Národní digitální knihovna

Jan Hutař, Marek Melichar, Bohdana Stoklasová / Národní knihovna ČR
e-mail: jan.hutar@nkp.cz; marek.melichar@nkp.cz; bohdana.stoklasova@nkp.cz

Úvod: Historie digitalizace, dlouhodobé ochrany a zpřístupnění digitálních dokumentů v knihovnách ČR

Historie digitalizace analogových dokumentů, „sklízení" a archivace českého webu i dlouhodobé ochrany a zpřístupnění digitálních objektů je poměrně dlouhá a koresponduje s historií digitalizace analogových dokumentů, „sklízení" a archivace webu a dlouhodobé ochrany a zpřístupnění digitálních objektů v nejvyspělejších zemích různých kontinentů. Několik národních grantových projektů umožnilo odstartovat projekty digitalizace v knihovnách ČR již počátkem 90. let minulého století. S archivací webu se začalo v roce 2000, od roku 2004 se tým českých expertů intenzivně věnuje problematice trvalé ochrany a zpřístupnění digitálních objektů.

Od samého počátku byly respektovány mezinárodní standardy a díky tomu je možné všechny výstupy lehce integrovat do různých národních (JIB) i nadnárodních portálů (TEL, EUROPEANA apod.). Ačkoliv je ČR malá země, vydobyla si celosvětové uznání svými dlouhodobými aktivitami v oblasti digitalizace a digitální ochrany: v roce 2005 obdržela Národní knihovna ČR (NK ČR) cenu UNESCO/JIKJI Memory of the World za svůj přínos k ochraně a zpřístupňování kulturního dědictví.

I přesto jsou digitalizace, archivace webu i problematika digitální ochrany v ČR významně pozadu za ostatními státy v důsledku nedostatku finančních prostředků a následně pomalého postupu digitalizace. Ze stejného důvodu zatím NK ČR dosud nevybudovala tzv. důvěryhodný digitální repozitář, který by byl schopen projít mezinárodní certifikací.

Digitální objekty určené k dlouhodobé ochraně a zpřístupnění zahrnují digitalizované analogové dokumenty a tzv. born digital dokumenty. Vznikají v rámci tří velkých národních projektů.

Manuscriptorium (http://www.manuscriptorium.com) je systém pro vytváření sbírek a zpřístupnění informací o historických a vzácných dokumentech na internetu, včetně virtuální digitální knihovny digitalizovaných dokumentů.

Kramerius (http://kramerius.nkp.cz) se zaměřuje na ochranu a zpřístupnění periodik, knih a ostatních dokumentů publikovaných od roku 1801. Velká část těchto dokumentů je silně ohrožena z důvodu tisku na kyselém papíře a/nebo častého používání.

WebArchiv (http://www.webarchiv.cz) je digitální archiv českých webových zdrojů, které jsou shromažďovány s cílem jejich dlouhodobé ochrany a zpřístupnění.

Národní digitální knihovna jako strategický projektový záměr pro čerpání finančních prostředků ze strukturálních fondů EU v rámci Smart Administration

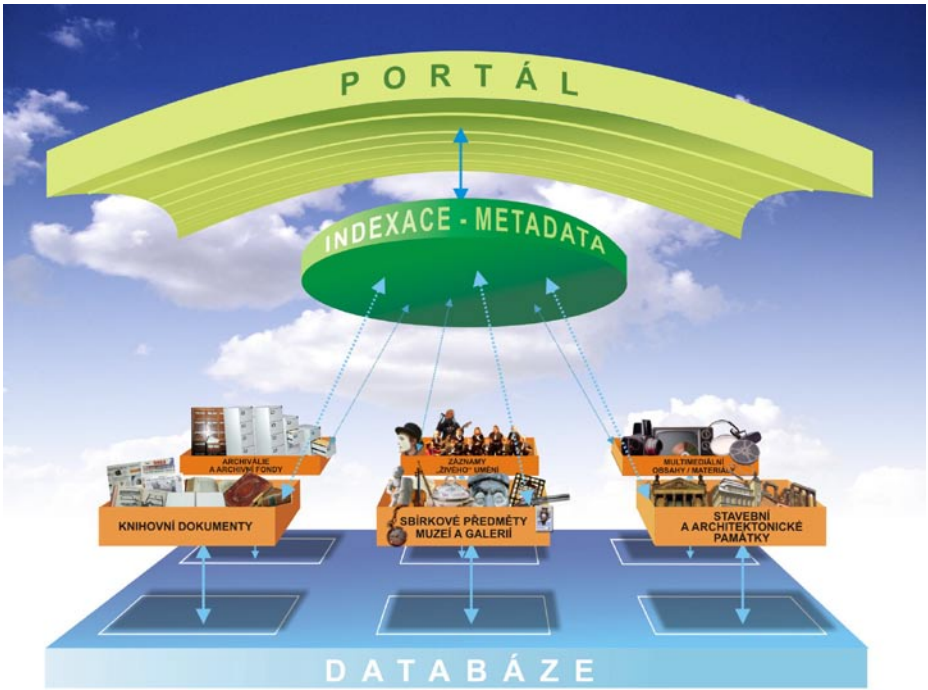
Z úvodní kapitoly vyplývá, že knihovny v ČR jsou na vybudování Národní digitální knihovny (NDK) v plném rozsahu velmi dobře teoreticky připravené. Existují zde expertní týmy pokrývající jednotlivé stavební komponenty, rozsáhlá a dlouhodobě prověřená kooperace českých knihoven, zkušenosti získané v rámci řešení domácích i zahraničních projektů, rozsáhlé mezinárodní kontakty, přednášková i publikační činnost doma i v zahraničí.

