

Projekt Vytvoření Národní digitální knihovny



V únoru 2010 podala Národní knihovna ČR (NK ČR) společně s Moravskou zemskou knihovnou (MZK) jako partnerem Projekt „Vytvoření Národní digitální knihovny“ (NDK). Projekt byl podán v rámci Výzvy 07 Integrovaného operačního programu „Elektronizace služeb veřejné správy“. V červnu 2010 byl projekt schválen. Jedná se o jeden ze základních stavebních kamenů konceptu eCulture, kterým sektor kultury významně přispívá k naplňování cílů Smart Administration.

Projekt NDK je financován z Integrovaného operačního programu EU částkou 255 milionů korun a spolufinancován z rozpočtu MK ČR částkou 45 milionů korun. Příspěvek 85% ze strukturálního fondu ERDF ve výši 254 946 300 Kč je doplněn 15% kofinancováním ze státního rozpočtu ve výši 44 990 700 Kč. Celkové způsobilé veřejné výdaje tedy činí 299 937 000 Kč.

NK ČR a MZK uchovávají ve svých fondech díky právu úplného povinného výtisku většinu monografií, periodik a dalších druhů dokumentů publikovaných na našem území (bohemika v užším slova smyslu), velké množství dokumentů vztahujících se k ČR publikovaných v zahraničí (bohemika v širším slova smyslu) a spravují bohaté historické fondy. Od roku 2000 spolupracují i na podchycení českého webu. Disponují tedy rozsáhlým a zároveň unikátním materiálem jedinečné kulturní, ale s ohledem na kontext Smart Administration především faktografické hodnoty.

Projekt NDK má tři hlavní cíle:

- 1. Digitalizace významné části bohemikální produkce 19.-21. století,** tj. knih vydaných na území České republiky, napsaných v češtině nebo pojednávajících o Česku. Celkem do konce roku 2019 zdigitalizujeme více než 50 milionů stran, tedy přibližně 300 000 svazků. Dosah projektu není zdaleka omezen jen na dobu jeho trvání, ale bude intenzivně pokračovat i po jeho ukončení v roce 2014 – nejen do roku 2019 v rámci povinné udržitelnosti projektu, ale i v dalších letech.
- 2. Dlouhodobé uložení dokumentů ve spolehlivém digitálním úložišti.** Úložiště poskytne prostor pro bezpečné umístění dosud digitalizovaných dokumentů i digitálních dokumentů vytvořených či získaných v projektu NDK i v rámci dalších projektů.
- 3. Zpřístupnění digitálních dokumentů v jednotném, uživatelsky vřídlném rozhraní** s vysokou mírou možné personalizace. Z jednoho místa budou přístupné digitalizované dokumenty i placené online databáze.

Systém NDK

Systém NDK tvoří čtyři základní subsystémy a řada subsystémů podpůrných. Všechny subsystémy budou vybudovány nad základní technologickou infrastrukturou, zahrnující servery, úložné prostředky, síť a základní řídicí prostředky. Tato infrastruktura vznikne rozšířením existujícího technologického prostředí provozovaného již nyní v NK ČR a MZK.

