Příjem, validace a ukládání dat z projektu VISK 7

Dobrá praxe a nejčastější chyby

Jan Bilwachs
Národní knihovna ČR
Oddělení správy a archivace digitálních dokumentů

Příjem dat z projektu VISK 7 v číslech

- **Ročně se přijímá 15-18 projektů**
- **Podstatná většina je předávána na přelomu kalendářního roku (prosinec-leden)**
- **Ročně se reklamuje 5-8 projektů**
- **Celkově (nové dodávky + opravené reklamace) se tedy jedná o přibližně 20-26 dodávek dat ke kontrole a zpracování**

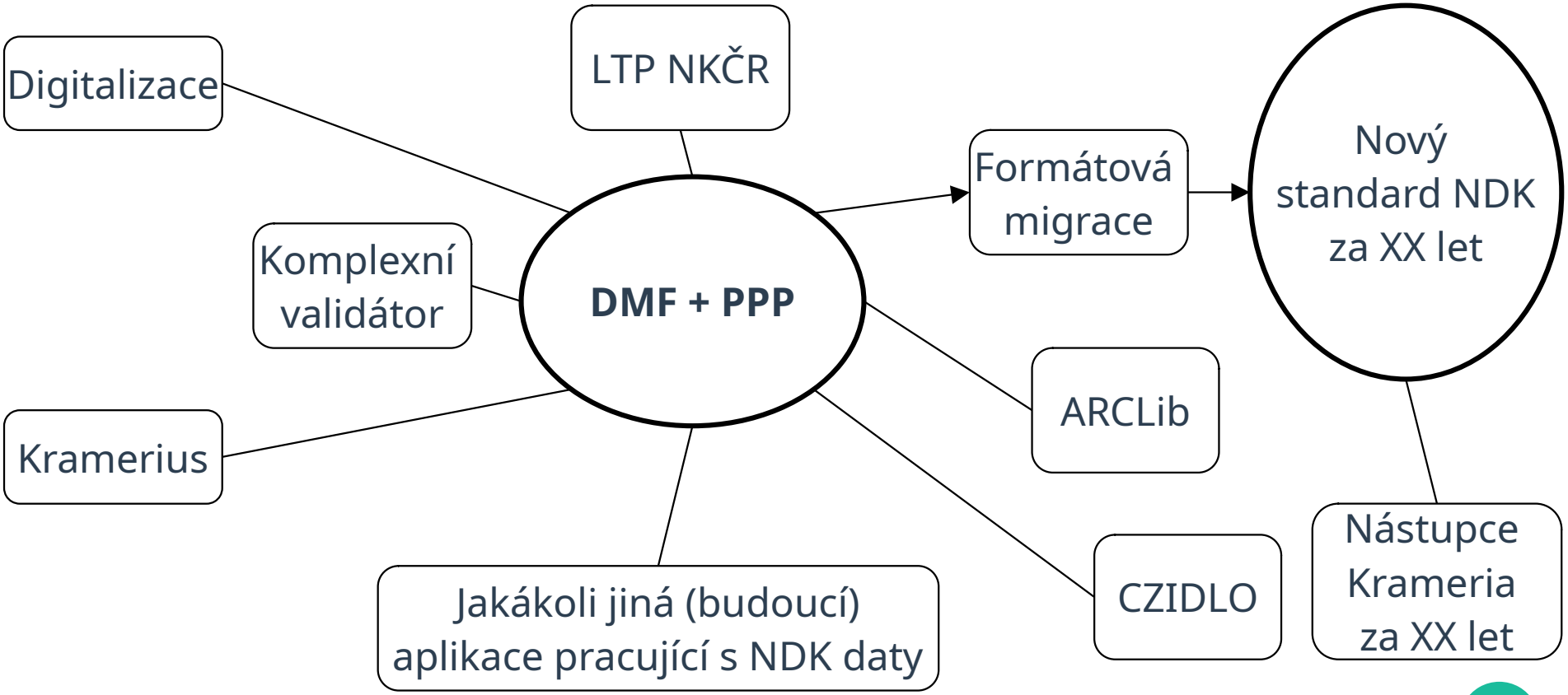
Proces příjmu dat

1) **Předání dat (HDD, FileSender, sFTP …)**
2) **Antivirová kontrola + kontrola integrity**