Brzdou rozvoje celé oblasti je prozatím chronický nedostatek finančních prostředků plynoucí z nedostatečné podpory nejvyšších státních orgánů. Ministerstvo kultury ČR ve spolupráci s NK ČR připravilo v roce 2005 *Koncepci trvalého uchování knihovních sbírek tradičních a elektronických dokumentů v knihovnách ČR do roku 2010*. Koncepce měla být projednána vládou ČR a její realizace měla být podpořena finančním objemem 210 mil. Kč, k čemuž bohužel nikdy nedošlo.
Celková situace by se mohla ale velmi brzy změnit. Ministerstvo kultury a česká vláda přijaly NDK za strategickou prioritu a jako kandidáta pro financování v rámci Integrovaného Operačního Programu – IOP (Smart Administration). Projekt *Vytvoření Národní digitální knihovny* je uveden v příloze vládního usnesení č. 536 ze 14. května 2008, o strategických projektových záměrech pro čerpání finančních prostředků ze strukturálních fondů Evropské unie v rámci Smart Administration – příloha ke strategii Efektivní veřejná správa a přátelské veřejné služby.

Klíčové postavení Národní digitální knihovny v rámci koncepce dlouhodobého uchování digitálních dokumentů (nejen) v knihovnách ČR

Ministerstvo kultury ČR připravuje Národní strategii digitalizace kulturního dědictví. Strategii bude vytvářet pracovní skupina, jejímiž členy jsou pracovníci Ministerstva kultury ČR a příspěvkových organizací, jichž se oblast digitalizace nejvíce týká. Přístup uživatelů k českému národnímu kulturnímu dědictví bude řešen přes jeden národní portál zahrnující knihovní dokumenty, archiválie, muzejní sbírky, architektonické památky, média a živé umění. Počet a uspořádání jednotlivých segmentů se může v průběhu přípravy strategie změnit. Základní rámec, korespondující s řešením obdobné situace v zahraničí, však zůstane podobný.

