

RD.CZ : EVIDENCE DIGITALIZOVANÝCH DOKUMENTŮ A SLEDOVÁNÍ PROCESU ZPRACOVÁNÍ

Pavel Kocourek, Incad Praha

Přestože mnohé knihovny v České republice digitalizují své dokumenty a další se na to chystají, neprobíhá doposud koordinace těchto aktivit. Národní knihovna ve spolupráci s partnery pracuje na Registru digitalizace České republiky označovaného RD.CZ. V současné době (srpen 2008) se na projektu podílí Národní knihovna a Knihovna Akademie věd. Technickou realizaci zastrešuje společnost Incad.

Cíl projektu

Hlavním cílem projektu RD.CZ je vytvořit zázemí pro evidenci kompletního procesu digitalizace. Počínaje výběrem dokumentů určených k digitalizaci, přes předávání jednotek zpracovatelům, mikrofilmování, skenování, úpravu obrázků, tvorbu metadat, importu dat do Krameria a promítnutí výsledného umístění digitálního dokumentu do katalogu „zdrojové“ knihovny a do Souborného katalogu. Účelem systému je podat ucelený obraz o průběhu digitalizace daného dokumentu včetně financování, údajů o postupu dílčích prací, reklamaci apod. Zároveň by měl sloužit k efektivnímu předávání informací mezi jednotlivými účastníky procesu digitalizace.

Vytvořit aplikační prostředí pro:

- evidenci digitalizovaných materiálů
- řízení výroby digitalizace
- evidence zakázek a sledování jejich stavu
- sledování úkolů a postupu prací na zakázkách

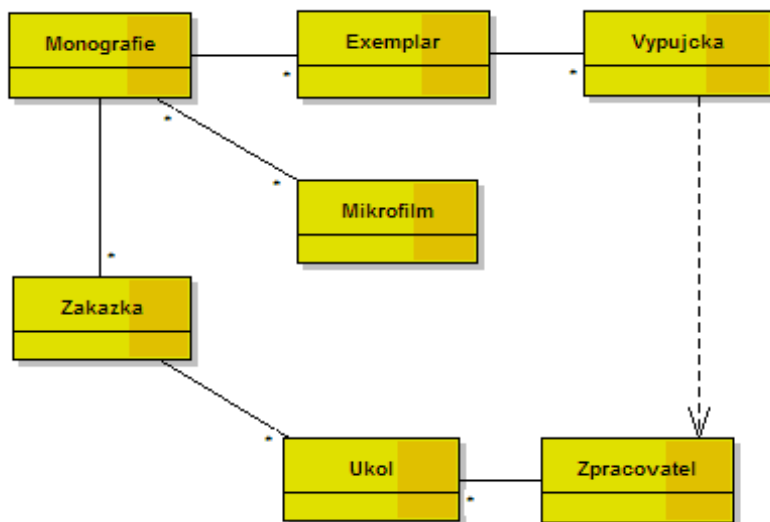
Aplikace bude poskytovat:

- přehled o zpracování a jeho průběhu
- celkový přehled o stavu a realizaci digitalizačních projektů s návazností na Souborný katalog
- zdroj dat údajů pro tvorbu metadat

Struktura evidovaných dat

Při tvorbě systému byl kladen důraz na robustnost a snadnou rozšiřitelnost systému. Snahou bylo zachovat objektový princip návrhu aplikace

a evidované údaje seskupovat na základě objektů vyskytujících se v reálném světě. Výsledkem je sada poměrně malých agend, které mají za cíl evidovat údaje týkající se jednoho typu objektu a s ním souvisejících operací. Mezi záznamy v jednotlivých agendách jsou pak vyvářeny vazby, které spojují všechny údaje v jeden celek. Stručný nástin vazeb mezi objekty je možné vidět na obrázku 1.