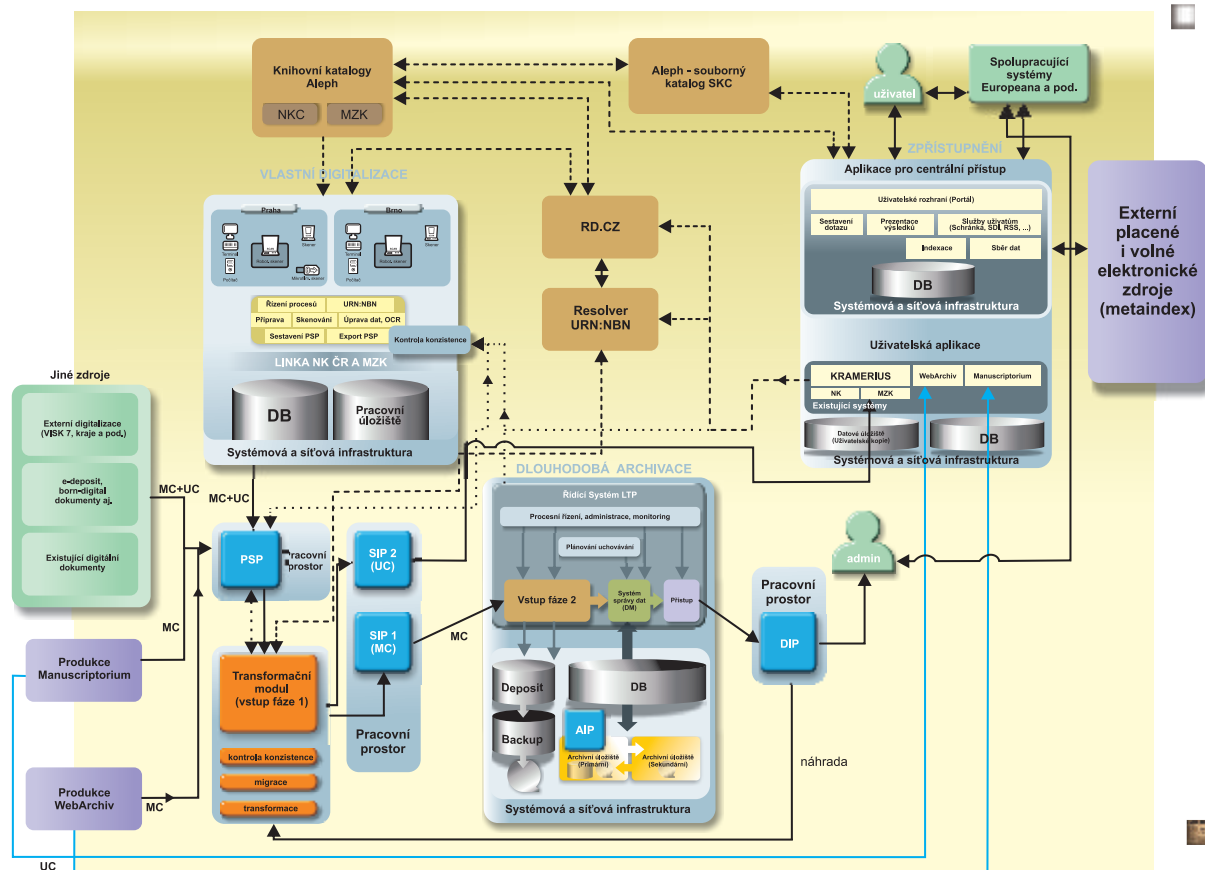


Základními subsystemy jsou:

- Subsystem digitalizace
- Subsystem dlouhodobého uložení dokumentů (LTP subsystem)
- Subsystem pro transformace a kontroly konzistence (transformační modul)
- Subsystem zpřístupnění informací a dokumentů (současné aplikace zpřístupnění + aplikace pro centrální přístup)

Součástí systému jsou technické a administrativní nástroje zajišťující propojení subsystemů mezi sebou a jejich integraci se stávajícími aplikacemi (integrační vrstva).

Role jednotlivých subsystemů v rámci systému NDK i jejich vzájemné vazby zachycuje následující grafické znázornění.



Legenda k obrázku:

- Digitalizace, Dlouhodobá archivace a Aplikace pro centrální přístup jsou moduly, kde existují hotová řešení (na základě požadavků ve výběrovém řízení bude vybráno řešení nejvhodnější)
- Transformační modul je dosud neexistující komponenta, kterou je třeba v rámci projektu NDK vyvinout (případně doprogramovat její funkcionality k použitelnému produktu)
- Systémy, které již existují (Aleph, Registr digitalizace, URN:NBN resolver, Kramerius, WebArchiv, Manuscriptorium) je třeba do systému NDK integrovat a propojit s nově budovanými systémy a moduly

MC = archivní kopie (master copy)

UC = uživatelská kopie (user copy)

PSP = Producer submission package – balíček dat a metadat z procesu digitalizace nebo dodaný z externích zdrojů

SIP = Submission information package – balíček dat a metadat vstupující do LTP systému

DIP = Dissemination information package – balíček dat a metadat vystupující z LTP systému

AIP = Archival information package – balíček dat a metadat v archivu

DB = databáze

Typy čar:

– plná čára - toky balíčků (data + metadata)

– přerušovaná čára - toky metadat, předání identifikátorů

– tečkovaná čára - kontroly záznamů

– modrá čára - uživatelské kopie projektů Manuscriptorium + WebArchiv (jdou mimo, zajišťuje se jen archivování MC z těchto projektů)

Postup dokumentů systémem NDK

(základní popis k obrázku)