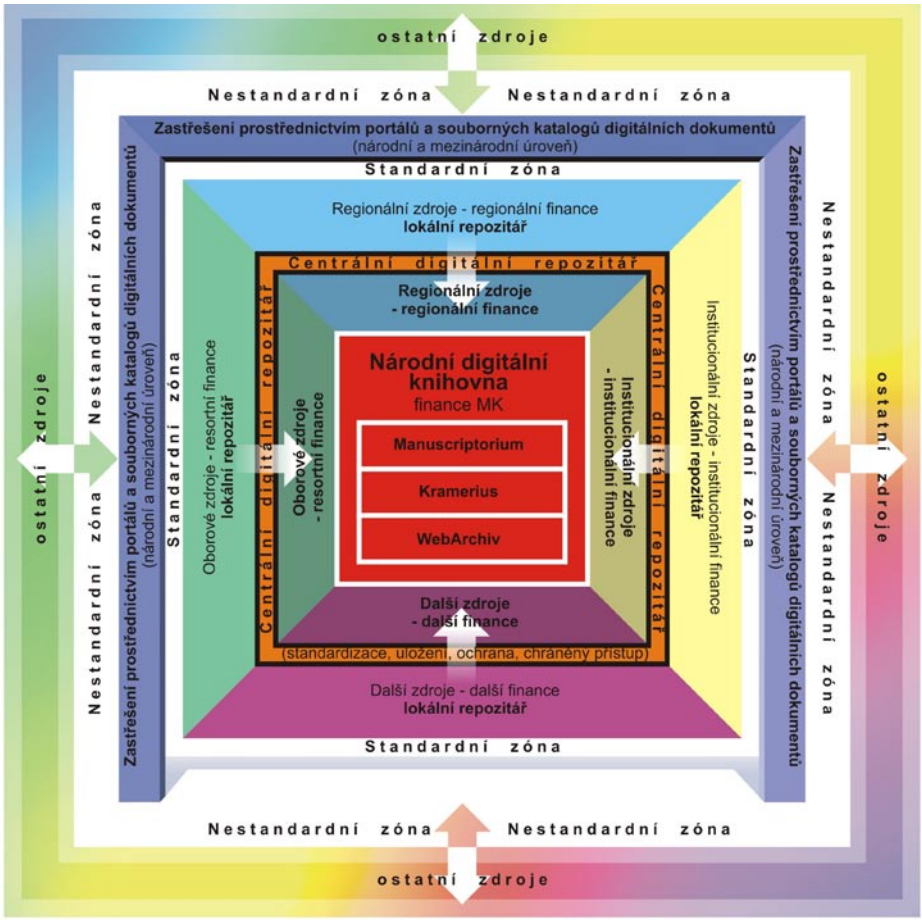
Obr. č. 1: Digitalizace a zpřístupnění národního kulturního dědictví



NDK bude zahrnovat významnou část národního kulturního dědictví, neboť, jak je patrné z obrázku výše, knihovní dokumenty jsou jedním z pilířů celého kulturního kontextu. NDK bude fungovat v širším kontextu České digitální knihovny.

Obrázek níže ilustruje celkový koncept České digitální knihovny. Začněme uprostřed diagramu. Střed (srdce) celého systému NDK obsahuje vybrané digitální objekty, které jsou považovány za jádro národního kulturního dědictví. Tyto digitální objekty určené k dlouhodobé ochraně a zpřístupnění jsou digitalizované analogové dokumenty nebo tzv. born digital dokumenty. Vznikají v rámci tří výše zmíněných národních projektů: Manuscriptorium, Kramerius a WebArchiv.

