] (dále jen CPK). Cílem metodiky je nabídnout poskytovatelům fulltextových zdrojů návod na přípravu a dodávání plných textů do CPK. Pokyny metodiky vycházejí ze způsobu zpracování plných textů Správcem zdrojů (nástroj pro stahování, deduplikaci a indexaci dat pro CPK vyvíjený Moravskou zemskou knihovnou v Brně) a opírají se o zkušenosti řešitelského týmu CPK s již zapojenými fulltextovými zdroji. Postupy popsané v metodice byly aplikovány při zapojování digitální knihovny MZK a jednoho neknihovního zdroje; mimo to vycházejí i ze zkušeností s několika zapojenými atypickými metadatovými zdroji. Metodika patří k výstupům projektu *Využití sémantických technologií pro zpřístupnění kulturního dědictví prostřednictvím Centrálního portálu knihoven (DG16P02R006)*[2] z programu NAKI II financovaného Ministerstvem kultury ČR. Na projektu spolupracuje Moravská zemská knihovna v Brně a Vysoké učení technické v Brně.

[1] *Knihovny.cz*. Dostupné z: https://www.knihovny.cz.

[2] CPK – Využití sémantických technologií pro zpřístupnění kulturního dědictví prostřednictvím Centrálního portálu knihoven. In: *Informační systém výzkumu, experimentálního vývoje a inovací*. Dostupné z: https://www.rvvi.cz/cep?s=jednoduche-vyhledavani&ss=detail&n=0&h=DG16P02R006.

Definice pojmů

Pro správné pochopení pokynů metodiky je nutné úvodní stanovení terminologie použité v tomto dokumentu. Následující definice předepisují význam pojmů pouze v rámci tohoto dokumentu a nejsou tedy univerzálně platné. Toto omezení umožňuje odlišit užití termínů v kontextu CPK od různorodé terminologie z externích zdrojů a přiřadit obecným pojmům užší význam pouze pro potřeby metodiky. Formulace těchto definicí se částečně opírají o hesla z TDKIV.[3]

V metodice budou opakovaně používány tyto pojmy (seřazeno dle významové návaznosti):

- **Centrální portál knihoven** (zkráceně **CPK**) je pracovní název portálu Knihovny.cz. Portál Knihovny.cz je oficiálním označením CPK pro prezentaci portálu veřejnosti.
- **Administrátor portálu** je řešitelský tým, který vyvíjí *CPK* a realizuje zapojování *fulltextových zdrojů* do *CPK* ve spolupráci s jejich *poskytovateli*.
- **Poskytovatel** je strana s kompetencí poskytnout obsah *fulltextového zdroje* ke zpracování *CPK*.
- **Fulltextový zdroj** je kolekce *fulltextů* a *metadat dokumentů* spravovaná *poskytovatelem*. *Poskytovatel* spravuje jeden či více *fulltextových zdrojů*.
- **Dokument** je zaznamenaná informace plnící funkci informačního pramene. Obsahem dokumentu je *fulltext*. *Dokument* lze popsat *metadaty*.
- **Digitální knihovna** je *fulltextový zdroj* obsahující *dokumenty* popsané *metadatovými záznamy* ve *formátu* umožňujícím předávání *údajů* bibliografického popisu (jmenných a pokud možno i věcných). *Digitální knihovny* uchovávají fulltexty ve vlastním indexu či databázi a mohou je poskytovat v podobě vyžadující minimální či žádnou úpravu před zaindexováním do *CPK*.
- **Fulltextová databáze** je *fulltextový zdroj* obsahující *dokumenty*, které kromě *plného textu* obsahují i svá vlastní *metadata*. *Fulltextové databáze* jsou schopny dodávat *fulltextové záznamy* pomocí API, protokolu nebo exportů.
- **Úložiště plných textů** je jakýkoliv *fulltextový zdroj*, který neodpovídá popisu *digitální knihovny* nebo *fulltextové databáze*. Za *úložiště fulltextů* se považují např. úložiště souborů obsahujících *digitální text* nebo klasické webové stránky s informačně přínosnými články. *Úložiště* mohou (a nemusí) být schopna dodávat *metadatové záznamy*, ale nemají vlastní technické řešení pro dodávání *fulltextů*.
- **Digitální text** je text sestávající ze znaků kódovaných strojově čitelným způsobem (např. UTF-8). Digitálním textem není vizuální zachycení tištěného textu (obrázkový soubor nebo soubor obsahující obrázky bez textové vrstvy).

---

[3] *KTD: Česká terminologická databáze knihovnictví a informační vědy (TDKIV)*. Dostupné z: http://aleph.nkp.cz/cze/ktd.

- **Plaintext** je *digitální text* obsahující pouze znaky *fulltextu* a žádné informace o formátování nebo posloupnosti znaků specifické pro různé souborové formáty mimo formátu TXT.
- **Údaj** je informace o konkrétní vlastnosti *dokumentu*. *Údaj* je sémantická jednotka – existuje nezávisle na *metadatových formátech*, které pouze stanovují způsob jeho zápisu.
- **Metadata** představují soubor *údajů* popisujících *dokument*. Obsahová struktura *metadat* je předepsána *metadatovým formátem*.
- **Metadatový záznam** obsahuje *metadata dokumentu* a pro jeho získání je nutno použít odlišný požadavek než pro získání *fulltextu dokumentu*.
- **Fulltext** (nebo také **plný text**) je obsah *dokumentu* v podobě *digitálního textu*. Může se jednat o text born-digital *dokumentu* (tj. *dokumentu*, který je vytvořen a prezentován v digitální podobě), o textovou vrstvu digitalizovaného fyzického *dokumentu* nebo o opis fyzického *dokumentu* (např. u rukopisů nerozpoznatelných OCR softwarem).
- **Fulltextový záznam** obsahuje *fulltext* i *metadata dokumentu*. *Fulltextový záznam* lze získat jedině vcelku, použitím téhož požadavku.
- **Formát dat** (zkráceně **formát**) předepisuje obsahovou strukturu *metadatových* a *fulltextových záznamů*. Pro poskytování *metadat* do *CPK* jsou relevantní *formáty* k zaznamenání popisných (a částečně i administrativních) *metadat*, jako například *formát* MODS nebo Dublin Core. Volba *formátu* pro poskytování *fulltextů* většinou závisí na specifických možnostech *zdroje* a je předmětem individuální domluvy *poskytovatele* a *administrátora portálu*.
- **Datový metaformát** (zkráceně **metaformát**) určuje formální strukturu *metadatových* a *fulltextových záznamů*. Pro zpracování v CPK je nejobvyklejší dodávání *dat v metaformátu* XML, ale poskytovat lze i v metaformátech JSON, DSV (v případě velmi jednoduchých záznamů) nebo jinak (dle domluvy *poskytovatele* s *administrátorem portálu*).
- **Datové schéma** (zkráceně **schéma**) předepisuje způsob zápisu konkrétního *formátu dat* v konkrétní *syntaktické podobě*. Např. u dat se XML *syntaxí* jde o XSD šablony a u JSONu o JSON Schema.
- **Pole** je část *metadatového* či *fulltextového záznamu* definovaná v rámci *formátu dat* pro účely zaznamenání a) konkrétního *údaje* o *dokumentu*; b) *fulltextu* či části *fulltextu dokumentu*. Při poskytování dat jsou *pole* reprezentována konzistentně pojmenovanými datovými prvky (XML elementy, JSON objekty apod.)
- **Kvalita metadat** je určena jejich formální a obsahovou správností. *Kvalitní metadata* odpovídají standardům pro svůj deklarovaný *formát* a sdělují maximum relevantních, strojově rozlišitelných údajů o *dokumentu*. Pro *kvalitní metadata* je zvolen vhodný *formát*, jehož struktura umožňuje efektivní zpracování zaznamenaných *údajů*. Klíčovou

vlastností *kvalitních metadat* je jejich konzistentnost, tj. uniformita zápisu každého *údaje* o každém *dokumentu* ve *zdroji*.

- **Analytická popisná jednotka** je stať ve sborníku, článek v časopise či novinách nebo kapitola či jiná část v monografické publikaci, opatřená vlastním názvem, která je vydána jako součást *zdrojového dokumentu* (tj. sborníku, časopisu, novin či monografie).
- **Zdrojový dokument** je *dokument* obsahující *analytické popisné jednotky*.

Seznam použitých zkratek

CPK	Centrální portál knihoven (portál Knihovny.cz)
CSV	Comma-Separated Values
čČNB	Číslo České národní bibliografie
ČDK	Česká digitální knihovna
DCMI	Dublin Core Metadata Initiative
DSV	Delimiter-Separated Values
EDM	Europeana Data Model
ESE	Europeana Semantic Elements
FOXML	Fedora Object XML
HTML	Hypertext Markup Language
ISBN	International Standard Book Number
JSON	JavaScript Object Notation
NK ČR	Národní knihovna České republiky
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
OCN	OCLC Control Number
OCR	Optical Character Recognition
MARC 21	MAchine-Readable Cataloging 21
MODS	Metadat Object Description Schema
MZK	Moravská zemská knihovna v Brně
NAKI II	Program na podporu aplikovaného výzkumu a vývoje národní a kulturní identity na léta 2016 až 2022
PID	Persistentní identifikátor
TEI	Text Encoding Initiative
TSV	Tab-Separated Values
URL	Uniform Resource Locator
UTF-8	Unicode Transformation Format – 8-bit
UUID	Univerzálně Unique Identifier
VUT	Vysoké učení technické v Brně
XML	Extensible Markup Language
XSD	XML Schema Definition

Použité standardy a doporučená referenční dokumentace

Metodika pro zapojování metadatových zdrojů do Centrálního portálu knihoven (zkráceně Metodika 1)

Do CPK jsou zapojeny zejména zdroje obsahující metadatové záznamy, a nikoliv plné texty dokumentů. Pro poskytovatele metadatových zdrojů byla vypracována metodika obsahující pokyny, jak dodávat metadata do CPK, aby je bylo možné efektivně zpracovat a prezentovat uživatelům.

Tyto pokyny se vztahují rovněž na poskytovatele fulltextových zdrojů. Každý fulltext je totiž obsahem dokumentu, který lze identifikovat a popsat pomocí metadat. CPK může zpracovávat jen takové fulltexty, pro něž dodá poskytovatel metadata s dostatečnou vypovídací hodnotou.

Metodika pro zapojování fulltextových zdrojů do Centrálního portálu knihoven se nevěnuje tématům, která již byla rozebírána v předcházející metodice. Proto je *Metodika 1* klíčovým referenčním materiálem rovněž pro poskytovatele fulltextových zdrojů.

- KURFÜRSTOVÁ, Jana, Petr ŽABIČKA a Petra ŽABIČKOVÁ. *Metodika pro zapojování metadatových zdrojů do Centrálního portálu knihoven* [online]. Brno: MZK, 2017 [cit. 2018-07-13]. Dostupné z: http://invenio.nusl.cz/record/373491/files/nusl-373491_1.pdf.

Dokumentace pro digitální knihovnu Kramerius

Instituce, které používají pro správu svých digitalizovaných dokumentů software Kramerius, mají k dispozici aktuální dokumentaci na GitHubu.

- Kramerius. *GitHub.com* [online]. [cit. 2018-07-31]. Dostupné z: https://github.com/ceskaexpedice/kramerius/wiki.

Pro zapojování do CPK jsou nejrelevantnější následující části:

- Client API. *GitHub.com: Kramerius* [online]. [cit. 2018-07-31]. Dostupné z: https://github.com/ceskaexpedice/kramerius/wiki/ClientAPIDEV.
- Remote API. *GitHub.com: Kramerius* [online]. [cit. 2018-07-31]. Dostupné z: https://github.com/ceskaexpedice/kramerius/wiki/RemoteAPI.
- Datový model. *GitHub.com: Kramerius* [online]. [cit. 2018-07-31]. Dostupné z: https://github.com/ceskaexpedice/kramerius/wiki/Data.
- Práva. *GitHub.com: Kramerius* [online]. [cit. 2018-07-31]. Dostupné z: https://github.com/ceskaexpedice/kramerius/wiki/Prava.
- Práva k datastreamům. *GitHub.com: Kramerius* [online]. [cit. 2018-07-31]. Dostupné z: https://github.com/ceskaexpedice/kramerius/wiki/PravaStreams.

NDK: Definice metadatových formátů pro digitalizaci

Dokumentace vytvořená pro účely Národní digitální knihovny obsahuje kapitoly s podrobnými pokyny pro vytváření metadat ve formátu MODS a Dublin Core. Informace z těchto kapitol jsou přínosné především pro knihovny, které vytvářejí metadata k fulltextovým zdrojům konverzí ze záznamů ve formátu MARC 21.

- HUTAŘ, Jan, Pavlína KOČIŠOVÁ, Natalie OSTRÁKOVÁ, Zdeněk VAŠEK, Iveta LODROVÁ, Pavla ŠVÁSTOVÁ a Jaroslav KVASNICA. *Definice metadatových formátů pro digitalizaci monografických dokumentů: monografií, kartografických dokumentů, hudebnin* [online]. Verze 1.3.1. Praha: NK ČR, 2018 [cit. 2018-08-31]. Dostupné z: https://www.ndk.cz/standardy-digitalizace/dmf_monografie_1-3-1.
- HUTAŘ, Jan, Pavla ŠVÁSTOVÁ, Pavlína KOČIŠOVÁ, Natalie OSTRÁKOVÁ, Iveta LODROVÁ a Jaroslav KVASNICA. *Definice metadatových formátů pro digitalizaci periodik* [online]. Verze 1.7.1. Praha: NK ČR, 2018 [cit. 2018-08-31]. Dostupné z: https://www.ndk.cz/standardy-digitalizace/dmf_periodika_1-7-1.