Obrázek 1 – Stručný nástin vazeb mezi agendami systému

Monografie + Exemplář

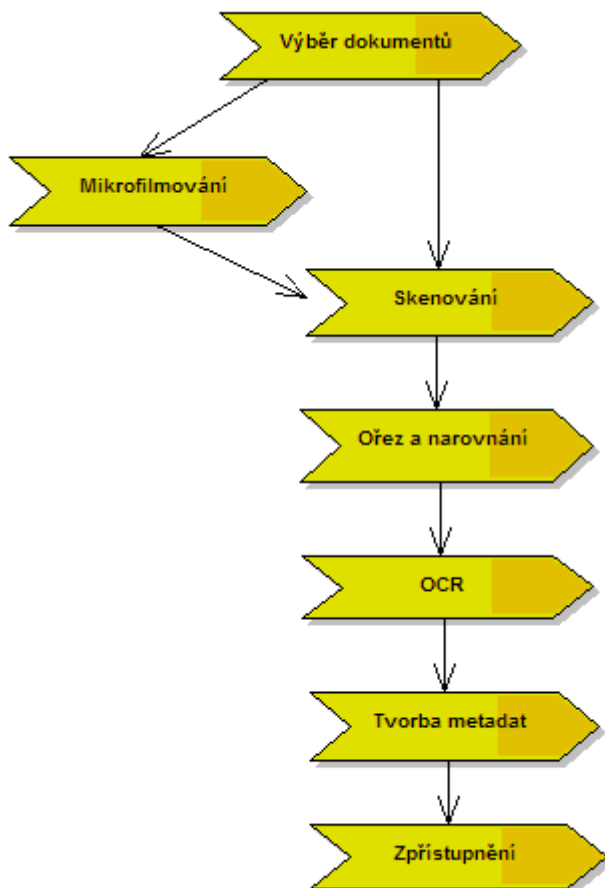
Prvním, co je v evidenci digitalizovaných dokumentů registrováno, je bibliografický záznam dokumentu a informace o jeho jednotkách. Čárový kód digitalizované jednotky je poté v mnoha funkcích považován za unikátní identifikátor dokumentu. Pro potřebu orientace uživatele v systému a některých tiskových výstupů jsou vyextrahovány základní údaje jako název, autoři, rok vydání, čárový kód, signatura apod. Zároveň je v systému uchován kompletní záznam ve formátu MARC-XML rozšířeném o elementy popisující digitalizovanou jednotku.

Výpůjčka

V rámci agendy Výpůjčka je evidováno předání dokumentů zpracovatelům a jejich řádné vrácení.

Mikrofilm

Pokud je součástí digitalizace také tvorba mikrofilmu, pak jsou editovány i kompletní informace o mikrofilmu. Záznam obsahuje jedinečný identifikátor mikrofilmu, tj. číslo archivního negativu, kameru, operátora atd. Součástí záznamu je i umístění dokumentů na mikrofilmu včetně unikátního identifikátoru Mf E, jejich pořadí na svitku, počet polí hrubých a čistých, atd. K evidenci mikrofilmu je možno připojit záznam o revizích technických parametrů. Součástí agendy Mikrofilm je i několik tiskových výstupů, např. průvodka mikrofilmu, návěští mikrofilmu apod.



Obrázek 2– Příklad možného workflow

Zakázka + Úkol

Záznam o zakázce je vždy připojen k bibliografickému záznamu dokumentu, který byl předán do procesu digitalizace. Unikátním identifikátorem zakázky je IDD (Identifikátor digitalizovaného dokumentu) nazývaný také zakázkové číslo. IDD je součástí názvů souborů výsledného digitálního dokumentu. Celá zakázka se skládá z dílčích úkolů, jejichž zpracovatelé se mohou navzájem lišit. Struktura úkolů i jejich pořadí se mohou navzájem lišit.

Zpracovatel

Evidence zpracovatelů slouží především jako zdroj pro tvorbu úkolů.