Obr. č. 2: Česká digitální knihovna

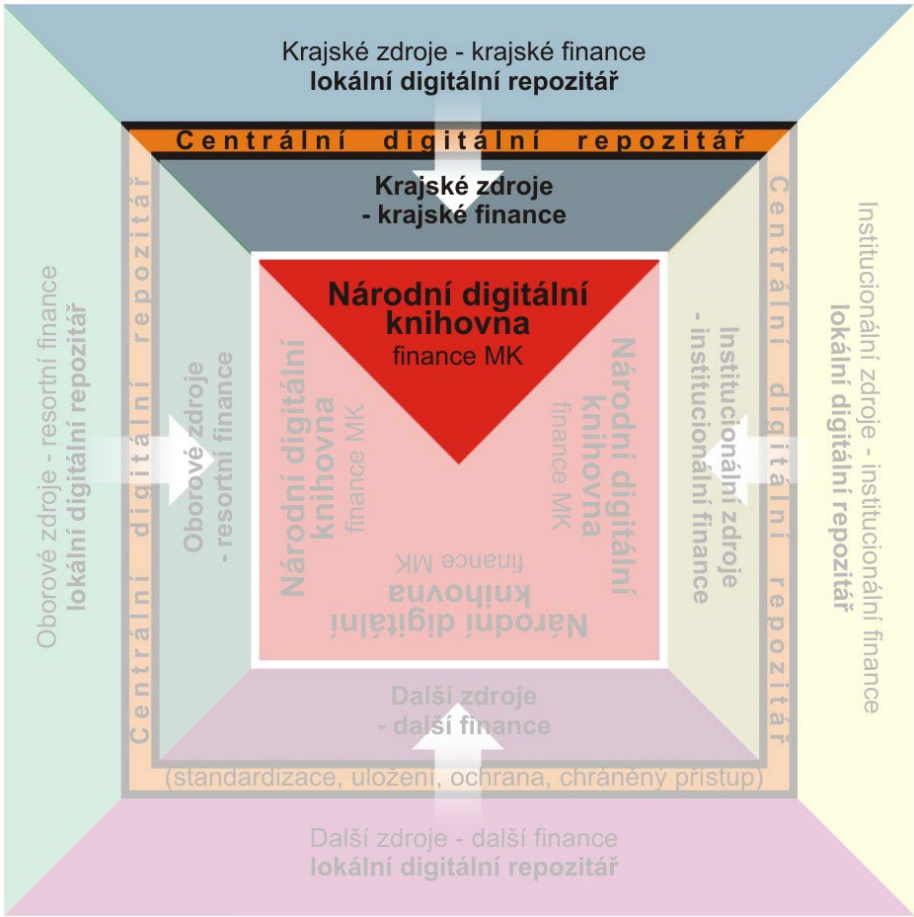


Dokumenty, které vlastní nebo vytvoří jakákoliv česká knihovna, muzeum, archiv nebo jiná podobná instituce, mohou být vybrány za součást NDK. Digitalizace, vytváření metadat a ochrana těchto vybraných dokumentů budou (a v rámci projektů již jsou) financovány Ministerstvem kultury ČR.

Instituce s digitálními objekty, které nebyly vybrány do NDK, budou i tak moci uložit svá data v Centrálním digitálním repozitáři, pokud o to budou stát. Bude ovšem požadována finanční spoluúčast ostatních ministerstev (v závislosti na různých oborech), regionů i samotných institucí. Digitální data jsou vytvářena samozřejmě i v ostatních institucích, které ovšem nemusí mít zájem o uložení svých dat v Centrálním repozitáři. Data uložená v lokálních repozitářích provozovaných takovými institucemi financovanými různými ministerstvy, samotnými institucemi, místními úřady nebo firmami, mohou být integrována do národního, mezinárodního portálu nebo jiných integračních nástrojů, pokud budou podporovat obecně uznávané a domluvené standardy.

Situaci názorně ilustruje modifikace schématu České digitální knihovny pro regionální (krajskou) úroveň, na které vznikají navazující projekty IOP:

Obr. č. 3: Národní digitální knihovna a krajské zdroje



Hlavní funkční požadavky na systém dlouhodobého uchování digitálních dokumentů v repozitáři Národní digitální knihovny

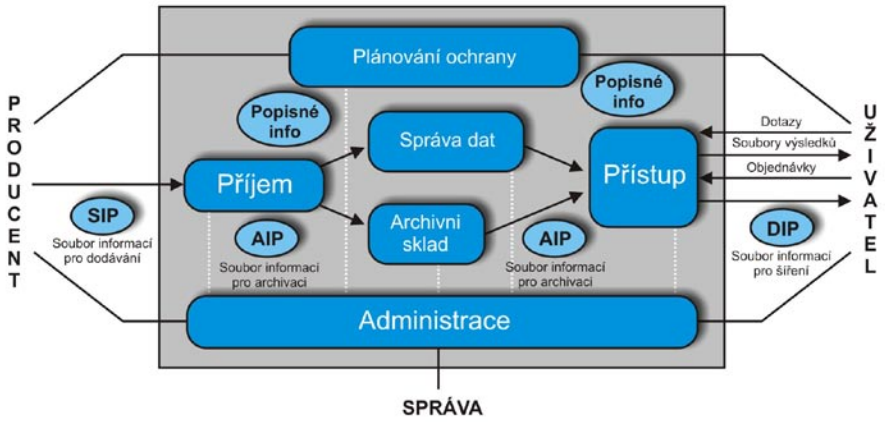
Dlouhodobá ochrana digitálních dokumentů je nový a dynamicky se rozvíjející obor, který hledá cesty, jak čelit nebo předcházet dopadům zastarávání či selhávání hardware, datových nosičů a zastarávání software nebo formátů souborů. Cílem digitální ochrany je nejen zachovat „bity“ (geograficky distribuovaným zálohováním dat, přesunem dat na nové datové nosiče či do nových technologií ap.), ale rovněž zajistit technickou použitelnost a sémantickou srozumitelnost archivovaných digitálních dokumentů i po velmi dlouhé době.