MODS

Oficiální dokumentace k metadatovému formátu MODS.

- *MODS: Metadata Object Description Schema* [online]. Washington: Library of Congress, [cit. 2018-08-31]. Dostupné z: http://www.loc.gov/standards/mods/.
    - XML schéma pro MODS:
      MODS 3.7 Schema, In: *MODS: Metadata Object Description Schema* [online]. Washington: Library of Congress, [cit. 2018-08-31]. Dostupné z: https://www.loc.gov/standards/mods/v3/mods-3-7.xsd.

Dublin Core

Oficiální dokumentace k metadatovému formátu Dublin Core.

- *DCMI: Dublin Core Metadata Initiative* [online]. ASIS&T, [cit. 2018-08-29]. Dostupné z: http://dublincore.org/.
    - XML schéma pro kvalifikovaný Dublin Core:
      Qualified DC XML Schema. In: *Dublin Core Metadata Initiative* [online]. DCMI, 2008, Version 2008-02-11 [cit. 2018-08-29]. Dostupné z: http://dublincore.org/schemas/xmls/qdc/2008/02/11/dcterms.xsd.
    - XML schéma pro nekvalifikovaný Dublin Core:
      Simple DC XML schema, In: *Dublin Core Metadata Initiative* [online]. DCMI, 2008, Version 2002-12-12 [cit. 2018-08-29]. Dostupné z: http://dublincore.org/schemas/xmls/simpledc20021212.xsd.

TEI

Oficiální dokumentace k metadatovému formátu TEI.

- P5: Guidelines for Electronic Text Encoding and Interchange. In: *TEI: Text Encoding Initiative* [online]. Version 3.4.0. Last updated on 23rd July 2018, revision 1fa0b54 [cit. 2018-08-31]. Dostupné z: http://www.tei-c.org/Vault/P5/current/doc/tei-p5-doc/en/html/.

EDM a ESE

Dokumentace k metadatovým formátům používaným v rámci projektu Europeana (formát ESE je považován za zastaralý).

- Europeana Data Model Documentation. In: *Europeana Pro* [online]. Posted on Tuesday November 18, 2014 [cit. 2018-08-31]. Dostupné z: https://pro.europeana.eu/resources/standardization-tools/edm-documentation.
- Europeana Semantic Elements Documentation. In: *Europeana Pro* [online]. Posted on Thursday December 4, 2014 [cit. 2018-08-31]. Dostupné z: https://pro.europeana.eu/page/ese-documentation.

OAI-PMH

Oficiální dokumentace k protokolu OAI-PMH

- LAGOZE, Carl, Herbert VAN DE SOMPEL, Michael NELSON a Simeon WARNER. The Open Archives Initiative Protocol for Metadata Harvesting. In: *Open Archives Initiative* [online]. Ithaca: Cornell University Library, Document Version 2015-01-08 [cit. 2018-08-20]. Dostupné z: https://www.openarchives.org/OAI/openarchivesprotocol.html.

ResourceSync

Oficiální dokumentace k protokolu ResourceSync. Jeho použití mohou zvážit jednak poskytovatelé spravující digitální knihovny či repozitáře pomocí softwaru již podporujícího tento způsob komunikace a jednak poskytovatelé, kteří zatím nepotřebovali automatizovaně distribuovat plné texty a hledají vhodné technické řešení pro tento úkol.

- ResourceSync Framework Specification - Table of Contents. In: *Open Archives Initiative* [online]. Ithaca: Cornell University Library, 22 February 2017 [cit. 2018-08-31]. Dostupné z: http://www.openarchives.org/rs/toc.

SWORD

Oficiální dokumentace k protokolu SWORD. Jeho použití mohou zvážit jednak poskytovatelé spravující digitální knihovny či repozitáře pomocí softwaru již podporujícího tento způsob

komunikace a jednak poskytovatelé, kteří zatím nepotřebovali automatizovaně distribuovat plné texty a hledají vhodné technické řešení pro tento úkol.

- *SWORD* [online]. JISC [cit. 2018-08-31]. Dostupné z: http://swordapp.org/.

1 Úvod

Řada knihoven v České republice vede nebo se podílí na digitalizačních projektech v souladu se svou úlohou uchovávat svěřené kulturní dědictví a umožnit jeho zpřístupnění současným čtenářům i budoucím generacím. Aktuální přehled těchto institucí je dostupný na stránkách Registru digitalizace[4] zřízeného pro koordinaci digitalizace v ČR. Digitální knihovny a repozitáře nejen digitalizovaných tištěných děl, ale i born-digital dokumentů spravují vědecké, výzkumné a vzdělávací instituce. Ačkoliv zdaleka ne všechny dokumenty v těchto zdrojích jsou veřejně dostupné bez omezení, stále mohou být pro uživatele hledajícího informační zdroje přínosnější než pouhé metadatové záznamy, lze-li je prohledávat.

V *Koncepci rozvoje knihoven na léta 2011 – 2015* byl stanoven cíl vytvořit Centrální portál, jehož prostřednictvím by uživatelé přistupovali k tištěným i digitálním informačním zdrojům českých knihoven, aniž by museli prohledávat online katalogy a digitální sbírky jednotlivých institucí.[5] Inspirací pro vznik portálu byly obdobné projekty například z Finska[6], Dánska[7], Nizozemska[8], Švýcarska[9] nebo Austrálie[10].

Vývojem portálu byla pověřena Moravská zemská knihovna v Brně. CPK byl oficiálně spuštěn 26. října 2016 jako portál Knihovny.cz.[11] Uživatelé mají nyní k dispozici discovery systém, který jim zprostředkuje funkce katalogů zapojených knihoven a umožňuje prohledávání řady dalších informačních zdrojů, včetně těch plnotextových.[12]

Pro účely zpracování metadat a fulltextů ze zdrojů zapojených do portálu byl v MZK vyvinut specializovaný nástroj Správce zdrojů. Správce zdrojů sklízí či importuje metadata a fulltexty, seskupuje je, vztahují-li se rozpoznatelně ke stejnému dokumentu, a indexuje jejich obsah za účelem snadné dohledatelnosti dokumentů pro uživatele CPK.

S daty získanými pomocí Správce zdrojů pracuje také tým na Fakultě informačních technologií Vysokého učení technického v Brně v rámci společného projektu MZK a VUT spadajícího do programu Ministerstva kultury ČR NAKI II.[13] Cílem projektu s názvem *CPK – Využití sémantických technologií pro zpřístupnění kulturního dědictví prostřednictvím Centrálního portálu knihoven* je vyvinutí nástroje pro automatické obohacování neúplných

---

4 Přehled institucí zapojených do registru digitalizace. In: *Registr digitalizace*. Dostupné z: http://registrdigitalizace.cz/rdcz/info/prehled_instituci.

5 *Koncepce rozvoje knihoven ČR na léta 2011 – 2015 včetně internetizace knihoven*. Dostupné z: http://files.ukr.knihovna.cz/200000077-a8cc8a9c7b/Koncepce_PIK_Rozp.doc.

6 *Finna.fi* [online]. Dostupné z: https://finna.fi/?lng=en-gb.

7 *Bibliotek.dk* [online]. Dostupné z: https://bibliotek.dk/eng.

8 *Bibliotheek.nl* [online]. Dostupné z: https://www.bibliotheek.nl.

9 *Swissbib* [online]. Dostupné z: https://www.swissbib.ch/.

10 *Trove* [online]. Dostupné z: http://trove.nla.gov.au.

11 *Knihovny.cz*. Dostupné z: https://www.knihovny.cz.

12 Zapojené knihovny a zdroje. In: *Knihovny.cz*. Dostupné z: https://www.knihovny.cz/Portal/Page/zapojene-knihovny-a-zdroje.

13 CPK – Využití sémantických technologií pro zpřístupnění kulturního dědictví prostřednictvím Centrálního portálu knihoven. In: *Informační systém výzkumu, experimentálního vývoje a inovací*. Dostupné z: https://www.rvvi.cz/cep?s=jednoduche-vyhledavani&ss=detail&n=0&h=DG16P02R006.

metadatových záznamů na základě sémantické analýzy informací obsažených v metadatech i fulltextech.

CPK nezavazuje poskytovatele plných textů povinností dodržovat konkrétní kritéria kvality dat. Cílem metodiky není omezovat zájemce o zapojení předepisováním podmínek. Tento dokument by měli poskytovatelé chápat jako sadu doporučení k dosažení optimálního výsledku zapojení fulltextového zdroje do CPK. Bude-li podoba a způsob poskytování dat v rozporu s doporučeními metodiky a tato skutečnost zabrání Správci zdrojů v optimálním zpracování záznamů, metodika poskytovateli poslouží jako referenční dokument k identifikaci podstaty problému a navedení na správné řešení v případech, kdy je to možné.

Je třeba brát v úvahu, že metodika nemůže pokrýt všechny eventuality, k nimž může dojít při zapojování různých typů fulltextových zdrojů. Předpokládá se, že technická řešení na straně poskytovatelů mohou mít svá specifika, která budou vyžadovat individuální zacházení.

Primárním účelem metodiky je usnadnit poskytovatelům dosažení konkrétního cíle, tj. zapojení fulltextového zdroje do CPK. Doporučení metodiky obecnějšího charakteru však lze využít i k jiným účelům – jejich dodržování poskytovatelům pomůže při zapojování do jakýchkoliv agregátorů prohledávajících fulltextové zdroje.

Metodika vychází z dosavadních zkušeností se zapojováním fulltextových a zčásti i metadatových zdrojů z knihovního i neknihovního prostředí a reaguje na potřebu existence podkladů sloužících jako vodítko pro další poskytovatele, kteří by měli zájem o zapojení do CPK. V neposlední řadě je metodika vypracována za účelem vytvoření dosud chybějící podrobné oficiální dokumentace k tématu zapojování fulltextových zdrojů do CPK.

2 Cílové skupiny

Cílovou skupinou metodiky jsou subjekty spravující kolekce dokumentů v plném textu, se zájmem stát se poskytovateli plnotextových zdrojů pro CPK. Metodika cílí především na technické administrátory těchto kolekcí, ať už jde o digitalizované fondy knihoven, institucionální repozitáře nebo jiné databáze a úložiště plných textů.

Poskytovatele lze dále dělit do užších skupin podle typu jimi spravovaných plnotextových zdrojů:

Digitální knihovny postavené na systému Kramerius od verze 4[14]
Do této skupiny se řadí většina digitalizačních projektů velkých knihoven v ČR, včetně MZK, jejíž digitalizované plné texty lze už nyní prohledávat v CPK.
V době psaní tohoto dokumentu se pracuje na zapojení České digitální knihovny[15] (dále ČDK) do CPK, což povede k současnému zapojení digitalizovaných fondů všech knihoven, které do ní přispívají. Jedním z cílů ČDK je totiž právě poskytování dat pro CPK. Knihovnám s novějšími verzemi Krameria lze tedy doporučit vstup do ČDK, čímž bude vyřešeno i jejich zapojení do CPK.
V metodice jsou však obsaženy i pokyny pro knihovny s Krameriem, které by usilovaly o to stát se poskytovatelem pro CPK nezávisle na ČDK.

Ostatní digitální knihovny
Do CPK lze zapojit i digitální knihovny a repozitáře postavené za pomoci jiného softwaru, než je Kramerius. V této metodice jsou jako digitální knihovny označeny takové zdroje, které poskytují metadatové záznamy oddělené od fulltextů a zároveň umožňují Správci zdrojů automatizované sklízení fulltextů. Za digitální knihovny lze považovat i zdroje s dokumenty, které jsou popsány metadatovými záznamy ve formátech používaných mimo prostředí knihoven. Metodika obsahuje obecný popis možností, které lze zvolit k zapojování zdrojů tohoto typu.

Databáze plných textů
Za databáze plných textů jsou považovány takové zdroje, které jsou schopny dodávat záznamy obsahující jak metadata, tak fulltext. U dokumentů z takových zdrojů se předpokládá, že mají spíše kratší fulltexty (zhruba do velikosti novinového článku) – a to právě proto, aby je bylo možno posílat celé v jednom záznamu bez rizika technických problémů. Pro zapojení databáze plných textů je nezbytné, aby poskytovatel dodal administrátorovi portálu podrobnou specifikaci formátu fulltextových záznamů a dokumentaci pro jejich sklízení prostřednictvím API či vybraného protokolu. V metodice jsou obecně popsány způsoby, kterými lze docílit zapojení fulltextových databází – konkrétnější formulaci pokynů brání rozmanitost technických řešení na straně potenciálních poskytovatelů těchto zdrojů.

[14] V době tvorby této metodiky zbývá v provozu jen několik instalací Krameria 3 – většinou se zamýšleným upgradem na novější verzi.
[15] *Česká digitální knihovna*. Dostupné z: https://www.czechdigitallibrary.cz/cs/.

Úložiště fulltextů v HTML, PDF, DOC a dalších formátech
Řada plnotextových zdrojů existuje pouze v podobě úložišť se stažitelnými soubory nebo webových stránek s obsahem, který lze považovat za dokument v plném textu. V případě zájmu správců úložiště a vysoké informační hodnoty takových dokumentů pro uživatele CPK a zapojit i takové zdroje. Do této skupiny se řadí i zdroje, které jsou schopny poskytovat samostatné metadatové záznamy, ale neumožňují sklízení plných textů v plaintextové podobě.

Důležitým faktorem ovlivňujícím úspěšnost zapojení fulltextového zdroje do CPK je schopnost poskytovatele spolupracovat s administrátorem portálu. Základní podmínkou efektivní komunikace je jmenování technického kontaktu z řad pracovníků na straně poskytovatele. Technickým kontaktem by měla být osoba, která nejlépe rozumí fungování systému, v němž jsou spravovány plné texty dokumentů.