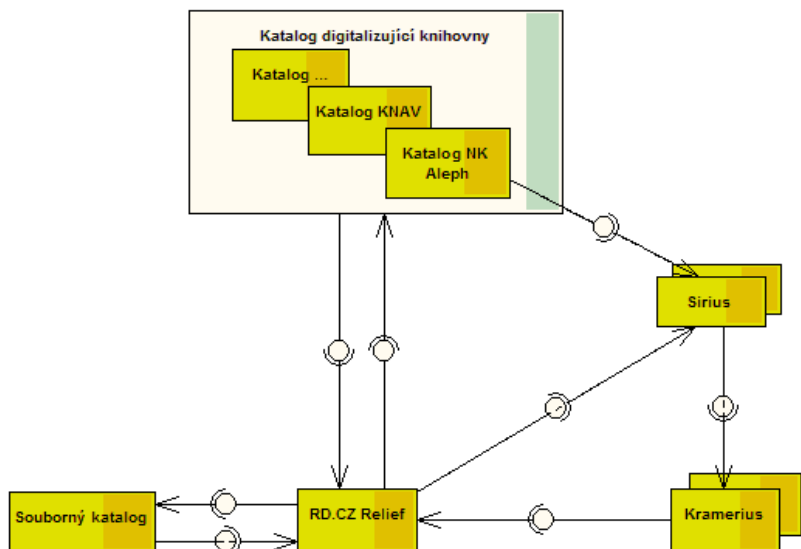
Digitalizační Workflow

Digitalizační workflow se může lišit v závislosti na zvyklostech instituce, která digitalizuje své dokumenty a v rámci jedné instituce i podle typu dokumentu. Nejčastějším rozdílem je použití mikrofilmu jako podkladu pro skenování nebo přímá digitalizace dokumentu. Rozdílná může být i požadovaná podrobnost členění činností, které by měly být sledovány. Z tohoto důvodu byla struktura úkolů vytvořena velice volně, kdy je možné průběžně definovat „typy práce“ a libovolně je kombinovat v rámci zakázky. Je možné definovat i návaznost jednotlivých prací. Na obrázku 2 je příklad jednoduchého stanovení typů prací.

Integrace

Cílem projektu je propojit existující nástroje, které jsou v procesu digitalizace už nyní využívány. Nesnaží se nahradit funkce stávajících nástrojů, ale spíše je vhodně integrovat, aby mohly být jako celek maximálně využity.

1. Import bibliografických dat z katalogu digitalizující knihovny.
2. Předání všech potřebných údajů externím zpracovatelům, v závislosti na druhu práce, která má být provedena. Tj. bibliografické údaje rozšířené o údaje o mikrofilmu, zakázce apod.
3. Pravidelné přebírání údajů z Krameria pomocí OAI-PMH a jejich doplnění do RD.CZ.
4. Předávání URL digitalizovaného dokumentu zpět do katalogu digitalizující knihovny na základě unikátního identifikátoru z jejího katalogu.
5. Předávání údajů o zahájení digitalizace a posléze i URL digitálního dokumentu do Souborného katalogu.



Obrázek 3 – Návaznost RD.CZ na ostatní nástroje používané v procesu digitalizace

Popis výchozího stavu

Evidence prací souvisejících s digitalizací dokumentů byla až do nedávna a v některých případech stále ještě je evidována pomocí řady excelových souborů, které si vede každé oddělení zvlášť, a to pouze v rozsahu údajů, které jsou pro danou problematiku nutné. Zkompletovat proto celkový obraz o digitalizaci těchto dílčích dokumentů je velice obtížné, ne-li nemožné. V souborech chybí unikátní identifikátory, pomocí kterých by bylo možné dokumenty identifikovat napříč všemi soubory, a tak dohledat údaje, které jsou o nich vedeny.