Digitální repozitář musí zajistit, aby uživatel mezi stovkami milionů nebo dokonce miliardami digitálních dokumentů dokázal snadno a rychle identifikovat, vyhledat a získat ty, které potřebuje. Repozitář musí dále zaručit, že tyto dokumenty jsou autentickými verzemi nějakého konkrétně identifikovatelného digitálního dokumentu první instance, jehož původ je jasně dokumentovaný. Repozitářem je myšlena komplexní organizace obsahující technologické řešení dlouhodobé archivace digitálních dat a zahrnující vše, co chod repozitáře může i minimálně ovlivnit. Celá organizace musí být institucí vhodně

financovanou a řízenou. V oblasti dlouhodobé ochrany a zpřístupnění digitálních dokumentů musí mít jasnou strategii a disponovat nejen adekvátním technologickým vybavením, ale i dostatkem kvalifikovaných zaměstnanců k jejímu naplnění. V souvislosti s dlouhodobou ochranou a zpřístupněním hraje důležitou roli i dlouhodobá udržitelnost – dlouhodobá ochrana digitálních dokumentů je permanentní proces. Tuto skutečnost ve svých příspěvcích opakovaně zdůrazňuje zástupce UNESCO Abdelaziz Abid: „*Pro ochranu analogových dokumentů většinou stačí uložení v optimálních podmínkách, dostatečná kontrola fyzického stavu a minimální využívání. U digitálních dokumentů je situace podstatně složitější. Jejich ochranu lze přirovnat k udržování ohně – je nutné se mu věnovat neustále, udržovat ho a kontrolovat. Jinak zhasne a nenávratně zmizí. Při správné péči ale může být věčný*“.

Součástí projektu NDK je kromě zavedení masové digitalizace také vybudování důvěryhodného repozitáře pro dlouhodobou ochranu digitálních dokumentů. Systém repozitáře musí vycházet z požadavků na důvěryhodný digitální repozitář (tj. odpovídat konceptuálnímu modelu Open Archival Information System – OAIS).

Obr. č. 4: Model OAIS



Po dobudování a uvedení do plného provozu bude muset splňovat mezinárodní certifikační kritéria.

V současné době jsou pro základní řešení dlouhodobé archivace na trhu k dispozici především systémy DIAS od IBM, systém Rosetta od ExLibris a systém SDB od firmy Tessella. Tým pro přípravu projektu NDK všechny uvedené systémy monitoruje nejen na úrovni teoretických analýz a firemních prezentací, ale i formou několikadenních stáží přímo v místech jejich praktické aplikace. Kromě systémů, které jsou přímo určeny na ochranu digitálních dat ve smyslu dlouhodobé archivace, existují i další řešení implementující volně dostupný archivační software jako například Fedora Commons a další.

Ať již bude vybrán jakýkoli systém dlouhodobé ochrany digitálních dokumentů, bude jeho implementace vyžadovat napojení na existující subsystémy NDK (katalogy, digitální knihovny, portály), napojení na systémy nově budovaného pracoviště masové digitalizace a komplexní integraci stávajících sbírek digitálních dokumentů (například 6 milionů stránek z projektu Kramerius včetně jejich metadat a tzv. master souborů). Výběr systému by měl být proveden na základě posouzení schopnosti dodavatele naplnit funkční požadavky definované NDK. Dodavatel musí prokázat, že systém je schopen funkčně a technicky integrovat mezi další systémy knihovny a že bude poskytovat implementaci také dlouhodobou podporu. Začíná se ovšem ukazovat ze zkušeností zahraničních knihoven, že velmi podstatným hlediskem je i schopnost dodavatele prokázat, že sys-