Pokyny v metodice mají pomoci právě pracovníkům plnícím roli technického kontaktu. U technického kontaktu se nevyžaduje žádná konkrétní kvalifikace. Podstatný je především jeho aktivní přístup ke komunikaci s administrátorem portálu a zájem o řešení případných problémů spojených se zapojováním.

Úroveň spolupráce administrátora portálu s technickým kontaktem se přizpůsobuje odlišným možnostem pokročilé práce s daty na straně poskytovatele. Tyto možnosti nemusí být dány jen hloubkou odborných znalostí jmenovaného kontaktu, ale i technickými nebo licenčními podmínkami nakládání s dokumenty ve fulltextovém zdroji nebo se softwarem používaným ke správě tohoto zdroje.

Pro většinu pokynů metodiky platí následující doporučení: Nemůže-li poskytovatel dodržet technicky splnitelné pokyny, sdělí tuto skutečnost administrátorovi portálu. Administrátor zajistí dodržení takových pokynů individuálním přizpůsobením zpracování dat Správcem zdrojů v případech, kdy by to představovalo znatelný přínos pro uživatele CPK.

Kromě technických kontaktů poskytovatelů fulltextových zdrojů metodiku zčásti využijí i dodavatelé a vývojáři softwaru, kteří u svých klientů provádějí nastavení systémů pro zapojení do CPK v rámci technické podpory.

3 Důvody pro zapojení fulltextových zdrojů do CPK

Hlavní motivací správců fulltextových zdrojů k zapojení do CPK je především možnost zviditelnit svou službu pro všechny uživatele portálu, včetně těch, kteří o ní dosud nevědí. I uživatelé, kteří vyhledávají v CPK bez zaměření na konkrétní zdroj, budou nacházet dokumenty dostupné u daného poskytovatele a využijí jeho služeb, aby se dostali k nalezeným informacím. Zvyšování využívanosti služeb, digitálních i klasických, je důležité nejen pro soukromé subjekty, ale i pro instituce provozované z veřejných financí. Pokud např. knihovna vynakládá značné prostředky na digitalizaci svého fondu, je zcela v souladu s jejím posláním, aby byl výsledek tohoto úsilí co nejužitečnější pro veřejnost.

Zapojením do CPK pomohou svým uživatelům především ti poskytovatelé, kteří na webových stránkách své služby neumožňují vyhledávat v plných textech. Uživatelé budou moci v CPK zjistit, zda se jimi hledaná informace nachází v některém dokumentu z daného fulltextového zdroje, a pro přístup k onomu dokumentu pak využijí službu poskytovatele.

4 Vyjádření míry závaznosti plnění požadavků

Metodika klade poskytovatelům jen nutné minimum podmínek, jejichž dodržení je pro zapojení do CPK nezbytné. Většina pokynů metodiky má pouze doporučující charakter – jejich plnění je vítáno, nikoliv však vyžadováno. V metodice je použit systém klíčových slov pro usnadnění orientace v těchto podmínkách. Jedná se o značení obvyklé pro dokumentaci podobného typu.

Způsob označení míry závaznosti plnění pokynů metodiky vychází z terminologie v *RFC 2119: Key words for use in RFCs to Indicate Requirement Levels*.[16] Jedná se o následující klíčová slova:

- *Musí* – požadavek, jehož plnění je nezbytné.
- *Nemůže* – omezující podmínka, jejíž porušení je vyloučeno.
- *Mělo by* – doporučení, jehož nedodržení lze akceptovat v opodstatněných případech.
- *Nemělo by* či *nelze* – nežádoucí jev, který lze akceptovat v opodstatněných případech.
- *Může* – doporučení, jehož dodržení je přínosné, ačkoliv nedodržení není nedostatkem.

Kdykoliv jsou tyto výrazy psány kurzívou, znamená to, že mají význam klíčového slova pro vyjádření závaznosti plnění pokynů. Klíčová slova se mohou vyskytovat ve všech tvarech daných ohebností českého jazyka.

[16] BRADNER, Scott. RFC 2119. In: *The Internet Engineering Task Force (IETF®)*. Dostupné z: https://tools.ietf.org/html/rfc2119.

5 Formální požadavky pro zapojení do CPK

Každý poskytovatel fulltextových zdrojů pro CPK *musí* splňovat následující podmínky:

1. Poskytovatel fulltextových zdrojů *musí* souhlasit se zpracováním plných textů za účely jejich prohledávání a případně i prezentace v CPK.
2. Poskytovatel fulltextových zdrojů *může* souhlasit s použitím plných textů k experimentálním účelům v rámci projektu *CPK – Využití sémantických technologií pro zpřístupnění kulturního dědictví prostřednictvím Centrálního portálu knihoven*.
3. Poskytovatel *musí* být oprávněn nakládat s fulltextovým zdrojem v rozsahu umožňujícím naplnění podmínky z bodu 1 (a případně i bodu 2).

Prezentací dat (viz bod 1) je myšleno zobrazování krátkých úryvků o rozsahu cca 100 znaků obsahujících výraz vyhledávaný uživatelem. V CPK se nezobrazují delší pasáže ze zaindexovaných fulltextů, ani jejich plné znění. Výjimkou mohou být plné texty o rozsahu kratšího článku či hesla, které lze zobrazovat v CPK celé – poskytovatel s tím však *musí* souhlasit a zaručit, že nedojde k rozporu s autorským zákonem.

Administrátor portálu získává data vždy nejbezpečnějším možným způsobem, který je podporován na straně uživatele. Je-li zabezpečení přenosu dat na straně poskytovatele nedostatečné, administrátor portálu na to poskytovatele upozorní a data stáhne až poté, co poskytovatel zjedná nápravu nebo udělí souhlas k přenosu dat s přijetím vlastní zodpovědnosti za případný únik. Administrátor portálu uchovává plné texty výhradně v indexu CPK, který je zabezpečen proti riziku porušení autorského zákona únikem dat.

Experimentálními účely (viz bod 2) je myšleno například využití plných textů při strojovém učení, které je nezbytné pro vývoj softwaru k rozpoznávání významového obsahu dokumentů. Tento vývoj probíhá na Fakultě informačních technologii VUT, kde jsou zpracovávané fulltexty zabezpečeny proti riziku porušení autorského zákona únikem dat.

Potřeba uzavření smlouvy mezi poskytovatelem a administrátorem portálu závisí na charakteru fulltextového zdroje. Smlouva je nezbytná při zapojování zdrojů s dokumenty, které nejsou volně dostupné online nebo se na ně vážou autorská práva. V ostatních případech záleží na preferencích poskytovatele. Přesné znění smlouvy o poskytování dat je předmětem domluvy obou stran.

6 Obecné pokyny a informace k přípravě fulltextových zdrojů na zapojení do CPK

Následují obecné pokyny k technické stránce zapojování fulltextových zdrojů do CPK platné pro všechny poskytovatele. Obecné pokyny budou v dalších kapitolách zpřesňovány doporučeními pro konkrétní typy fulltextových zdrojů. Účelem těchto pokynů je obeznámit poskytovatele s žádoucí podobou dat a ozřejmit důvody, které mohou vést k problematickému začlenění obsahu zdroje do CPK. Doporučení o správné podobě dat by měla rovněž poskytovatele pobídnout k tomu, aby v rámci svých možností napravovali nedostatky odstranitelné hromadnou úpravou, nebo o nich alespoň informovali administrátora portálu.

6.1 Zajištění přístupu k plným textům pro Správce zdrojů

Poskytovatel *musí* upravit zabezpečení přístupu k plným textům i metadatům tak, aby byly dosažitelné pro Správce zdrojů CPK.

Je-li fulltextový zdroj zabezpečen firewallem, poskytovatel v něm *musí* nastavit výjimku pro IP adresy CPK.[17]

Je-li k získávání fulltextů nutné přihlašování, poskytovatel *musí* zřídit pro CPK účet s patřičnými právy a sdělit administrátorovi portálu přihlašovací údaje.

Je-li poskytovatelův systém zabezpečen omezením počtu dotazů nebo omezením množství přenášených dat, poskytovatel *musí* zajistit, aby se tato opatření nevztahovala na CPK.

6.2 Plánování sklizní

Je v zájmu poskytovatele, aby sklízení dat nezpůsobovalo přílišné zatěžování jeho systému, obzvláště v době nejvyšší návštěvnosti jím poskytované služby. Před počáteční úplnou sklizní *musí* poskytovatel sdělit administrátorovi portálu, zda je nutné tuto činnost rozložit do více etap. **Poskytovatel se *musí* dohodnout s administrátorem portálu na optimálním čase sklizní a na frekvenci aktualizací** (typicky jednou měsíčně).

6.3 Hlášení změn

Aby mohli uživatelé CPK prohledávat plné texty a metadata dokumentů aktuálně obsažených v zapojeném zdroji, poskytovatel *musí* informovat administrátora portálu o veškerých změnách, které by mohly ovlivnit sklízení dat. Administrátor portálu pak upraví konfiguraci Správce zdrojů tak, aby bylo zachováno fungující zapojení fulltextového zdroje do CPK.

[17] IP serverů Centrálního portálu knihoven Knihovny.cz jsou: `195.113.155.74`, `195.113.155.141` a `195.113.155.142`.

K informacím, které je nutné sdělovat administrátorovi portálu, patří především:

- **Nová adresa serveru** pro získávání dat (včetně případných změn v autentizaci).
- **Změny ve formátu požadavků** pro získávání dat (např. nové API).
- **Změna doby**, kdy poskytovatel provádí na serveru úkony, při nichž není vhodné sklízet data.
- **Změny struktury metadat nebo plných textů** (např. nové názvy datových polí obsahujících sklízená data).
- **Hromadné změny identifikátorů dokumentů**, které Správce zdrojů používá k získávání dat, (v případě sklízení obsahu webových stránek může jít i o změny související s novým způsobem tvorby URL na webu, například o přechod na tzv. přátelská URL).
- **Další hromadné úpravy dat**, zejména těch, které se týkají umístění a dostupnosti plných textů.

6.4 Možnosti aktualizací

Při aktualizaci získává Správce zdrojů data, která byla vytvořena či pozměněna v období, jež uplynulo od předcházející aktualizace.

6.4.1 Zdroje s přibývajícím neměnným obsahem

Do řady plnotextových zdrojů jsou přidávány nové dokumenty s tím, že se jejich plné texty ani metadata již nebudou později měnit (s výjimkou migrací, hromadných úprav apod.)

Poskytovatelé takových zdrojů *by měli* umožnit Správci zdrojů vybrat si ke sklizni dokumenty na základě data jejich vložení do zdroje. Do CPK pak bude sklízen jen obsah, který ve zdroji přibyl od poslední aktualizace.

Není-li u tohoto typu zdroje možné detekovat dokumenty přibyvší od poslední aktualizace, poskytovatel *musí* administrátora portálu informovat, že jde o zdroj, v němž se přibývající data již dále nemění. Správce zdrojů pak bude pravidelně žádat o seznam identifikátorů dokumentů ve zdroji (pokud možno za poslední rozlišitelné časové období) a stahovat nová data na základě porovnání se seznamem identifikátorů již sklizených fulltextů. **Toto se nevztahuje na zdroje malého rozsahu, které lze sklízet pokaždé celé znovu.**

6.4.2 Zdroje s proměnlivým obsahem

Charakter některých plnotextových zdrojů je takový, že data, které jsou do něj vkládána, mohou být později dále upravována.

Poskytovatelé malých fulltextových zdrojů *by měli* a poskytovatelé rozsáhlejších fulltextových zdrojů *musí* umožnit Správci zdrojů vybrat si ke sklizni dokumenty na

základě data jejich vložení do zdroje nebo poslední úpravy. Do CPK pak bude sklízen jen obsah, který ve zdroji přibyl nebo se změnil od poslední aktualizace.

Pouze u fulltextových zdrojů malého rozsahu, které lze sklízet pokaždé celé znovu, je přípustné neumožňovat detekci nových a pozměněných dat.

6.5 Doporučené kódování znaků a metaformáty dat

Preferované kódování znaků pro fulltexty i metadata poskytovaná do CPK je **UTF-8**. O použití jakéhokoliv jiného kódování je třeba informovat administrátora portálu.

Pro dodávání dat v jakémkoliv formátu je nejvhodnější použití metaformátu **XML**. Přípustné jsou však i jiné možnosti – např. metaformát JSON nebo některá varianta DSV. Poslední jmenovaný metaformát *by neměl* být používán pro dodávání záznamů obsahujících text v přirozeném jazyce (fulltexty, abstrakty, anotace aj.). Použití DSV je vyloučeno u zdrojů, jejichž záznamy obsahují opakovatelná pole. Obsahy opakovatelných polí *nelze* spojovat do jednoho pole za účelem jejich poskytování v DSV.

6.6 Poskytování metadat

Ke každému dokumentu zapojovaného fulltextového zdroje *musí* být poskytována metadata, která popisují jeho základní charakteristiky. S metadaty patřícími k plným textům zachází Správce zdrojů stejně jako se záznamy z metadatových zdrojů zapojených do CPK, tj. pokusí se je zdeduplikovat s ostatními záznamy v CPK a zaindexuje je, aby umožnil jejich prohledávání, filtrování a prezentaci uživatelům.