Použité technologie

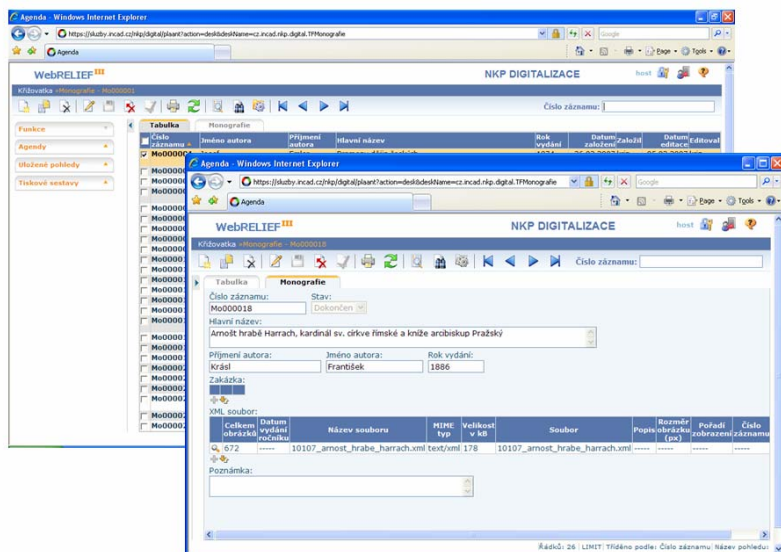
Systém je postaven na aplikačním frameworku RIII (J2EE) a data jsou ukládána do relační databáze Oracle.

Pro zpřístupnění aktuální informace o stavu digitalizace koncovým uživatelům je použit vyhledávací nástroj Fast RW.

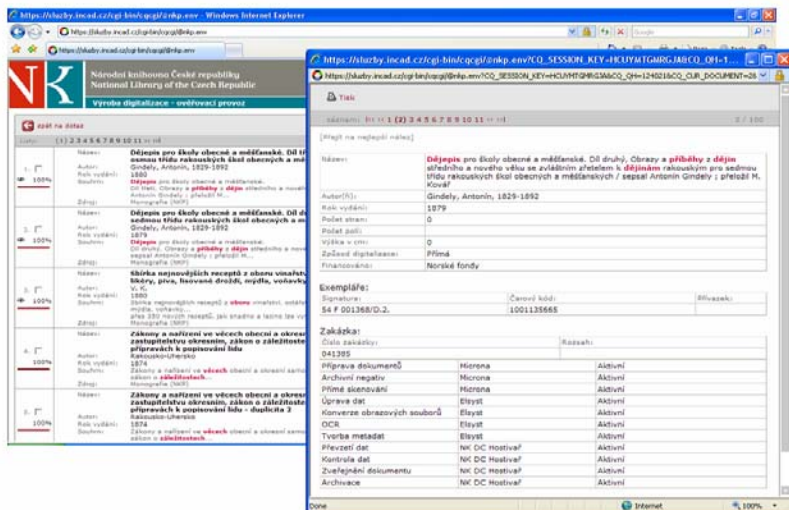
Všechny uživatelské přístupy jsou realizovány prostřednictvím webové aplikace, což neklade žádné nároky na klientská zařízení, kromě existence webového prohlížeče.

V současnosti je systém provozován v ASP modelu ze serverového prostředí dodavatele. V budoucnu se předpokládá samostatné serverové pro-

středí na páteřní síti v jedné z účastnických institucí, pravděpodobně Národní knihovně.



Obrázek 4 – Editaci prostředí RIII RD.CZ



Obrázek 5 – Ukázka vyhledávání

Výhled

Spolu s plánovaným přechodem Národní knihovny na hromadnou digitalizaci pomocí robotických skenerů bude systém rozšířen o podporu automatického sklizení dat vzniklých na těchto zařízeních a jejich následné zpracování a evidenci v aplikaci RD.CZ.

Dle požadavků dalších účastnických institucí bude rozšířen záběr evidovaných dat a procesů. Zároveň by měly být zjednodušeny procesy, které vyžadují uživatelský zásah.

Po zaběhnutí vstupu dat se plánuje rozšíření systému o řadu kontrolních mechanismů a statistik, které umožní organizační náhled na proces digitalizace.

Vyřešeny budou muset být také pravidla sdílení digitalizovaných dokumentů.