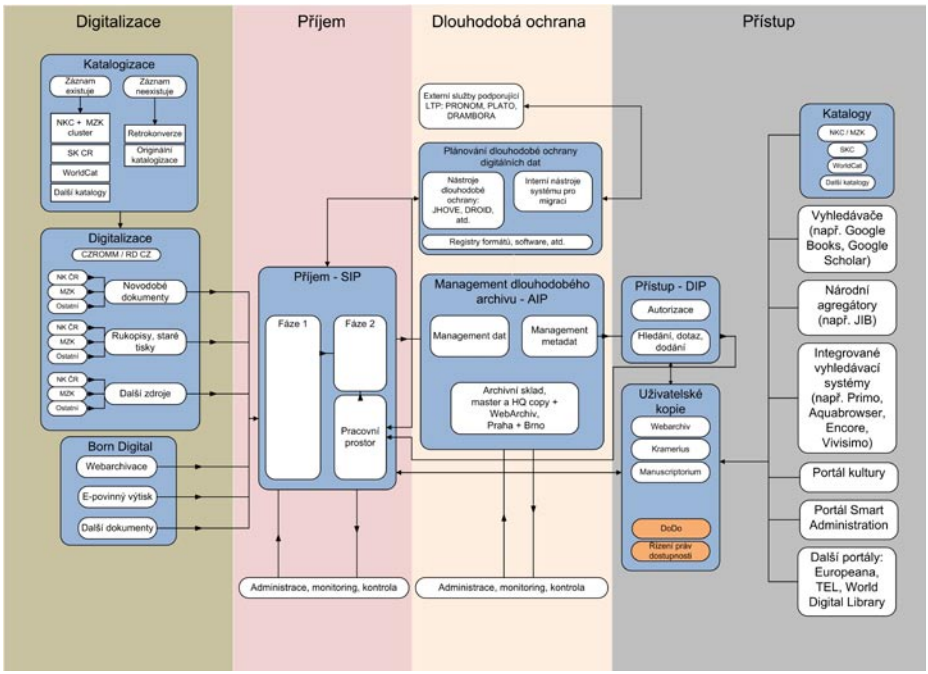
tém má naplánovaný rozvoj do budoucna a bude reflektovat požadavky již existujících uživatelů i obecné uživatelské komunity, a to v souladu s vývojem v oboru dlouhodobé ochrany digitálních dat. Jedná se o velmi komplexní systém, náklady na jeho vývoj a rozvoj jsou vysoké, přičemž skupina potencionálních uživatelů je relativně malá. Systém samotný by měl být v provozu alespoň v několika institucích, musí mít za sebou silný řešitelský tým i zkušený tým uživatelů a musí mít zaručenou perspektivu dalšího vývoje.

V tomto příspěvku není prostor pro podrobný popis funkčních požadavků na systém dlouhodobé ochrany digitálních dokumentů pro NDK, obecně však systém musí vyhovět především následujícím kritériím:

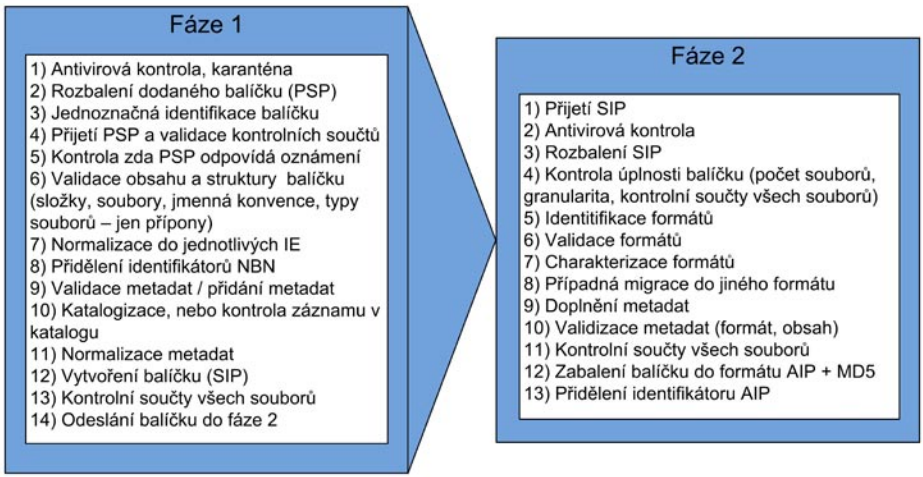
- systém musí být otevřený a vysoce interoperabilní, tj. musí umožňovat odpojení a připojení různých částí systému bez omezení funkcionality jiných, musí umožňovat jednoduchou a otevřenou integraci externích nástrojů (přes API a SDK) a jejich využití v pracovních postupech systému pro případnou migraci, emulaci, validaci formátů, ale i pro spolupráci s katalogem, deposit, zpřístupňovací aplikace, vyhledávací nástroje apod.;
- systém musí podporovat otevřené komunikační protokoly;
- systém musí být kvalitně a kompletně zdokumentován;
- systém musí být rozšířitelný, tj. musí být možné pružně měnit parametry systému, jak z hlediska celkové ukládací datové kapacity, tak z hlediska počtu a typů uložených objektů, jejich velikosti a z hlediska datové prostupnosti kritických míst systému;
- systém musí být v mnoha ohledech nastavitelný (NDK například bude i přes relativní homogenitu materiálů přicházejícího z masové digitalizace potřebovat různé definice informačních balíčků SIP, AIP a DIP, systém musí být schopen je zpracovávat paralelně, musí umožňovat snadnou integraci nových formátů a nových projektů);
- systém musí mít pravidly řízené pracovní postupy (rule based workflow);
- řešení musí být zcela nezávislé na použitých typech hardware (architektura řešení by neměla být svázána s jedním určitým typem archivního skladu (archival storage), jedním výrobcem …);
- řešení musí být pokud možno maximálně automatizované (ve smyslu zpracování hromadných dodávek dokumentů, automatická musí být i distribuce, konverze), ovšem při maximálním zachování možnosti manuálního zpracování nastavení;
- systém musí umožňovat provádět ochranné akce hromadně na administrátorem definovaných skupinách objektů;
- řešení musí podporovat řízení práv dostupnosti, podporovat systémy typu Onelog, Shibboleth, udržovat databázi uživatelů i dodavatelů dat, zaznamenávat jejich aktivity;
- systém musí mít maximálně propracovaný modul pro monitorování a sledování akcí probíhajících na vstupu do repozitáře, dále při archivaci, distribuci, správě a administraci;
- řešení musí podporovat zapojení nástrojů třetích stran pro plánování dlouhodobé ochrany, realizaci hromadných ochranných aktivit a hodnocení jejich úspěšnosti;
- systém musí umožňovat vytváření logických souborů digitálních objektů uložených v repozitáři a následně i manipulaci s nimi.