- Dokument vybraný pro digitalizaci prochází zpracováním v subsystému digitalizace s použitím nástrojů tzv. „digitalizačního workflow“; metadata se získávají z knihovního systému Aleph pomocí načtení čárového kódu a předávají se do digitalizačního workflow a do Registru digitalizace (dále RD); digitalizační workflow také přiděluje dokumentům identifikátory (URN:NBN), které dále spravuje aplikace Resolver URN:NBN.
- Výstupem procesu digitalizace je balíček dat PSP, který obsahuje jak data a metadata pro zpřístupnění, tak data a metadata určená k archivaci. Tento balíček je vložen do sdíleného pracovního prostoru.
- V pracovním prostoru balíčky PSP dále zpracovává transformační modul. Jsou zkontrolována metadata a vytvořeny balíčky SIP1 pro LTP systém a SIP2 pro zpřístupňující systém (transformací metadat a struktury balíčků). Transformační modul přiděluje URN:NBN externím dokumentům, které nepřicházejí z digitalizačního workflow.
- SIP1 je dále zpracován v LTP systému: výstupem je balíček AIP a jeho uložení do archivního modulu LTP systému.
- Aplikace Kramerius (v NK ČR i MZK) zpracuje balíček SIP2 a zajistí zpřístupnění uživatelských kopií (UC). V aplikacích Kramerius budou umístěny uživatelské kopie vyprodukované oběma pracovišti digitalizace (Praha, Brno) i uživatelské kopie externích dat.
- Digitalizační workflow kontroluje konzistenci – zda UC i MC skončily v cílovém místě – a zajistí smazání PSP balíčků z pracovního prostoru.
- Systémy RD a Resolver URN:NBN sklízí informace z aplikací Kramerius a RD posléze poskytuje URL uživatelské kopie knihovnímu katalogu Aleph (NKC, MZK), odkud se předávají dále do SKC.
- Aplikace pro centrální přístup sklízí data ze zpřístupňujících aplikací (Kramerius, Manuscriptorium, WebArchiv), odkud uživatel získává i obrazová data a full texty, případně rozšířený popis.
- V případě potřeby nahradit uživatelské kopie jsou archivní data vyexportována přes nastavitelný DIP do transformačního modulu a jsou vložena do aplikací pro zpřístupnění (migrace metadat i dat může proběhnout v rámci LTP systému a jeho workflow pro DIP).
- Koncový uživatel nemá přístup k archivním datům, pouze k UC přes reprezentace v aplikacích pro zpřístupnění. Pokud koncový uživatel přesto potřebuje data ve vysoké kvalitě, může je manuální cestou (prostřednictvím správce systému) získat na objednávku z archivu.
- Data z projektů Manuscriptorium a WebArchiv nebudou v transformačním modulu upravována pro zpřístupňující aplikace. Zde vložení do zpřístupňující aplikace probíhá „nezávisle na projektu NDK“ mimo transformační modul, před či po vložení dat do LTP systému.
- Data z jiných zdrojů určená k archivaci a zpřístupnění v systému NDK, jsou vkládána do pracovního prostoru transformačního modulu, který zajišťuje jejich převod do balíčků SIP1 a SIP2 a jejich odeslání do LTP a aplikací zpřístupnění.
- Transformační modul sleduje tok dokumentů z externích zdrojů a kontroluje konzistenci UC a MC mezi LTP a aplikacemi zpřístupnění. V případě, že je vše v pořádku, zajistí smazání PSP balíčků externích dat z pracovního prostoru.

Popis subsystémů projektu NDK

Subsystém digitalizace

Subsystém digitalizace zajišťuje provoz digitalizačních pracovišť, která jsou umístěna ve dvou lokalitách (Praha-Hostivař a Brno). Na těchto pracovištích probíhá převod – skenování



analogových podkladů (papírových dokumentů, či mikrofilmových předloh) do digitální podoby a příprava dat pro dlouhodobé uložení a prezentaci. Vstupem do digitalizačního pracoviště jsou jednak vlastní analogové předlohy budoucích digitálních objektů, dále metadata jednotlivých objektů, která vznikají převodem z knihovního katalogu NK ČR a MZK. Výpůjčka dokumentu pro skenování se provádí v knihovním katalogu, metadata jsou stažena z katalogu po záznamu do RD. Systém RD udržuje informace o digitalizovaných dokumentech v celé ČR, z jednotlivých komponent subsystému pro zpřístupnění získává informace o lokacích uživatelských kopií (URL). Tyto informace je schopen zpětně předat knihovnímu systému.

Po naskenování analogových předloh probíhají další úpravy naskenovaných obrazů (narovnání, ořez apod.) a především kompletace informací o dokumentu – metadat. Nakonec vznikne pro každý dokument informační balíček PSP (producer submission package) zahrnující veškerá metadata náležející k dokumentu i vlastní naskenovaný obraz (=data).