Žádoucí míra podrobnosti metadat se liší podle charakteru dokumentů ve zdroji. Zatímco od poskytovatelů zdrojů s digitalizovanými tištěnými dokumenty se očekávají metadata v rozsahu odpovídajícím běžným bibliografickým záznamům, u poskytovatelů jednodušeji řešených zdrojů se počítá s tím, že mohou být schopni dodávat jen několik základních údajů. Pro každý zdroj navíc mohou být relevantní jiné údaje – některé informace, které jsou zásadní např. pro šedou literaturu, mohou být zbytečné pro legislativní dokumenty, a naopak.

Přesto existují doporučení týkající se podoby metadatových záznamů, která jsou společná pro většinu zdrojů zapojovaných do CPK. Tato doporučení jsou shrnuta v *Metodice 1*.

Poskytovatelů fulltextových zdrojů se týkají především tyto **kapitoly *Metodiky 1***:

- **5.1** **Zpracování metadat** – s. 18-21
- **5.2** **Obecné požadavky na kvalitu metadat** – s. 21-24
- **Příloha 1:** **Údaje pro deduplikaci**
- **Příloha 2:** **Údaje pro indexaci**
- **Příloha 3:** **Indexace pro různé účely** (Doplňující materiál k *Příloze 2*.)
- **Příloha 6:** **Doporučení pro záznamy dokumentů ve formátu Dublin Core** (Pomůcka pro poskytovatele, kteří mají funkční, ale nezkonfigurovaný generátor Dublin Core záznamů nebo dosud nepoužívají žádný

standardizovaný formát pro distribuci metadat. Detaily k interpretaci obsahu přílohy jsou uvedeny v kapitole **5.4 Požadavky na kvalitu metadat neknihovních zdrojů** – s. 34-35.)

Metadata k plným textům lze dodávat do CPK dvojím způsobem:

6.6.1 Poskytování metadat nezávisle na fulltextech, jako samostatné metadatové záznamy

Tato varianta je obvyklá u digitálních knihoven, ale lze ji aplikovat i na jiné zdroje.

Metadatové záznamy lze dodávat prostřednictvím protokolu OAI-PMH, jiného podobně použitelného protokolu nebo přes vlastní API poskytovatele. V případě potřeby se lze domluvit i na poskytování metadat formou exportů.

Pokyny pro dodávání metadat popisuje kapitola *6 Způsoby poskytování metadat do CPK* v *Metodice 1* (od s. 35)**.** Na fulltextové zdroje se však nevztahuje řazení těchto způsobů od nejvhodnějšího k méně vhodným: U fulltextových zdrojů není problém poskytování metadat přes API. Na druhou stranu, posílání exportů se záznamy je varianta, k níž je vhodné se uchýlit, až když není jiná možnost.

Ať už poskytovatel zvolí jakýkoliv způsob dodávání metadat, *musí* zajistit, aby metadatový záznam každého dokumentu obsahoval identifikátor, na jehož základě lze k záznamu přiřadit odpovídající fulltext.

Pro poskytovatele upřednostňující použití protokolu OAI-PMH je relevantní kapitola ***6.1 Pokyny pro poskytování metadat pomocí protokolu OAI-PMH* v *Metodice 1*** (s. 36-41).

Na poskytování metadat přes API se vztahují v podstatě stejná doporučení jako na poskytování přes OAI-PMH, s tím rozdílem že poskytovatel *musí* dodat administrátorovi portálu podklady k sestavování požadavků pro stahování metadatových záznamů. **U API pro poskytování metadatových záznamů je podstatná podpora následujících funkcí:**

1 **Získání sady metadatových záznamů:**
Silně doporučeno.
Nejsou-li podporovány body 2 a 3, povinné.

1.1 **Získání všech metadatových záznamů:**
Možnost stažení všech metadatových záznamů ve zdroji bez udání dalších parametrů.
V rámci bodu 1 doporučeno.
Není-li v rámci bodu 1 podporován bod 1.2, povinné.

1.2 **Získání metadatových záznamů vytvořených či upravených v daném časovém rozmezí:**
Možnost stahovat záznamy podle data (zejména kvůli aktualizacím).
V rámci bodu silně doporučeno.
Není-li v rámci bodu 1 podporován bod 1.1, povinné.

1.3 **Stránkování pro oba předchozí body:**
Možnost získávání záznamů v krocích po několika desítkách kusů.
V rámci bodu 1 silně doporučeno.

2 **Získání seznamu identifikátorů metadatových záznamů:**
Volitelné.
Není-li podporován bod 1, povinné.

2.1 **Získání identifikátorů všech metadatových záznamů:**
Možnost získat seznam všech metadatových záznamů ve zdroji bez udání dalších parametrů.
V rámci bodu 2 doporučeno.
Není-li v rámci bodu 2 podporován bod 2.2, povinné.

2.2 **Získání identifikátorů metadatových záznamů vytvořených či upravených v daném časovém rozmezí:**
Možnost získat seznam metadatových záznamů podle data (zejména kvůli aktualizacím).
V rámci bodu 2 silně doporučeno.
Není-li v rámci bodu 2 podporován bod 2.1, povinné.

2.3 **Stránkování pro oba předchozí body:**
Možnost získávání identifikátorů záznamů v krocích po několika desítkách kusů.
V rámci bodu 2 doporučeno. U rozsáhlejších zdrojů silně doporučeno.

3 **Získání metadatového záznamu s konkrétním identifikátorem:**
Možnost získat samostatný metadatový záznam na základě zadání jeho identifikátoru.
Silně doporučeno.
Není-li podporován bod 1, povinné.

4 **Detekce smazaných záznamů:**
Možnost identifikovat záznamy, které byly odstraněny ze zdroje.
Silně doporučeno pro zdroje, ve kterých někdy dochází k mazání záznamů.
Poskytovatel má dvě možnosti, jak toho dosáhnout:

4.1 Umožnit získání seznamu identifikátorů smazaných záznamů, pokud možno i s datem jejich smazání. V optimálním případě umožnit získání tohoto seznamu za určité časové období (kvůli aktualizacím).

4.2 Zachovávat ve zdroji hlavičky smazaných záznamů s označením `deleted` (nebo s jiným předem domluveným příznakem) a s datem jejich smazání (ve smyslu data poslední úpravy záznamu), aby je bylo možné získat tak, jak je to popsáno v bodu 1.2 nebo 2.2.

6.6.2 Poskytování metadat společně s fulltexty, jako součást záznamů obsahujících všechna data dokumentu dodávaná do CPK

Tato varianta je použitelná u všech zdrojů, které jsou schopny distribuovat fulltext celého dokumentu i s metadaty uvnitř jednoho záznamu. Různé způsoby dodávání metadat spolu s plnými texty jsou popsány dále v této metodice.

6.7 Informace o dostupnosti dokumentů

V metadatech fulltextových zdrojů je obzvláště důležité poskytovat údaje o dostupnosti dokumentů. Najde-li uživatel na CPK dokument s existujícím plným textem v digitální podobě, musí být schopen se dozvědět, kde a za jakých podmínek se tomuto plnému textu může dostat. Poskytovatel tedy *musí* ke každému plnému textu dodat metadata, která to uživateli umožní, tj.:

6.7.1 Odkaz na plný text

Metadata každého dokumentu ve fulltextovém zdroji *musí* obsahovat pole s URL odkazem nebo alespoň s údajem umožňujícím sestavení URL odkazu, který *by měl* vést přímo na plný text nebo alespoň na stránku, odkud se lze k plnému textu jednoduše dostat (za splnění podmínek nutných pro daný režim přístupu, viz dále).

Jestliže je plný text rozčleněn do více části s různými URL, pak lze uvádět buďto jen odkaz na první část (pokud se z ní uživatel dokáže sám dostat na pokračování) nebo odkazy na všechny části dokumentu.

Více URL odkazů v metadatech lze uvádět také tehdy, je-li plný text dostupný na více adresách nebo v několika různých formátech (např. jako HTML, PDF a EPUB).

Odkazy od sebe *musí* být v metadatech rozpoznatelně odděleny, pokud možno použitím opakovatelného pole, a *měly by* být zapsány ve správném pořadí.

6.7.2 Režim přístupu k plnému textu

U odkazu na plný text *musí* být uvedena informace o tom, zda je celý dokument volně dostupný pro kohokoliv na internetu nebo se na něj vztahují nějaká omezení.

Volně dostupný dokument si může uživatel prohlížet v plném rozsahu, odkudkoliv a bez přihlašování.

Autorsky chráněný dokument si lze zpravidla prohlížet jen na počítači v budově instituce vlastnící originál dokumentu. Pro ostatní uživatele bývá dostupný pouze náhled či úryvek. Po uplynutí lhůty trvání majetkových autorských práv se dokument stává volně dostupným.

Další zdroje, mohou mít plné texty dokumentů **dostupné pouze po přihlášení** nebo je uvolňují až **po uplynutí určitého časového období**. Poskytovatel *musí* sdělit administrátorovi portálu, jak bude tyto informace uvádět v metadatech.

Údaj o režimu přístupu k plnému textu může být uváděn dvojím způsobem:

a) **Explicitně** – např. `policy:public` vs. `policy:private`.
b) **Datem uvolnění plného textu** – na rozdíl od varianty *a)* je tento způsob vhodný i pro zdroje, u kterých dochází ke změnám dostupnosti dokumentů, ale nelze u nich provádět aktualizace upravených záznamů.

Tento údaj *musí* být jednoznačně přiřaditelný ke konkrétnímu odkazu, je-li jich v záznamu více, a *měl by* se nacházet v jiném datovém poli než odkaz.

Pokud jsou všechny fulltexty ve zdroji dostupné za stejných podmínek, stačí s tím obeznámit administrátora portálu při zapojování zdroje.

6.7.3 Typ odkazovaných informací

Nevedou-li všechny odkazy obsažené v metadatech dokumentu na plný text, poskytovatel *by měl* uvést, jaké informace se na dané adrese nacházejí (tj. zda jde o obsah, nějaký typ redukovaného textu, rozcestník atd.)

Tento údaj *musí* být jednoznačně přiřaditelný ke konkrétnímu odkazu, je-li jich v záznamu více, a *měl by* se nacházet v jiném datovém poli než odkaz.

Vedou-li odkazy ve všech záznamech na stejný typ informací, stačí s tím obeznámit administrátora portálu při zapojování zdroje.

6.8 Poskytování obsahů a redukovaných, částečných či pro čtenáře znehodnocených plných textů

Do CPK lze zapojit i zdroje, které obsahují jen úryvky plných textů, redukované texty či obsahy dokumentů. V případě, že jsou tato data dobře čitelná pro uživatele, se lze domluvit na zobrazování celého tohoto textu v CPK. Poskytovatel s tím však *musí* souhlasit a *musí* mít právo takový souhlas udělit.

Do CPK lze zapojit i zdroje, jejichž poskytovatelé upřednostňují posílání plných textů v člověku nesrozumitelné podobě (např. slova v náhodném nebo abecedním pořadí, texty s vypuštěnými nevýznamovými slovy atd.). Je však třeba mít na paměti, že možnosti prohledávání takových textů jsou značně omezené. Tuto cestu lze zvolit tehdy, omezují-li poskytovatele složité smluvní vztahy s držiteli autorských práv. Není však nutné posílat podobně upravené texty kvůli obavě z úniku fulltextů mezi uživatele CPK – ti vidí vždy jen zhruba stoznakové úryvky.

6.9 Minimální kvalita OCR textů

Správci fulltextových zdrojů, v nichž převažují digitalizované dokumenty s nekvalitními OCR texty, by měli zvážit možnost zapojit se do CPK jen jako poskytovatelé metadatových zdrojů. Uživatel portálu by pak stále by mohl vyhledávat dokumenty alespoň podle údajů jmenného či věcného popisu a zjišťovat informace o jejich dostupnosti. Poskytovatelé, kteří vědí, že se problém nekvalitních OCR týká jen určité části dokumentů ve zdroji (např. těch, které byly digitalizovány před nasazením novějšího softwaru), mají možnost dohodnout se s administrátorem portálu na parametrech selektivního sklízení fulltextů.

6.10 Omezení rozsahu plných textů indexovaných v CPK

Stávající technické řešení CPK neumožňuje indexaci více než 55 000 stran plných textů na jeden dokument. U dokumentů, které přesahují tento limit (typicky u periodik), je indexováno jen prvních 55 000 stran. Omezujícím faktorem je zde maximální velikost záznamu v Solru CPK, kdy s rostoucí velikostí záznamu dochází ke zpomalení jeho odezvy. Zvolený limit je kompromisem mezi rychlostí a množstvím periodik, které není možné vcelku zaindexovat. Předpokládá se, že maximální možný počet stran na dokument v budoucnosti stoupne nebo bude nalezeno jiné vhodné technické řešení tohoto problému.

7 Digitální knihovny Kramerius (od verze 4)

Digitální knihovny vybudované pomocí softwarového nástroje Kramerius lze zapojit do CPK trojím způsobem. Optimální varianta počítá s kolektivním zapojením digitálních knihoven přispívajících do ČDK, ale možné je i individuální zapojování za použití krameriovského API nebo umožněním přímých dotazů do Solru.

7.1 Zapojení prostřednictvím ČDK

Česká digitální knihovna[18] je národní agregátor digitálních knihoven provozovaných v ČR. Úkolem ČDK je umožnit uživatelům vyhledávat a prohlížet dokumenty digitalizované v různých knihovnách z jednotného rozhraní a fungovat jako primární národní zdroj dat pro další projekty, včetně CPK.[19]

CPK bude sklízet plné texty všech knihoven zapojených do ČDK. **Knihovna, která dodává data do ČDK, již nepotřebuje podnikat žádné další kroky, aby své fulltexty poskytla i do CPK.**

Knihovny, které mají zájem poskytovat fulltexty CPK a zároveň zvažují jejich zpřístupnění v ČDK, *by měly* zvolit právě tuto variantu zapojení.