Základní složky systému dlouhodobé archivace budou v rámci NDK obsahovat následující funkční moduly.

Obr. č. 5: Modifikace modelu OAIS pro workflow NDK



Obr. č. 6: Příjem – specifikace fází 1 a 2



Přijetí materiálu – první fáze:

Příprava dokumentů pro vstup do repozitáře bude začínat již na pracovišti masové digitalizace nebo při harvestingu ve WebArchivu. Odtud budou data přicházet již se základními metadaty (popisnými, strukturálními, administrativními, po kontrole kvality obrazu a zpracování OCR). Pro pracoviště masové digitalizace a WebArchivu musí existovat nezávislé aplikace pro zpracování příjmu, které budou v podstatě automatizované a budou se lišit od ručního vkládání do systému. V budoucnu může vzniknout potřeba archivovat

i další typy dat (například audiovizuální data), což může vyžadovat zapojení dalších aplikací pro vstup dat nebo minimálně flexibilitu deposit modulu systému, který by si měl umět poradit s jakýmkoliv typem souboru. Systém bude v této fázi provádět: kontroly úplnosti dat, antivirové kontroly, validaci struktury balíčků, vytvoření intelektuálních entit, přidělení URN:NBN:CZ nebo i dalších vnitřních identifikátorů, validaci metadat, kontrolu záznamu v registru digitalizace, doplnění metadat, vytvoření balíčku pro další fázi, odeslání do další fáze zpracování. V každém okamžiku v případě problému bude dokument vrácen buď producentovi, nebo zaměstnanci, který rozhodne o jeho dalším osudu.

Přijetí materiálu – druhá fáze:

Druhá fáze vstupu do repozitáře by měla být společná pro všechny typy vkládaných materiálů a pro všechny dodavatele. Zde půjde především o provedení automatické identifikace a validace formátů souborů, tzv. enrichment (obohacení metadat automatickou cestou), případnou migraci do preferovaných formátů (normalizace), opětnou antivirovou kontrolu. Toto bude probíhat zapojením služeb jako je JHOVE, PRONOM/DROID případně s využitím NZME (New Zealand Metadata Extractor). Průběh pohybu vstupního balíčku by měl být vidět v online monitorovacím modulu systému, kde bude informace dostupná správci repozitáře i dodavateli dat a tato data se budou dále archivovat.

Správa archivního modulu

Jedná se o jádro systému dlouhodobé ochrany. Základem je systém archivace (archival storage) – vlastní technologie uchovávání digitálních dokumentů na fyzických úložištích. Současné metody archivace mají řadu funkcionalit, které dlouhodobou archivaci podporují. Ačkoli využívají ultra levná úložná média, díky pokročilému clustrování, inteligentnímu nakládání s daty, monitoringu a dalším funkcím jsou z hlediska dlouhodobé ochrany velmi vhodné. Nad archivací musí mít systém komplexní data a metadata management, administrativní rozhraní. To vše s odpovídajícím GUI rozhraním.