    a) Údaje v předávacím protokolu
    b) Seznam urn:nbn
    c) Kontrolní součty md5
3) **Kontrola formátu a obsahu**
    a) Plošná kontrola Komplexním validátorem
    b) Výpis popisných metadat a jejich kontrola
    c) Manuální kontrola reprezentativních vzorků (různé tituly, přílohy, interní součásti)
    d) Záznam v RD a u periodik také kontrola titulových UUID
4) **Testovací import (kvůli kompatibilitě s prostředím NDK)**
5) **Import do LTP a Krameria**
6) **Kontrola úspěšnosti importu**

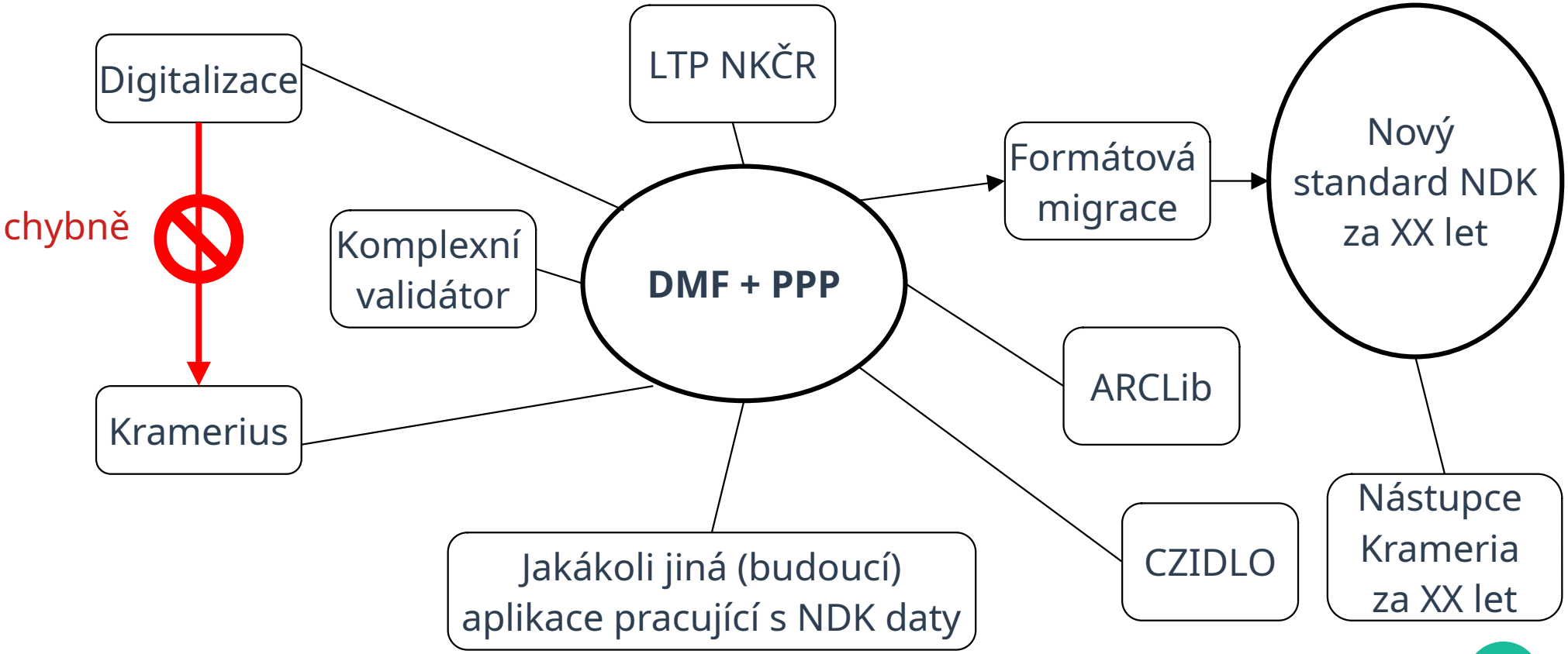
Úloha Standardů NDK, aneb proč je nutné je dodržovat