O podmínkách zapojení do ČDK se lze informovat prostřednictvím kontaktu uvedeného na webových stránkách projektu.[20]

Zapojování digitálních knihoven prostřednictvím ČDK je preferovaný postup.

7.2 Alternativní způsoby zapojení

Knihovny, které z různých důvodů neplánují zpřístupnit své digitalizované fondy prostřednictvím ČDK, mají možnost poskytovat fulltextové zdroje do CPK jedním ze dvou alternativních způsobů, přes API Krameria nebo dotazováním do Solru (viz dále).

7.2.1 Datové modely digitálních objektů

V obou případech je nutné, aby měl Správce zdrojů k dispozici **seznam modelů digitálních objektů FOXML**[21], které jsou v zapojované digitální knihovně použity pro různé typy dokumentů.

---

[18] *Česká digitální knihovna*. Dostupné z: https://www.czechdigitallibrary.cz/cs/.

[19] LHOTÁK, Martin. Česká digitální knihovna. *Duha*. Dostupné z: https://duha.mzk.cz/clanky/ceska-digitalni-knihovna.

[20] Zapojte se. In: *Česká digitální knihovna*. Dostupné z: https://www.czechdigitallibrary.cz/cs/zapojte-se/.

[21] Datový model. In: *GitHub.com: Kramerius*. Dostupné z: https://github.com/ceskaexpedice/kramerius/wiki/Data.

Správce zdrojů předpokládá využití následujících modelů:

- **`monograph`** pro monografie (součást všech instalací Krameria),
- **`periodical`** pro periodika (součást všech instalací Krameria),
- `map` pro mapy,
- `archive` pro archivní dokumenty,
- `manuscript` pro rukopisy,
- `sheetmusic` pro hudebniny,
- `soundrecording` pro zvukové záznamy,
- `graphic` pro obrazové dokumenty.

Výše jmenované digitální objekty sestávají z dílčích objektů, které mají své vlastní modely. U seriálových publikací a vícedílných dokumentů může být tato hierarchie vícestupňová, s objekty reprezentujícími ročníky, čísla, díly apod. Základní úroveň hierarchie tvoří objekty pro jednotlivé stránky dokumentů (datový model **`page`**), z nichž jsou získávány plné texty dokumentů.

Pokud poskytovatel používá jiné či jinak pojmenované modely pro typy dokumentů nebo potřebuje některé modely vyloučit ze zpracování, *musí* s tím obeznámit administrátora portálu.

7.2.2 Poskytování metadatových záznamů

Pro oba způsoby poskytování fulltextů jsou společné počáteční kroky, kdy si Správce zdrojů přes API Krameria stáhne metadatové záznamy k dokumentům, jejichž plné texty mají být sklizeny.

7.2.2.1 Doporučení pro přípravu metadat ke sklizni

Krameriovské záznamy ve formátu Dublin Core nebo MODS vznikají v naprosté většině případů konverzí z formátu MARC 21 a jejich kvalita vychází z kvality originálních záznamů.

Doporučení pro optimální podobu záznamů ve formátu MARC 21 jsou obsažena v ***Metodice 1***, v kapitole ***5.3.1 Požadavky na kvalitu metadat bibliografických knihovních zdrojů*** (s. 25-33).

Podrobné informace o nastavení konverze MARCových záznamů do formátů MODS a Dublin Core lze čerpat z následujících kapitol dokumentace pro přispívání do Národní digitální knihovny:

- 7.3 METS část <dmdSec> - Bibliografická metadata – MODS a Dublin Core. HUTAŘ, Jan a spol. *Definice metadatových formátů pro digitalizaci monografických dokumentů: monografií, kartografických dokumentů, hudebnin* [online]. Verze 1.3.1. Praha: NK ČR, 2018, s. 21-73 [cit. 2018-08-31].
  Dostupné z: https://www.ndk.cz/standardy-digitalizace/dmf_monografie_1-3-1.

- 7.3 METS část <dmdSec> - Bibliografická metadata – MODS a Dublin Core. HUTAŘ, Jan a spol. *Definice metadatových formátů pro digitalizaci periodik* [online]. Verze 1.7.1. Praha: NK ČR, 2018 s. 23-58 cit. 2018-08-31]. Dostupné z: https://www.ndk.cz/standardy-digitalizace/dmf_periodika_1-7-1.

U digitálních knihoven institucí, které nejsou zapojeny do CPK jako knihovní metadatový zdroj **nebo digitalizují i dokumenty dosud nezapojených knihoven, je vhodnější sklízet záznamy v podrobnějším formátu MODS**, který Správci zdrojů umožňuje efektivnější deduplikaci a indexaci.

Sklízení záznamů v méně podrobném formátu **Dublin Core dostačuje u takových institucí, které už jsou zapojeny do CPK** jako knihovní metadatový zdroj **a jejichž digitální knihovny neobsahují téměř žádné dokumenty, pro něž by v CPK neexistoval originální záznam** ve formátu MARC 21. I v těchto případech je **MODS preferovaným formátem**.

Pro zapojování každé digitální knihovny obsahující alespoň nějaké dokumenty, jejichž originální MARCové záznamy už jsou v CPK, platí následující pravidlo:

Poskytovatel *by měl* usilovat o to, aby záznamy z jeho Krameria a knihovního katalogu bylo možné zdeduplikovat v CPK. To znamená, že poskytovatel *by měl* dodržovat doporučení týkající se kvality metadat (viz výše jmenovaná dokumentace) a nerozcházet se při uvádění klíčových údajů napříč metadatovými formáty. **Při zapojování Krameria do CPK se proto doporučuje provést kontrolu shody následujících metadat:**

- Zásadní: **UUID** (v MARCu 21 jako součást URL v podpoli `856$u` nebo `911$u`)
- Zásadní: **Název, podnázev, číslo části a název části** (porovnává se na podobnost)
- Významné: **ISBN, ISSN, čČNB, OCN** (tj. číslo OCLC)
- Významné: **Rok vydání**
- Významné: **Hlavní autor**
- Významné: **Počet stran** (porovnává se na podobnost)

Poskytovatel *by měl* **věnovat zvláštní pozornost vícesvazkovým dokumentům**, u kterých mohou vyvstat problémy v případě rozdílného způsobu popisu (více v *Metodice 1*, s. 29-31). S různými možnostmi, které mohou nastat, se lze vypořádat následujícími způsoby:

a) **Dokument popsán v knihovním katalogu shora a v Krameriu zdola:**
Do záznamu v katalogu uvést odkazy se UUID všech svazků (opakovatelné pole `856`).
V záznamu v Krameriu dodržovat pravidla pro zápis názvových údajů (tj. používat správná pole pro název celého díla a pro číslo a název části).

b) **Dokument je popsán v knihovním katalogu zdola a v Krameriu shora:**
Do záznamu každého dílu v katalogu uvést jako první UUID celého díla. UUID části lze uvádět jako další v pořadí nebo nemusí být zapsáno vůbec.
V záznamu každého dílu v katalogu dodržovat pravidla pro zápis názvových údajů do podpolí `100$a`, `100$b`, `100$n` a `100$p`.

c) **Dokument je popsán v knihovním katalogu i v Krameriu shora:**
Do záznamu v katalogu uvést jako první UUID celého díla. UUID jednotlivých svazků lze uvádět jako další v pořadí nebo nemusí být zapsána vůbec.

d) **Dokument je popsán v knihovním katalogu i v Krameriu zdola:**
Uvádět konzistentně názvové údaje v záznamech pro každou část v knihovním katalogu i v Krameriu.

7.2.2.2 Postup sklízení metadatových záznamů

Sklízení metadatových záznamů probíhá následujícím způsobem:

V prvním kroku Správce zdrojů získá seznam PIDů dokumentů určených ke sklizení. Správce zdrojů si může stáhnout buďto seznam všech dokumentů v digitální knihovně nebo jen seznam nových a aktualizovaných dokumentů za určité období – záleží na tom, zda bude následovat úplná sklizeň nebo půjde jen o pravidelnou aktualizaci. PIDy jsou stahovány po skupinách (typicky po dvaceti) posloupností dotazů na `SearchResource` krameriovského API.[22]

Ve druhém kroku Správce zdrojů stahuje metadatové záznamy dokumentů posloupností dotazů na `ItemResource` krameriovského API.[23] PIDy získané v předchozím kroku jsou použity jako parametry těchto dotazů. Metadatové záznamy lze stahovat jednak ve formátu ve formátu MODS a jednak ve formátu Dublin Core.

Poté, co Správce zdrojů získá, zdeduplikuje a zaindexuje metadatové záznamy, dojde na samotné sklízení fulltextů. Fulltexty, které už do CPK dodala jiná knihovna, se nesklízejí. Totožné elektronické verze dokumentů se rozpoznávají podle UUID. Uživatelé však v CPK stále uvidí informace o dostupnosti těchto fulltextů ve všech zapojených digitálních knihovnách. Právě proto *by měl* poskytovatel zachovávat UUID titulů při migraci a replikaci digitalizovaných dokumentů.

7.2.3 Poskytování fulltextů přes API Krameria

Tento postup je vhodný pro digitální knihovny menší a střední velikosti (cca do 3 milionů stran). Jedná se o způsob, který je sice pomalejší, ale jednodušší, co se týče požadavků na poskytovatele, protože při něm **Správce zdrojů využívá funkce dostupné v každé instalaci Krameria**.

Aby bylo možné sklízet přes API Krameria, **poskytovatel *musí* Správci zdrojů umožnit stahovat plné texty i u autorsky chráněných děl**. Tzn. vytvořit pro tento účel v Krameriovi uživatelský účet s patřičnými právy a poslat administrátorovi portálu autorizační token, který bude posílán jako parametr dotazů na API Krameria.

---

[22] Client API. In: *GitHub.com: Kramerius*. Dostupné z: https://github.com/ceskaexpedice/kramerius/wiki/ClientAPIDEV.

[23] Client API, tamtéž.

Sklízení probíhá formou řady dotazů na `ItemResource` krameriovského API.[24] Správce zdrojů si nejprve opatří PIDy stránek dokumentů určených ke sklizení. Aby získal i PIDy stránek dokumentů členěných na ročníky, čísla, díly apod., doptává se na potomky digitálních objektů vždy tak dlouho, dokud jsou potomci jiného typu než `page`. Pomocí PIDů digitálních objektů typu `page` se Správce zdrojů dotáže na OCR všech stran sklízených dokumentů. Všechna OCR jsou pak v rámci každého dokumentu pospojována a zaindexována do pole `fulltext` v solrovém záznamu v CPK. Výjimkou jsou velmi obsáhlé dokumenty, u nichž je nakonec zaindexováno jen prvních 50 000 stran.

7.2.4 Poskytování fulltextů ze Solru

Tento postup je vhodný pro rozsáhlé digitální knihovny (cca nad 3 miliony stran), jejichž fulltexty představují velký objem dat. Jedná se o způsob, který je sice rychlejší, ale poněkud komplikovanější, co se týče požadavků na poskytovatele.

Situace v době psaní této metodiky je taková, že **poskytovatel *musí* Správci zdrojů povolit přímý přístup do Solru**. Ke sklízení plných textů je totiž nutné zpracovávat dotazy s parametrem `fl=fulltext`, který je Krameriem ignorován. Není však vyloučeno, že v budoucnosti bude tento problém dále řešen a CPK již nebude muset pro tento způsob sklizně požadovat přímý přístup do Solru. Poskytovatelé a zájemci o poskytování fulltextových zdrojů budou informování o dalším vývoji v řešení této otázky.

Sklízení plných textů je uskutečňováno posloupností dotazů požadujících OCR digitálních objektů typu `page`, jejichž rodiče nebo vzdálenější předci mají PIDy získané ve fázi stahování metadat. Stahuje se maximálně 50 000 stran OCR na jeden dokument, přičemž jeden dotaz je omezen na 1 000 stran. OCR v rámci každého dokumentu jsou následně pospojována a zaindexována do pole `fulltext` v solrovém záznamu v CPK.

Varianta B je aplikována i při sklízení fulltextů z ČDK.

[24] Client API. In: *GitHub.com: Kramerius*. Dostupné z: https://github.com/ceskaexpedice/kramerius/wiki/ClientAPIDEV.

8 Další digitální knihovny

Možnosti zapojení fulltextových zdrojů používajících jakýkoliv jiný systém digitální knihovny jsou velmi podobné jako v případě Krameria. Hlavní rozdíl spočívá v tom, že poskytovatel *musí* dodat administrátorovi portálu podrobnější informace o struktuře metadat i plných textů a o možných způsobech jejich poskytování.

8.1 Poskytování metadatových záznamů

U klasických digitálních knihoven bývá obvyklé, že **poskytují metadatové záznamy dokumentů nezávisle na fulltextech** některým ze způsobů popsaných v kapitole 6.6.1. Ať už jde o jakýkoliv způsob, poskytovatel *musí* sdělit administrátorovi portálu informace potřebné k získávání metadatových záznamů obsahujících všechny potřebné údaje ve vhodném formátu.

U digitálních knihoven se předpokládá, že podporují distribuci metadatových záznamů alespoň v jednom standardizovaném, všeobecně známém formátu, jako je MODS, Dublin Core, EDM či ESE. Je-li dostupných více formátů, pro poskytování metadat do CPK *by měl* být zvolen ten z nich, který umožňuje nejpodrobnější a nejpřesnější popis dokumentů (tj. MODS před Dublin Corem, kvalifikovaný Dublin Core před jednoduchým Dublin Corem, EDM před ESE apod.).