Balíček PSP je z pracovního prostoru vymazán na základě kontroly přítomnosti dat v LTP a v aplikaci zpřístupnění, kterou provede kontrolní modul digitalizačního workflow (v případě externích dat, která nebudou procházet digitalizačním workflow, provede tuto kontrolu transformační modul).

Subsystém dlouhodobého uložení dokumentů (LTP subsystém)

LTP subsystém je důvěryhodný digitální repozitář umožňující efektivní správu, ochranu a uložení dat a metadat vznikajících v projektu NDK a dalších dat, která se stanou součástí NDK. Repozitář zajišťuje jak fyzickou ochranu dat (ochranu bit streamu), tak logickou ochranu dat (zachování použitelnosti, srozumitelnosti pro blízkou i vzdálenou budoucnost) podle normy OAIS (ISO 14721:2003). Subsystém LTP bude ukládat archivní kopie zdigitalizovaných dokumentů NK ČR, MZK i některých externích dodavatelů, archivní kopie born-digital dokumentů z archivace webu, soubory dalších dokumentů, především born-digital (e-depozit).

Ke každému dokumentu jednotlivě i jako k logické entitě budou ukládána metadata popisná (v omezené míře), strukturální, technická i administrativní. Jakákoliv metadata budou muset odpovídat aktuálním obecně přijímaným standardům.

Subsystém pro transformace a kontroly konzistence (Transformační modul)

LTP subsystém bude na svém vstupu vyžadovat informační balíčky SIP v určitém definovaném tvaru. Řada již existujících zdigitalizovaných dokumentů je však uložena v jiných formátech a strukturách metadat vyplývajících z dosavadní praxe digitalizace v NK ČR a MZK. Podobný problém s nevyhovujícím formátem obrazových dat a metadat může nastat i u dokumentů přicházejících z externí digitalizace. Pokud mají být stávající a externí dokumenty správně uloženy do LTP subsystému, bude třeba jejich formáty transformovat do preferované podoby. Transformační modul musí umožnit přizpůsobení různých metadatových formátů i formátů digitálních dokumentů. Z pohledu použití se bude jednat o individuální nastavení či vývoj pro konkrétní zdroje dokumentů. Protože i systémy pro zpřístupnění vyžadují na vstupu data v určitém definovaném tvaru, transformační modul bude zajišťovat též konverzi dat z různých zdrojů do formátu SIP pro zpřístupnění uživatelské kopie.

Tento subsystém, nezbytný pro komplexní chod celého systému, bude předmětem vývoje ze strany systémového integrátora celého projektu a bude nutně vytvářen ve spolupráci s odborníky z NK ČR a MZK. Bude muset být dále rozvíjen a udržován vlastními zaměstnanci obou institucí i po skončení doby trvání projektu.



Subsystém zpřístupnění informací a dokumentů

Subsystém zpřístupnění bude mít dvě vrstvy:

A. Stávající aplikace zpřístupnění – Kramerius, WebArchiv a Manuscriptorium

Základem subsystému zpřístupnění dokumentů budou tři již existující aplikace: Kramerius (instalace v NK ČR + instalace v MZK), WebArchiv a Manuscriptorium. Do jejich provozního úložiště budou uloženy zpřístupňované dokumenty jako uživatelské kopie. Uživatelská kopie dokumentu obsahuje obdobná metadata jako jeho archivní kopie, vlastní dokument je ale v úspornějším formátu postačujícím pro běžné zobrazení.

B. Aplikace pro centrální přístup

Kromě těchto aplikací bude v rámci systému zpřístupnění vybudována tzv. aplikace pro centrální přístup, která zajistí pohodlný přístup k dokumentům z různých aplikací pro běžné uživatele, a to v jediném, uživatelsky vlném rozhraní. Výhodou přístupu přes aplikaci centrálního přístupu bude, že uživateli umožní vyhledávat všechny dokumenty a informace najednou, aniž by musel vědět, ve které z výše zmíněných a jiných aplikací jsou příslušné informace uloženy a aniž by se musel seznamovat s řadou různých uživatelských rozhraní. Prostřednictvím Aplikace pro centrální přístup budou uživatelům dostupné nejen výstupy NDK, ale budou moci vyhledávat také informace o fyzických dokumentech ve fondech NK ČR a MZK a dalších knihoven nebo informace z předplacených externích elektronických zdrojů (jednotliví agregátoři nabízejí zdroje ze svých portfolií v různých uživatelských rozhraních, aplikace pro centrální přístup však již dnes umožňují přistupovat ke všem zdrojům v rámci jediného centrálního indexu).