- **Standardy NDK:**
  - Definice metadatových formátů (DMF)
  - Pravidla pro popis (PPP)
- **Jejich dodržování je klíčové pro dlouhodobou logickou ochranu dokumentů, jejich čtení, správu, migraci …**
- **Jejich dodržování je závazné pro všechny zúčastněné strany**
  - Producenti, vývojáři aplikací, knihovníci

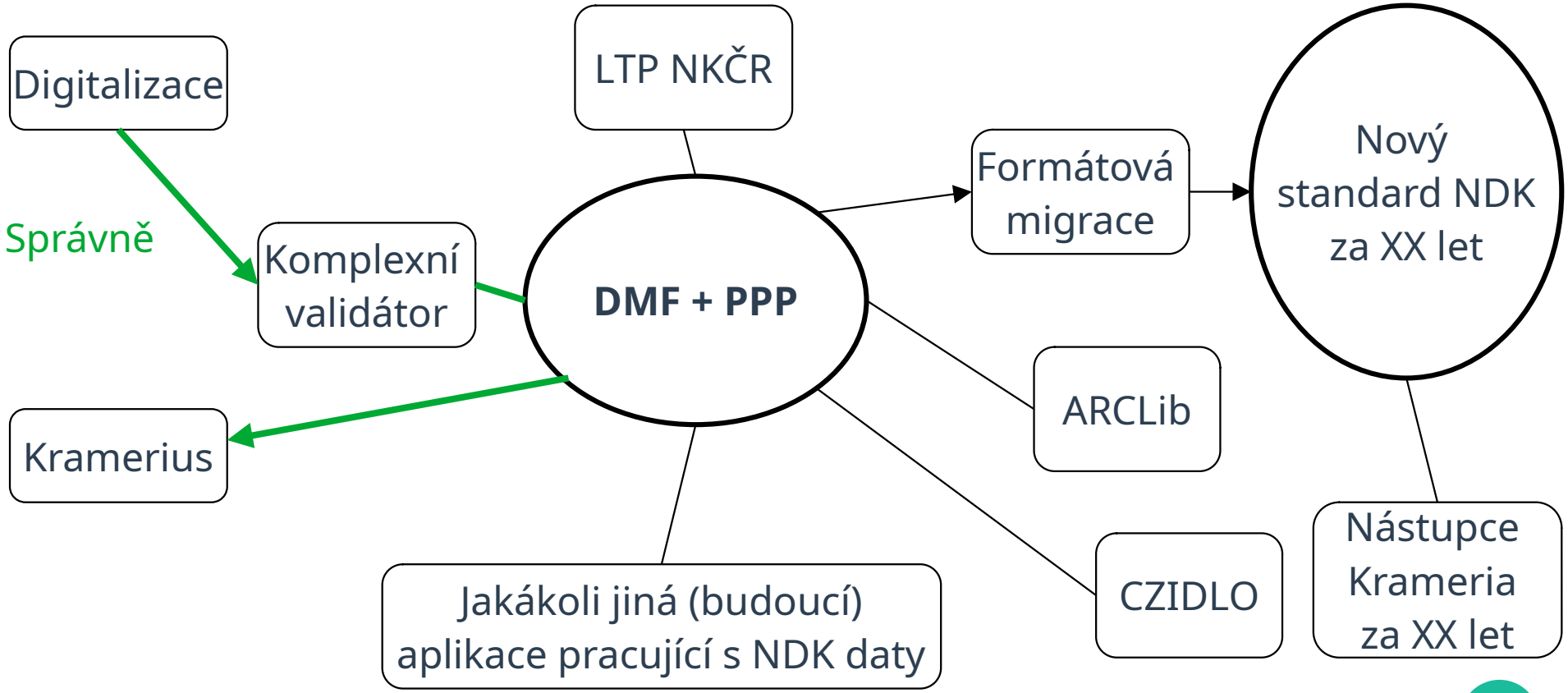
Úloha Standardů NDK, aneb proč je nutné je dodržovat