I v případě použití standardizovaných metadatových formátů *by měl* poskytovatel dodat administrátorovi portálu **úplný seznam všech metadatových polí, která jsou ve zdroji použita**. V případě dat v metaformátu XML je pro tento účel optimální šablona XSD, u dat zapsaných jako JSON potom JSON Schema. To platí obzvláště pro zdroje, v nichž se kromě polí definovaných daným standardem používají i nějaká další pole (jak je to obvyklé u zdrojů se záznamy v jednoduchém Dublin Coru). Je-li to možné, poskytovatel *by měl* zpřístupnit administrátorovi portálu i **předpisy, podle nichž jsou tvořeny metadatové záznamy**.

Od digitálních knihoven se očekává, že jejich metadata budou zpracovatelná lépe, než je tomu u zdrojů postavených na jednodušších technických řešeních. Dokumentace s pokyny pro dodávání kvalitních metadat je uvedena už v kapitole 6.6. U záznamů z digitálních knihoven s dokumenty, které bývají typickým předmětem bibliografického popisu (tj. monografie, časopisy, sborníky, články apod.), je doporučeno **věnovat zvláštní pozornost správnému zadávání údajů potřebných pro deduplikaci** (*Příloha 1* v *Metodice 1*) – digitální knihovny a repozitáře totiž mohou obsahovat dokumenty, pro něž existují záznamy v dalších zdrojích zapojených do CPK.

8.2 Poskytování fulltextů

Poskytovatel *musí* Správci zdrojů umožnit tvorbu požadavků na získávání plných textů pomocí identifikátorů, které jsou obsaženy v metadatových záznamech.

Pro digitální knihovny spravované jakýmkoliv softwarovým nástrojem přichází v úvahu obdobné způsoby poskytování fulltextů, jaké byly popsány v části o zapojování Krameriů:

- **Varianta 1:**
  Podporuje-li digitální knihovna distribuci plných textů **prostřednictvím vhodného protokolu** (např. ResourceSync nebo SWORD, pro kratší fulltexty i OAI-PMH), lze jej použít pro zapojení zdroje do CPK.

- **Varianta 2:**
  Existuje-li pro digitální knihovnu **API umožňující získávání plných textů**, lze jej použít pro zapojení zdroje do CPK.

- **Varianta 3:**
  Pokud digitální knihovna nepodporuje poskytování fulltextů prostřednictvím existujícího API či protokolu, pak lze **Správci zdrojů umožnit získávání plných textů přímým přístupem do indexu či databáze digitální knihovny**. Variantu přímého přístupu lze použít i tehdy, když je získávání fulltextů přes API nebo protokol příliš pomalé.

- **Varianta 4:**
  Není-li pro poskytovatele uskutečnitelná žádná z výše jmenovaných variant, pak je možné **dodávat fulltexty pravidelným posíláním exportů** ze systému digitální knihovny.

Ve všech případech *musí* poskytovatel seznámit administrátora portálu se strukturou fulltextů a s možnostmi jejich získávání. V optimálním případě *by měl* poskytovatel:

- **dodat datová schémata všech digitálních objektů**, s nimiž je třeba pracovat při sklízení fulltextů, a

- **poskytnout dokumentaci k použití API, podporovaného protokolu** (odklání-li se digitální knihovna od jeho oficiální specifikace) **nebo přímého přístupu do systému**.

Z materiálů, které dodá poskytovatel administrátorovi při zapojování, *musí* být patrné tyto informace:

1. Budou plné texty dokumentů poskytovány po stránkách nebo vcelku?
2. Budou plné texty poskytovány v plaintextu nebo s příměsí tagů nějakého značkovacího jazyka?
3. Existují ve zdroji hierarchicky členěné dokumenty (např. periodikum - ročník - číslo - článek - *stránka* - fulltext nebo monografie - svazek - kapitola - *stránka* - fulltext)? Pokud ano, pro které úrovně jsou vedeny samostatné metadatové záznamy? Jak má být s těmito záznamy zacházeno v CPK? (Viz kapitola 8.3.)
4. Jaký je optimální postup získání plného textu jednoho dokumentu?
5. Jaký je optimální postup pro získání plných textů všech dokumentů ve zdroji?
6. Je nutné při sklízení fulltextů používat různé požadavky pro různé dokumenty (v závislosti na typu dokumentu, struktuře dat apod.)?

7. Existují nějaká omezení na počet požadavků za určitý časový interval nebo na objem dat poskytnutých na základě jednoho požadavku, která poskytovatel nemohl odstranit před zapojováním zdroje do CPK?
8. Lze sklízení plných textů nějak krokovat, ať už v rámci celého zdroje nebo jednoho dokumentu?

8.3 Možnosti zpracování analytických popisných jednotek

Digitální knihovny nezřídka obsahují archivy časopisů nebo sborníků (tj. seriálových publikací), které vycházejí v instituci poskytovatele. Poskytovatel se *musí* domluvit s administrátorem portálu na způsobu zpracování těchto dokumentů s ohledem na zavedená pravidla popisu seriálových publikací v knihovnách, podle kterých lze pracovat se záznamy seriálů jakožto titulů a se záznamy jednotlivých článků, ale nikoliv se záznamy ročníků, čísel nebo svázaných čísel.

Poskytovatel si *musí* zvolit jednu z následujících variant v závislosti na svých technických možnostech a preferencích:

- **Varianta 1:**
  - *Situace:* Digitální knihovna je schopna dodávat metadatové záznamy o celých seriálech, ale ne o samostatných článcích.
  - *Řešení:* U každého seriálu sklidí Správce zdrojů všechny dostupné plné texty, spojí je dohromady a zaindexuje jako jeden fulltext pro daný seriál. Stejný postup se používá při zpracování seriálových dokumentů v digitálních knihovnách Kramerius.

- **Varianta 2:**
  - *Situace:* Digitální knihovna je schopna dodávat metadatové záznamy o článcích, ale ne o celých seriálech (záznamy pro jednotlivé ročníky či čísla jsou pro CPK irelevantní). K metadatovému záznamu každého článku lze přiřadit plný text téhož článku.
  - *Řešení:* V CPK není problém zaindexovat plné texty k záznamům jednotlivých článků. Poskytovatel *by měl* v tomto případě věnovat zvýšenou pozornost správnému zápisu údajů o zdrojovém dokumentu podle *Přílohy 1*, pokud možno i podle *Přílohy 2* a případně i podle *Přílohy 6 Metodiky 1*. Do CPK jsou totiž zapojeny oborové bibliografie obsahující řadu metadatových záznamů o článcích. Je žádoucí, aby bylo možné deduplikovat tyto záznamy se záznamy z fulltextového zdroje.

- **Varianta 3:**
  - *Situace:* Digitální knihovna je schopna dodávat metadatové záznamy o celých seriálech i o samostatných článcích. K metadatovému záznamu každého článku lze přiřadit plný text téhož článku.
  - *Řešení:* Záleží na poskytovateli, zda zvolí řešení pro *Variantu 1*, pro *Variantu 2* nebo pro obě varianty zároveň. V CPK totiž není na překážku zaindexovat tentýž plný text dvakrát: Jednou samostatně jako fulltext článku a podruhé dohromady s dalšími články jako fulltext seriálu.

- **Varianta 4:**
  - *Situace:* Digitální knihovna je schopna dodávat metadatové záznamy o celých seriálech i o samostatných článcích. K metadatovému záznamu každého článku nelze přiřadit plný text téhož článku.
  - *Řešení:* Na fulltexty se uplatní řešení popsané ve *Variantě 1*. Metadatové záznamy článků lze buďto ignorovat, nebo je zpracovat i bez fulltextů, jsou-li dostatečně kvalitní (viz *Varianta 2*).

- **Varianta 5:**
  - *Situace:* Digitální knihovna není schopna dodávat metadatové záznamy o celých seriálech ani o článcích, protože vede záznamy pouze o jednotlivých ročnících, číslech nebo svazcích čísel.
  - *Řešení:* K takovým případům je třeba přistupovat individuálně. V případě vysoké informační hodnoty fulltextů bude administrátor hledat způsob, jak se přiblížit řešení popsanému ve *Variantě 1*.

Tyto varianty se nevztahují jen na seriály, ale také na další dokumenty obsahující samostatně popsané části, kapitoly či články (samozřejmě s vynecháním částí o ročnících, číslech a svázaných číslech).

9 Fulltextové databáze

Možnosti zapojování fulltextových databází jsou velmi různorodé, ale doporučení pro poskytování dat se v základních principech příliš neliší od pokynů popsaných v předcházejících kapitolách. U přistupujících poskytovatelů těchto fulltextových zdrojů se předpokládá schopnost dodávat data přes API, pomocí vhodného protokolu nebo posíláním exportů. Metadata a fulltexty mohou přicházet v různě upravených metadatových formátech nebo ve formátu definovaném poskytovatelem. Klíčovým faktorem pro úspěšné zapojení zdroje je komunikace mezi poskytovatelem a administrátorem portálu.

9.1 Formát fulltextových záznamů

Fulltextové záznamy lze dodávat do CPK ve standardizovaných metadatových formátech nebo v interních formátech poskytovatele:

Z existujících metadatových formátů lze pro tyto účely použít např. Dublin Core rozšířený o pole pro posílání fulltextů (pojmenované třeba `cpk:fulltext`) nebo TEI[25].

U interních formátů se předpokládá, že umožňují předávání metadat v rozsahu potřebném pro popis daných typů dokumentu, a to v podobě umožňující strojové rozpoznání jednotlivých údajů.

V obou případech *by měl* poskytovatel dodat administrátorovi portálu **úplný seznam všech polí, která jsou ve zdroji použita.** V případě dat v metaformátu XML je pro tento účel optimální šablona XSD, u dat zapsaných jako JSON potom JSON Schema. Je-li to možné, poskytovatel *by měl* zpřístupnit administrátorovi portálu i **předpisy, podle nichž jsou tvořeny fulltextové záznamy**. Poskytovatel *by měl* sdělovat také informaci o tom, zda budou fulltexty dodávány v plaintextu nebo s příměsí tagů nějakého značkovacího jazyka.

U plnotextových databází se předpokládá, že dokumenty v nich obsažené obvykle nebývají složeny z více částí s vlastními fulltextovými záznamy. Tzn. **většinou platí, že co dokument, to jeden fulltextový záznam**.

Vyskytuje-li se ve zdroji nezanedbatelné množství dokumentů s daty rozloženými do více fulltextových záznamů, jejich zpracování závisí na možnostech poskytovatele zajistit jedno z následujících řešení:

- **Varianta 1:**
  Každý záznam části dokumentu *musí* mít alespoň jeden výskyt zvláštního opakovatelného pole (např. `cpk:linkage`[26]) pro zápis některého z těchto údajů:
  a) identifikátor záznamu kterékoliv jiné části dokumentu (např. první, předcházející či následující část),
  b) identifikátor nadřazeného záznamu, pokud takový existuje (např. obsah čili seznam částí dokumentu),

[25] Považovat TEI za metadatový formát je nepřesné, ale pro účely tohoto dokumentu dostačující.
[26] Název pole je předmětem domluvy mezi poskytovatelem a administrátorem portálu.

c) jiný společný jmenovatel všech částí, unikátní pro daný dokument v rámci zdroje (např. společný tag či identifikátor takového tagu).

- **Varianta 2:**
  Poskytovatel umožní zpracovat pouze záznam s jednou (nejlépe první) částí dokumentu. Toho docílí tím, že záznamy všech ostatních částí opatří příznakem `cpk0` ve zvláštním poli (dále zvaném např. `cpk:exclude`[27]). Nevýhodou je, že v CPK bude možné vyhledávat jen ve fulltextu první části dokumentu.

- **Varianta 3 (nulová varianta):**
  Poskytovatel zabrání zpracování všech záznamů s částmi dokumentů buďto tím, že neinformuje Správce zdrojů o jejich existenci, nebo je opatří příznakem `cpk0` v poli `cpk:exclude`. Tato varianta je sice radikální, ale nezbytná u zdrojů, které neumožňují řešení *Variantou 1* ani *Variantou 2* a přitom obsahují větší množství dokumentů rozdělených do mnoha částí.

Neprovede-li poskytovatel žádné z řečených opatření, Správce zdrojů zpracuje každý záznam části dokumentu zvlášť, jako by se jednalo o samostatný dokument. Tato varianta je přípustná v případech, kdy se ve zdroji vyskytuje jen málo rozdělených dokumentů, nebo jde alespoň o dokumenty, které jsou rozděleny jen do mála částí.

9.2 Poskytování fulltextových záznamů

Pokyny pro poskytování fulltextových záznamů jsou velmi podobné těm, které již byly popsány v kapitolách 6.6.1 a 8.2 (jde o kombinaci doporučení z těchto kapitol).

Fulltextové záznamy lze do CPK dodávat následujícími způsoby:

- **Varianta 1:**
  Prostřednictvím vhodného protokolu (např. ResourceSync nebo SWORD, pro kratší fulltexty i OAI-PMH). Pokud se implementace protokolu na straně poskytovatele nějak odklání od oficiální specifikace, poskytovatel o tom *musí* informovat administrátora portálu.

- **Varianta 2:**
  Pomocí vlastního API poskytovatele, které *musí* podporovat alespoň minimum funkcí z následujícího výčtu:[28]

  1 **Získání seznamu identifikátorů fulltextových záznamů:**
  Silně doporučeno.
  Není-li podporován bod 2, povinné.

  1.1 **Získání identifikátorů všech fulltextových záznamů:**
  Možnost získat seznam všech fulltextových záznamů ve zdroji bez udání

[27] Název pole je předmětem domluvy mezi poskytovatelem a administrátorem portálu.
[28] Výčet je velmi podobný tomu z kapitoly 6.6.1, ale není stejný. U fulltextových záznamů je totiž preferováno sklízení záznamů „po jednom“ (tj. postupem, kdy se nejprve stáhne seznam identifikátorů a potom jednotlivé záznamy).

dalších parametrů.
V rámci bodu 1 doporučeno.
Není-li v rámci bodu 1 podporován bod 1.2, povinné.