Úloha Standardů NDK, aneb proč je nutné je dodržovat



Úloha Standardů NDK, aneb proč je nutné je dodržovat



Nejčastější chyby

Narušená integrita balíčku

- **Nesouhlasící kontrolní součty md5**
  - Při výpočtu md5 je nutné brát do úvahy i md5 v souborech mets, amd_mets a info
  - Dodržení správně posloupnosti při výpočtu md5
- **Chybně uvedené velikosti souborů v mets**
- **Chybějící či nadbytečné soubory v NDK balíčcích**
- **Nerespektování XML syntaxe**

Nerespektování názvové konvence balíčku

- **mets_cnb000752014.xml – chybně, čćnb není unikátní, platí pro celý titul**
- **mets_abc123-000frs.xml – správně, urn:nbn je unikátní, reprezentuje danou intelektuální entitu, pochází ze základní úrovně dokumentu**
- **Do úvahy přichází použití:**
  - urn:nbn základní úrovně
  - UUID základní úrovně
- **Více viz: DMF, 6. kapitola**

Nepopsání titulu periodika (popis pod jiným titulem)

- **Často se jedná o přívazky v periodikách a konvoluty**
  - Vlastní bibliografické údaje (titul, vydavatel, ročníková řada…)
  - Vlastní ČČnb a issn
  - Vlastní katalogizační záznam
- **Pokud kat. záznam neexistuje, měl by se prvně před digitalizací vytvořit.**
  - Viz také: *Metodika pro tvorbu balíčků SIP se zaměřením na digitalizáty tištěných dokumentů*, 2019, s. 58.
- **Tituly „schované“ pod jiný titul jsou v dlouhodobém úložišti nedohledatelné → závažný problém nutný k řešení**

Nepopsané (nesprávně popsané) přílohy

- **Typy častých chyb:**
  - Chybné zanoření v logických strukturálních mapách
  - Chybějící metadata dmdSec s popisem přílohy
  - Příloha má vlastní číselnou řadu ročníků, avšak přesto je přidružena k některému číslu a není popsána jako samostatný titul
- **Pravidla pro popis periodik, 3. kapitola:**
  - Příloha, která se neskenuje (CD/DVD, pohlednice, plakát apod.)
  - Příloha, která se skenuje spolu s číslem (podobného typu, tvaru a velikosti jako popisované číslo)
  - Příloha, která se skenuje zvlášť (odlišného typu, tvaru a velikosti + vlastní číselná řada)
- **Vztah přílohy jakožto samostatného titulu a rodičovského titulu periodika lze popsat pomocí elementu „relatedItem“**

Nevhodné údaje v elementu „partNumber“ (číslo, ročník)

- **Definice elementu partNumber dle DMF pro periodika:**
  - „Pořadové číslo vydání (čísla), např. 40; u ročenek číslo řady/edice“
- **Nesprávné hodnoty:**
  - „40 (mimořádné číslo k výročí vzniku samostatného československého státu)“
    - Část v závorce přijde do elementu „partName“
  - „[2]“
    - Nejisté nebo chybné číslování se vyjádří v elementu „note“
  - Příloha ročníku
    - Bude v elementu „partName“
- **Správné hodnoty:**
  - 40
  - 21-22
- **Více viz: Pravidla pro popis periodik, kapitola 2.3**
- **Každý ročník se popisuje samostatně, pakliže záměrem vydavatele není dvoj- nebo víceročník (PPP, kap. 4.1)**

Nevhodné hodnoty v elementu dateIssued (úroveň čísla a ročníku)

Nesprávné hodnoty (číslo):

- [1950]
- 19??
- 1989-90
- 01.031989
- 2.6.1954
- 1954-1955-1956-1957-1958

Předepsané vzorce v PPP (číslo):