1.2 **Získání identifikátorů fulltextových záznamů vytvořených či upravených v daném časovém rozmezí:**
Možnost získat seznam fulltextových záznamů podle data (zejména kvůli aktualizacím).
V rámci bodu 1 silně doporučeno.
Není-li v rámci bodu 1 podporován bod 1.1, povinné.

1.3 **Stránkování pro oba předchozí body:**
Možnost získávání identifikátorů záznamů v krocích po několika desítkách kusů.
V rámci bodu 1 doporučeno. U rozsáhlejších zdrojů silně doporučeno.

2 **Získání sady fulltextových záznamů:**
Volitelné.
Nejsou-li podporovány body 1 a 3, povinné.

2.1 **Získání všech fulltextových záznamů:**
Možnost stažení všech fulltextových záznamů ve zdroji bez udání dalších parametrů.
V rámci bodu 2 doporučeno.
Není-li v rámci bodu 2 podporován bod 2.2, povinné.

2.2 **Získání fulltextových záznamů vytvořených či upravených v daném časovém rozmezí:**
Možnost stahovat záznamy podle data (zejména kvůli aktualizacím).
V rámci bodu silně doporučeno.
Není-li v rámci bodu 2 podporován bod 2.1, povinné.

2.3 **Stránkování pro oba předchozí body:**
Možnost získávání záznamů v krocích po několika desítkách kusů.
V rámci bodu 2 silně doporučeno.

3 **Získání metadatového záznamu s konkrétním identifikátorem:**
Možnost získat samostatný fulltextový záznam na základě zadání jeho identifikátoru.
Silně doporučeno.
Není-li podporován bod 2, povinné.

4 **Detekce smazaných záznamů:**
Možnost identifikovat záznamy, které byly odstraněny ze zdroje.
Silně doporučeno pro zdroje, ve kterých někdy dochází k mazání záznamů.
Poskytovatel má dvě možnosti, jak toho dosáhnout:

4.1 Umožnit získání seznamu identifikátorů smazaných záznamů, pokud možno i s datem jejich smazání. V optimálním případě umožnit získání tohoto seznamu za určité časové období (kvůli aktualizacím).

4.2 Zachovávat ve zdroji hlavičky smazaných záznamů s označením `deleted` (nebo s jiným předem domluveným příznakem) a s datem jejich smazání (ve

smyslu data poslední úpravy záznamu), aby je bylo možné získat tak, jak je to popsáno v bodu 1.2 nebo 2.2.

Má-li být ke sklízení záznamů použito API, poskytovatel k němu *musí* dodat dokumentaci administrátorovi portálu.

- **Varianta 3:**
  Pokud poskytovatel neumožňuje sklízení fulltextových záznamů prostřednictvím API či protokolu, je třeba **umožnit Správci zdrojů získávat metadata a plné texty přímým přístupem do indexu či databáze fulltextového zdroje**. V tom případě *musí* poskytovatel dodat poskytovateli potřebnou dokumentaci k bezproblémovému stahování dat z daného systému.
  Variantu přímého přístupu lze použít i tehdy, když je získávání fulltextů přes API nebo protokol příliš pomalé.

- **Varianta 4:**
  Dodávání dat posíláním exportů. K této variantě je vhodné se uchýlit, až když není jiná možnost.

U *Variant 1* až *3 musí* poskytovatel informovat administrátora portálu i tehdy, když:

- je potřeba používat různé požadavky pro různé dokumenty (v závislosti na typu dokumentu, struktuře dat apod.) nebo
- existují omezení na počet požadavků za určitý časový interval nebo na objem dat poskytnutých na základě jednoho požadavku, která poskytovatel nemohl odstranit před zapojováním zdroje do CPK.

10 Úložiště plných textů

Do této kategorie spadají všechny fulltextové zdroje, které nebyly popsány v žádné z předchozích kapitol. Tyto zdroje nemají nebo nemohou použít žádné z výše jmenovaných řešení pro dodávání fulltextů a jejich zapojení do CPK s sebou může nést potřebu získávat a zpracovávat data značně komplikovaným postupem. Řadíme sem jednak zdroje s plnými texty dodávanými v ne-plaintextových souborech a jednak webové stránky s obsahem dostupným jen ve formátu HTML. Zájemci o zapojení do CPK *by se měli* těmto způsobům poskytování dat vyhnout, kdykoliv je to možné.

10.1 Úložiště souborů s plnými texty

Fulltexty dokumentů ze zdrojů tohoto typu se vyskytují pouze uvnitř souborů různých formátů, které neobsahují pouze plaintext. Může jít např. o PDF či DjVu soubory s textovou vrstvou, o dokumenty vytvořené v textových procesorech jako MS Word či OpenOffice Writer nebo o elektronické knihy.

Soubory s plnými texty *by mělo* být možné stáhnout ve formátu, ze kterého lze extrahovat plaintext bez nutnosti další konverze (což se týká např. postscriptových souborů nebo formátů typu TEX). Vhodné nejsou ani různé proprietární souborové formáty, s nimiž pracuje pouze software používaný poskytovatelem. Soubory jednotlivých dokumentů *by neměly* být ani komprimovány nebo zaheslovány.

I pro dokumenty s fulltexty v různých souborech *musí* být do CPK dodávána metadata, a to pokud možno některým ze způsobů jmenovaných v kapitole 6.6.1.

V případě samostatného dodávání metadatových záznamů *by mělo* platit, že v metadata budou obsahovat odkazy na soubory, které budou použitelné nejen k zobrazení v CPK, ale také jako informace o umístění fulltextu určeného ke zpracování Správcem zdrojů. **Dodá-li poskytovatel soubory do CPK jinou cestou, *musí* zajistit jejich spárovatelnost s metadatovými záznamy jiným způsobem** (např. unikátními názvy souborů, které jsou uvedeny i v metadatových záznamech).

Nelze-li dodávat metadata samostatně, pak lze využít vlastnosti některých souborových formátů, které umožňují připojit k souboru základní metadata jako je autor, název, datum, popis nebo klíčová slova. Tato možnost přichází v úvahu pouze za předpokladu, že jsou metadata souborů vyplněna důsledně a smysluplně (tzn. autor není nutně osoba, která vytvořila soubor, název nemusí být vždy totožný s názvem souboru či s prvními slovy dokumentu atd.).

Poskytovatel zdroje s metadaty uvnitř souborů *musí* informovat administrátora portálu o všech lokacích, ze kterých má Správce zdrojů stahovat řečené soubory. **Správce zdrojů bude považovat umístění každého souboru zároveň i za adresu, která má být v CPK zobrazována uživateli jako odkaz na dokument.**

Pokud by měl Správce zdrojů získávat soubory obsahující metadata cestou, která není dostupná pro uživatele, pak *musí* poskytovatel dodat odkazy pro uživatele jiným způsobem: Buďto explicitním uvedením URL v metadatech, která jsou součástí souborů, nebo prostřednictvím adresářové struktury, která umožní sestavení URL přidáním jednotné předpony před cestu k souboru (např. ze souboru `dokument.pdf` v adresáři `/2018/01` se vytvoří URL `https://cokoliv.cz/2018/01/dokument.pdf`).

Jak již bylo řečeno v části o formálních náležitostech zapojení, CPK na své straně neukládá soubory dodané poskytovatelem – pouze z nich extrahuje plný text za účelem indexace. Uživatel CPK se k dokumentu dostane vždy jen přes odkaz na soubor uložený u poskytovatele.

10.2 Úložiště plných textů ve formátu HTML

Zájem o zapojení do CPK mohou projevit i správci různých **webových informačních portálů, kde jsou dokumenty představovány jednotlivými HTML stránkami**. Správce zdrojů je schopen procházet obsah stránek poskytovatele podobným způsobem jako crawlery webových vyhledávačů, ale na rozdíl od Googlu nebo Seznamu potřebuje mít k dispozici o něco lépe strukturovaná data. Potřebuje totiž rozpoznat, co má, respektive nemá považovat za samostatný dokument a kde má očekávat metadata popisující jednotlivé dokumenty. K tomu je nutné, aby se poskytovatel a administrátor portálu dohodli na následujících bodech.

1. **Jak bude Správce zdrojů získávat informace o tom, kde se nachází dokumenty ke zpracování?**

   Poskytovatel *by měl* být schopen udržovat jeden či více seznamů s odkazy na dokumenty nebo alespoň s údaji, ze kterých lze tyto odkazy vytvořit přidáním jednotného prefixu.

   V těchto seznamech *by neměly* být zahrnuty odkazy na cokoliv, co nemá být zpracováno jako dokument. Není-li to možné, pak je zajistit rozpoznatelnost různých typů odkazů.

   Stejnou cestou *by mělo* být zjistitelné i datum poslední aktualizace (a ideálně i případného smazání) dokumentu.

   Pro tyto účely lze použít existující sitemapy webu, které buďto obsahují pouze odkazy na dokumenty nebo zahrnují i jiné odkazy, přičemž odkazy na dokumenty lze jednoznačně rozpoznat od ostatních odkazů.

2. **Co bude považováno za identifikátor dokumentu?**

   Identifikátorem *by měl* být údaj, který je nejen unikátní, ale také neměnný.

   Je zřejmé, že v rámci zdroje je vždy unikátní cesta k dokumentu. Předpokládá se tedy, že celá cesta nebo její část může sloužit k tvorbě identifikátoru. K tomuto účelu je vhodné zvolit nejkratší část cesty, o níž je známo, že je unikátní a nebude se měnit.

Demonstrační příklad:
Na webu je vystaveny nové dokumenty s cestami
`/aktualni/zpravy?year=2018&title=jaro` a
`/aktualni/statistiky?year=2018&title=jaro`.
Po roce se dokumenty přesunou do archivu a jsou dostupné z umístění
`/archiv/zpravy?year=2018&title=jaro` a
`/archiv/statistiky?year=2018&title=jaro`.
Vhodnými identifikátory takových dokumentů pak mohou být např. řetězce `zpravy2018jaro` a `statistiky2018jaro`.

3. **Jaký je na webu podíl dokumentů rozdělených do více stránek? Existuje údaj, který by umožnil jejich sloučení v CPK?**

   U zdrojů tohoto typu se předpokládá, že dokumenty v nich obsažené obvykle nebývají složeny z více stránek s vlastními URL. Vyskytuje-li se ve zdroji nezanedbatelné množství dokumentů s daty rozloženými do více stránek, jejich zpracování závisí na možnostech poskytovatele zajistit jedno z řešení popsaných v kapitole 9.1; a to s tím rozdílem, že místo polí `cpk:linkage` lze použít jakýkoliv jiný strojově rozpoznatelný element nesoucí stejnou informaci a místo přidávání polí `cpk:exclude` stačí vyloučit nebo označit dotčené odkazy tak, aby je Správce zdrojů rozpoznal při identifikaci dokumentů ke stažení (viz bod *1*).

4. **Odkud má Správce zdrojů čerpat metadata dokumentu?**

   Poskytovatel *musí* sdělit administrátorovi portálu, které údaje mají být zpracovávány jako metadata dokumentu, kde přesně se nacházejí a jakým způsobem jsou zapisována.

   Je třeba si ujasnit, zda má být za název dokumentu považován titul v HTML hlavičce, hodnota uvnitř prvního elementu `h1` nebo jiný údaj. Aby mohl Správce zdrojů pracovat i s dalšími informacemi (autor, rok, klíčová slova atd.), jejich zápis být *musí* jednotný a rozpoznatelný ve všech dokumentech. Metadata lze čerpat jednak ze samotného HTML kódu a jednak z URL dokumentu.

   V prvním případě lze k identifikaci metadatového údaje použít buďto informaci v HTML tagu (např. cokoliv uvnitř elementu s tagem obsahujícím `class="subtitle"` bude považováno za podnázev dokumentu) nebo informaci z obsahu elementu (např. jméno autora bude vždy zapisováno v samostatném elementu jako `<p>Autor: Jan Novák</p>`).

   Z URL lze čerpat např. metadata o roku vzniku, tématu, jazyku nebo formě dokumentu, pokud to struktura URL dovoluje a poskytovatel dodá administrátorovi portálu kódovník k interpretaci jednotlivých údajů (nejde-li jen o rok). Např. `https://cokoliv.cz/env/2018/annualreports/zprava` by pak mohlo být interpretováno následovně: `env` = téma: životní prostředí, `2018` = rok: 2018, `annualreports` = forma: výroční zprávy.

Srovnání novosti postupů

Základní postupy popsané v metodice vznikaly už od roku 2015, kdy se do CPK zapojily první knihovny. Až nyní však byla zpracována podrobná dokumentace, kterou lze použít jako průvodce zapojením fulltextového zdroje do CPK i jako referenční materiál pro poskytovatele již zapojených zdrojů. Metodika se vztahuje na všechny typy zapojovaných fulltextových zdrojů, které mohou být po technické i obsahové stránce značně různorodé. Proto jsou pokyny metodiky formulovány spíše obecně a neobsahují přesný popis kroků vedoucích k zapojení zdroje. Výjimkou jsou postupy uplatňované při zapojování digitálních knihoven Kramerius, které bylo možno popsat konkrétně.

V metodice jsou obsažena jednak obecná doporučení, která se týkají všech zapojovaných zdrojů, a jednak specifické pokyny pro jednotlivé typy zdrojů. Požadavky jsou v metodice odstupňovány podle závaznosti – od povinných, přes doporučené, až po volitelné. V metodice převažují nezávazná doporučení nad závaznými nařízeními. Nepovinné pokyny jsou proto formulovány s důrazem na pochopitelnost výhod jejich dodržování.

Tento dokument navazuje na *Metodiku pro zapojování metadatových zdrojů do CPK* a odkazuje se na ni v kapitolách věnovaných metadatům a způsobům dodávání dat. Nejsou zde tedy znovu podrobně rozebírána témata, která již byla řešena v metodice pro metadatové zdroje.

Metodika inovuje způsob, jakým se poskytovatelé fulltextových zdrojů dostávají k informacím potřebným k zapojení do CPK. Poskytovatelé dosud neměli k dispozici oficiální dokumentaci věnovanou problematice dodávání fulltextů a detaily o podmínkách zapojení se dozvídali výhradně přímou komunikací s administrátorem portálu. Používání metodiky by tedy mělo zefektivnit proces zapojování nových poskytovatelů fulltextových zdrojů, kteří potřebují vodítko pro přípravu svých dat k efektivnímu zpracování Správcem zdrojů.

Popis uplatnění metodiky

Metodika je určena především poskytovatelům fulltextových zdrojů zapojovaných do CPK. Metodiku využijí rovněž správci fulltextových zdrojů, kteří své zapojení do CPK zatím jen zvažují – dokument jim pomůže zhodnotit vlastní technickou připravenost na tento proces. K potenciálním i aktuálním poskytovatelům fulltextových zdrojů se řadí jednak knihovny provádějící digitalizaci svého fondu a jednak správci digitálních knihoven, repozitářů a dalších zdrojů přínosných pro uživatele CPK.

K metodice lze přistupovat jako k sadě doporučení, z nichž většina není pro poskytovatele fulltextových zdrojů povinná. Čím více se bude poskytovatel držet uvedených pokynů, tím lépe bude možné zpracovat jeho data pro snadné nalezení uživatelem CPK. Do CPK se zapojují zdroje s již existujícími fulltexty a metadaty se strukturou vycházející z původních potřeb poskytovatele. Metodika má sloužit poskytovatelům především jako orientační příručka, která jim pomůže přizpůsobit data posílaná do CPK do podoby umožňující zpracování Správcem zdrojů.

Postupy uváděné v metodice byly implementovány při zapojování digitalizovaného fondu Moravské zemské knihovny v Brně a informačního webu Zákony pro lidi. Metodika také do značné míry vychází ze zkušeností řešitelského týmu CPK s metadatovými zdroji obsahujícími záznamy s delšími textovými úryvky. Způsob dalšího uplatnění řečených postupů závisí na charakteru nově zapojovaných zdrojů.

Výsledky popisovaného řešení byly prezentovány na konferencích a v odborné literatuře, nikoliv však separátně, ale vždy jako jeden z aspektů vývoje a provozu CPK. CPK včetně Správce zdrojů je vyvíjen jako open source software a od 26. října 2016 je v plném provozu pro uživatele na adrese www.knihovny.cz.

Použitá a související literatura

Bibliotek.dk [online]. Ballerup: Danish Bibliographic Centre [cit. 2018-09-03]. Dostupné z: https://bibliotek.dk/eng.

Bibliotheek.nl [online]. Haag: Koninklijke Bibliotheek [cit. 2018-09-03]. Dostupné z: https://www.bibliotheek.nl.

BRADNER, Scott. RFC 2119: Key words for use in RFCs to Indicate Requirement Levels. In: *The Internet Engineering Task Force (IETF®)* [online]. 1997 [cit. 2018-08-31]. Dostupné z: https://tools.ietf.org/html/rfc2119.

CPK - Využití sémantických technologií pro zpřístupnění kulturního dědictví prostřednictvím Centrálního portálu knihoven. In: *Informační systém výzkumu, experimentálního vývoje a inovací* [online]. RVVI, 2016 [cit. 2018-08-29]. Dostupné z: https://www.rvvi.cz/cep?s=jednoduche-vyhledavani&ss=detail&n=0&h=DG16P02R006.

Česká digitální knihovna [online]. Praha: Knihovna AV ČR [cit. 2018-09-03]. Dostupné z: https://www.czechdigitallibrary.cz/cs/.

DCMI: Dublin Core Metadata Initiative [online]. ASIS&T, [cit. 2018-08-29]. Dostupné z: http://dublincore.org/.

Europeana Data Model Documentation. In: *Europeana Pro* [online]. Posted on Tuesday November 18, 2014 [cit. 2018-08-31]. Dostupné z: https://pro.europeana.eu/resources/standardization-tools/edm-documentation.

Europeana Semantic Elements Documentation. In: *Europeana Pro* [online]. Posted on Thursday December 4, 2014 [cit. 2018-08-31]. Dostupné z: https://pro.europeana.eu/page/ese-documentation.

Finna.fi: The material of Finnish archives, libraries and museums with a single search [online]. Helsinki: National Library of Finland [cit. 2018-09-03]. Dostupné z: https://finna.fi/?lng=en-gb.

HUTAŘ, Jan, Pavlína KOČIŠOVÁ, Natalie OSTRÁKOVÁ, Zdeněk VAŠEK, Iveta LODROVÁ, Pavla ŠVÁSTOVÁ a Jaroslav KVASNICA. *Definice metadatových formátů pro digitalizaci monografických dokumentů: monografií, kartografických dokumentů, hudebnin* [online]. Verze 1.3.1. Praha: NK ČR, 2018 [cit. 2018-08-31]. Dostupné z: https://www.ndk.cz/standardy-digitalizace/dmf_monografie_1-3-1.

HUTAŘ, Jan, Pavla ŠVÁSTOVÁ, Pavlína KOČIŠOVÁ, Natalie OSTRÁKOVÁ, Iveta LODROVÁ a Jaroslav KVASNICA. *Definice metadatových formátů pro digitalizaci periodik* [online]. Verze 1.7.1. Praha: NK ČR, 2018 [cit. 2018-08-31]. Dostupné z: https://www.ndk.cz/standardy-digitalizace/dmf_periodika_1-7-1.

Jak se zapojit. In: *Knihovny.cz* [online]. Brno: MZK [cit. 2017-09-11]. Dostupné z: https://www.knihovny.cz/Portal/Page/jak-se-zapojit.

Knihovny.cz [online]. Brno: MZK [cit. 2018-08-20]. Dostupné z: https://www.knihovny.cz.

Koncepce rozvoje knihoven ČR na léta 2011 – 2015 včetně internetizace knihoven: Knihovny pro EVROPU 2020 [online]. Praha: ÚKR ČR, 2012 [cit. 2018-08-30]. Dostupné z: http://files.ukr.knihovna.cz/200000077-a8cc8a9c7b/Koncepce_PIK_Rozp.doc.

Kramerius. *GitHub.com* [online]. [cit. 2018-07-31]. Dostupné z: https://github.com/ceskaexpedice/kramerius/wiki.

KTD: Česká terminologická databáze knihovnictví a informační vědy (TDKIV) [online databáze]. Praha: Národní knihovna České republiky, 2003- [cit. 2018-08-31]. Dostupné z: http://aleph.nkp.cz/cze/ktd.

KURFÜRSTOVÁ, Jana, Petr ŽABIČKA a Petra ŽABIČKOVÁ. *Metodika pro zapojování metadatových zdrojů do Centrálního portálu knihoven* [online]. Brno: MZK, 2017 [cit. 2018-07-13]. Dostupné z: http://invenio.nusl.cz/record/373491/files/nusl-373491_1.pdf.

LAGOZE, Carl, Herbert VAN DE SOMPEL, Michael NELSON a Simeon WARNER. The Open Archives Initiative Protocol for Metadata Harvesting. In: *Open Archives Initiative* [online]. Ithaca: Cornell University Library, 2002, Document Version 2015-01-08 [cit. 2018-08-20]. Dostupné z: https://www.openarchives.org/OAI/openarchivesprotocol.html.

LHOTÁK, Martin. Česká digitální knihovna. *Duha* [online]. 2016, **30**(3) [cit. 2018-07-31]. ISSN 1804-4455. Dostupné z: https://duha.mzk.cz/clanky/ceska-digitalni-knihovna.

MODS: Metadata Object Description Schema [online]. Washington: Library of Congress, [cit. 2018-08-31]. Dostupné z: http://www.loc.gov/standards/mods/.

Obálky knih [online]. JVK České Budějovice, MZK [cit. 2018-09-03]. Dostupné z: https://obalkyknih.cz/.

P5: Guidelines for Electronic Text Encoding and Interchange. In: *TEI: Text Encoding Initiative* [online]. Version 3.4.0. Last updated on 23rd July 2018, revision 1fa0b54 [cit. 2018-08-31]. Dostupné z: http://www.tei-c.org/Vault/P5/current/doc/tei-p5-doc/en/html/.

Přehled institucí zapojených do registru digitalizace [online]. In: *Registr digitalizace*. [cit. 2018-08-20]. Dostupné z: http://registrdigitalizace.cz/rdcz/info/prehled_instituci.

ResourceSync Framework Specification - Table of Contents. In: *Open Archives Initiative* [online]. Ithaca: Cornell University Library, 22 February 2017 [cit. 2018-08-31]. Dostupné z: http://www.openarchives.org/rs/toc.

Swissbib [online]. Basel: Universität Basel [cit. 2018-09-03]. Dostupné z: https://www.swissbib.ch/.

SWORD [online]. JISC [cit. 2018-08-31]. Dostupné z: http://swordapp.org/.

Trove [online]. Canberra: National Library of Australia [cit. 2018-09-03]. Dostupné z: http://trove.nla.gov.au.

Zapojené knihovny a zdroje. In: *Knihovny.cz* [online]. Brno: MZK, Aktualizováno 11. září 2017 [cit. 2018-09-03]. Dostupné z: https://www.knihovny.cz/Portal/Page/zapojene-knihovny-a-zdroje.

Zapojte se. In: *Česká digitální knihovna*. Dostupné z: https://www.czechdigitallibrary.cz/cs/zapojte-se/.

Seznam publikací předcházejících metodice a další výstupy

Správce zdrojů – GitHub repozitář. In: *GitHub* [online]. Moravská zemská knihovna [cit. 2018-09-05]. Dostupné z: https://github.com/moravianlibrary/RecordManager2.

CPK – GitHub repozitář. In: *GitHub* [online]. Moravská zemská knihovna [cit. 2018-09-05]. Dostupné z: https://github.com/moravianlibrary/CPK.

KURFÜRSTOVÁ, Jana, Petr ŽABIČKA a Petra ŽABIČKOVÁ. *Metodika pro zapojování metadatových zdrojů do Centrálního portálu knihoven* [online]. Brno: MZK, 2017 [cit. 2018-07-13]. Dostupné z: http://invenio.nusl.cz/record/373491/files/nusl-373491_1.pdf.

STOKLASOVÁ, Bohdana, Petr ŽABIČKA, Petra ŽABIČKOVÁ, Lenka MAIXNEROVÁ, Jan POKORNÝ, Karolína KOŠŤÁLOVÁ a Martin LHOTÁK. *Knihovny.cz – Centrální portál českých knihoven: Projektový záměr verze 5*. 2016 [cit. 2018-09-05]. Dostupné také z: https://www.knihovny.cz/Portal/Page/projektove-zamery.

STOKLASOVÁ, Bohdana, Petr ŽABIČKA, Iva BUREŠOVÁ, Pavlína DOLEŽALOVÁ, Karolína KOŠŤÁLOVÁ, Martin LHOTÁK, Pavlína LONSKÁ a Jaroslav MEIXNER. *Knihovny.cz: Centrální portál* českých *knihoven: Projektový záměr verze 4*. 2014 [cit. 2018-09-05]. Dostupné také z: https://www.knihovny.cz/Portal/Page/projektove-zamery.

Technické požadavky na zapojení knihovny do portálu [online]. Verze 2.0. Brno: MZK, 2015 [cit. 2018-09-05]. Dostupné z: https://goo.gl/sDV4Zb.

ŽABIČKOVÁ, Petra a Michal MERTA. Portál knihovny.cz stručně a jednoduše. *Duha* [online]. 2015, **29**(4) [cit. 2018-09-05]. ISSN 1804-4255. Dostupné z: http://duha.mzk.cz/clanky/portal-knihovnycz-strucne-jednoduse.

ŽABIČKOVÁ, Petra a Petr ŽABIČKA. Knihovny.cz - Discovery portal for Czech libraries = Knihovny.cz - Centrální portál pro české knihovny. *Libraries V4* in *the Decoy of Digital Age: proceedings of the 6th Colloquium of library and information experts of the V4+ countries held from 31st May - 1st June 2016 in Brno*. Brno: Moravian Library in Brno, 2016. s. 295-305. ISBN 978-80-7051-216-6.

ŽABIČKA, Petr, ŽABIČKOVÁ, Petra, KRAVEC, Martin. Knihovny.cz: spuštění se blíží [online]. In: *Sborník INFORUM*, 2016 [cit. 2018-09-05]. ISSN 1801-2213. Dostupné z: http://sdruk.mlp.cz/data/xinha/sdruk/2015/knihovny_soucasnosti_2015.pdf.

ŽABIČKA, Petr, Petra ŽABIČKOVÁ, a Martin KRAVEC. Knihovny.cz - Začínáme. *Knihovny současnosti 2016* [online]. Praha: Sdružení knihoven ČR, 2016. s. 322-330 [cit. 2018-09-05]. ISBN 978-80-86249-80-3. Dostupné z: http://sdruk.mlp.cz/data/xinha/sdruk/2016/KKS/sbornik/Knihovny_soucasnosti_2016.pdf.