- DD.MM.RRRR
- MM.RRRR
- RRRR
- RRRR-RRRR
- MM.-MM.RRRR
- MM.RRRR-MM.RRRR
- DD.MM.-DD.MM.RRRR
- DD.MM.RRRR-DD.MM.RRRR
- DD.-DD.MM.RRRR

- **Na úrovni ročníku se zapisuje pouze rok (RRRR) nebo rozmezí let (RRRR-RRRR)**
- **Nejisté (odvozené) datum vydání je vhodné (namísto závorek či otázníků) vyjádřit následovně:**
  - <mods:dateIssued qualifier="approximate">10.06.1946</dateIssued>

Kontrolované slovníky a jejich dodržování

- **Sigly, typy stran, hodnoty v mods:form …**
- **Nutné dodržovat i malá a velká písmena**
- **Např u sigel přebíraných z databáze ADR:**
  - ABA001 (správně)
  - aba001 (špatně)

Další časté chyby a nevhodné praxe

- **Nepodporovaná verze MODS pro danou verzi DMF (viz DMF kap. 1.4)**
- **Identifikátor deaktivovaný v MODS je ponechán v DC beze změny**
  - (typicky issn)
- **Nerespektování metodiky resolveru**
  - Při úpravách signifikantních metadat ukládaných v resolveru je nutné postupovat v souladu s metodikou resolveru
- **Obsah elementu originInfo neodpovídá typu katalogizačního záznamu (RDA nebo AACR)**
- **Pořadová čísla v identifikátorech a názvech souborů nezačínají od „1“, např.:**
  - mc_abc123-000fwe_0043.jp2 (jako první obrázek)
  - MODS_ISSUE_0032 (úroveň čísla může být v balíčku pouze jedna)
- **Neúplná hodnota atributu LABEL v kořenovém elementu souboru mets**

Doporučení k elementu „location“

U periodik je vhodné uvádět údaj o lokaci předlohy (mods:location) nejen na úrovni titulu, ale i na základní úrovni (čísla nebo přílohy)

Kritéria pro (ne)přijetí dodávky

Závažné chyby (k reklamaci):

- Narušená integrita
- Chyby (neúplnosti) v popisných metadatech
- Nerespektování metodiky resolveru
- Nevalidní obrázky
- Nejednoznačné nebo chybějící údaje důležité při dlouhodobém uchování (např. o použité verzi MODS, DMF…)

„Měkké“ chyby (k toleranci):

- (Nadbytečné elementy nepopsané v DMF, pakliže nerozporují jiné povinné elementy a nejsou v rozporu s Library of Congress)
- Case sensitivita (malá a velká písmena)
- UUID na titulové úrovni se neshoduje s předchozí téhož titulu

Dostupné nástroje

Dostupné nástroje – Komplexní validátor

- **Kontroluje zejména:**
  - Formát metadat, (ne)přítomnost elementů, jejich umístění
  - Formát identifikátorů
  - Validitu obrázků pomocí externích nástrojů
  - Kontrolované slovníky
- **Co Komplexní validátor neodhalí:**
  - Chyby v popisných metadatech, nesoulad s údaji v předloze, překlepy

Další dostupné nástroje

- **mets2xlsx**
  - Skript vypisující hodnoty z hlavních metadatových souborů (mets) do excelové tabulky
  - Snaha o částečné vyplnění limitů Komplexního validátoru při kontrole popisných metadat
  - Snaha o zlepšení přehlednosti nad velkým množstvím dokumentů
- **validatorLog2html**
  - Skript, který seskupuje výsledky Komplexního validátoru z množství XML logů do jednoho přehlednějšího html souboru
- **Skripty jsou zveřejňovány zde: https://github.com/NLCR/visk7-nastroje**

Závěrem

- **Kontrola se vyplatí nejen u ručně vytvářených metadat, ale i v případě metadat vytvářených z katalogizačního záznamu**
- **Prosba o kontrolu dat ještě před odesláním do NK**
- **V případě zjištěných problémů s Komplexním validátorem prosíme o využití „issues“ na platformě GitHub**
  - https://github.com/NLCR/komplexni-validator/issues

Děkuji za pozornost!

Email: Jan.bilwachs@nkp.cz