Plánování ochrany

Systém by měl monitorovat základní vlastnosti vkládaného materiálu a inteligentně pomáhat správcům repozitáře s plánováním dlouhodobé ochrany (musí uchovat informace o vložených formátech a platformách, na kterých fungují, o použitých metodách komprese a dalších souvisejících technologiích, které mohou mít potenciálně dopad na použitelnost archivovaného materiálu). Systém musí umožňovat, aby oprávněný správce repozitáře mohl vyexportovat z archivu různě definované množiny objektů pro účely například migrace nebo jiné ochranné akce mimo repozitář (ve fázi testování migračních nástrojů atp.), případně obsahovat základní nástroje pro migraci dat. Jinou možností je, že systém umožní provádět hromadné ochranné akce použitím externích nástrojů. Další vývoj v oblasti dlouhodobé ochrany pravděpodobně přinese nové praktické nástroje, jež provádění ochranných akcí podpoří nebo usnadní. Proto musí být systém dostatečně otevřený, aby tyto nástroje bylo možné v budoucnu využívat.

Přístup

Systém repozitáře bude muset umožňovat vyhledávání a dodání archivovaných dokumentů a jejich metadat v různě strukturovaných balíčcích DIP při dodržení přístupových práv. Systém musí být propojený s přístupovou vrstvou, která bude odpovídat požadavkům definovaným na úrovni IOP – Smart Administration pro všechny dodavatele dat financované z tohoto zdroje, moderním systémům knihovních portálů a požadavkům uživatelů, prostřednictvím používaných komunikačních protokolů. Systém musí také nabídnout provázání s již existujícími databázemi knihovny (katalog NK, případně Souborný katalog ČR, digitální knihovny) tak, aby v rámci těchto systémů bylo možné přistupovat k archivovanému obsahu velmi jednoduchým a bezpečným způsobem.

Administrace, Monitoring

Systém musí mít propracovaný modul pro monitoring. Měl by podporovat sledování pohybu dokumentů v jednotlivých fázích životního cyklu balíčků SIP>AIP>DIP, měl by monitorovat a zaznamenávat všechny prováděné akce, měl by monitorovat použité formáty a software. Měl by také podporovat monitoring distribuce dokumentů z repozitáře. Bylo by vhodné, aby monitoring bylo možné použít v obchodním modelu repozitáře, tj. sledovat náklady na skladování dokumentů, vykazovat náklady jednotlivým skupinám uživatelů a dodavatelů dat.

Problém formátů archivovaných digitálních dokumentů

Digitální dokumenty se skládají z jedniček a nul, avšak v praxi nepracujeme s pouhými soubory těchto dvou číslic. Formáty jsou pravidla, podle kterých se soubory jedniček a nul strukturují (tzv. endianita), přičemž sama tato pravidla jsou často obsažena v datech samotných (někdy jako přípona souboru, ale to samo o sobě málokdy stačí). Formáty stejně jako hardware nebo datové nosiče rychle zastarávají.

Pro dlouhodobou ochranu digitálních dokumentů je klíčové vybrat vhodné formáty. To se ovšem pojí s otázkou, zda omezovat typy formátů, které budou do systému přijímány, či přijímat vše. Z podstaty věci by měl systém NDK přijímat všechny typy dokumentů, bude ovšem mít seznam preferovaných typů. Obecně se doporučuje, aby formát, který má sloužit pro dlouhodobé uchovávání, byl otevřený, tj. musí k němu existovat volně dostupná a otevřená dokumentace. Otevřený formát může být jak proprietární (např. PDF/A patřící firmě Adobe), tak neproprietární (např. OpenOfficeXML). Otevřenost formátu do určité míry zaručuje, že přestane-li se nějaký formát používat, zůstane alespoň možnost zrekonstruovat způsob, jak formát zpracovat na základě analýzy dokumentace. Volba formátu je také důležitá v případě, že se rozhodneme migrovat data z jednoho formátu do druhého, protože první již zastarává nebo se objevil nějaký nový, vhodnější. Komunity zabývající se ochranou digitálního kulturního dědictví často vydávají doporučení nebo studie, zkoumající efektivnost převodů a výběr vhodných formátů, proto je nejlepší (a v našem kontextu de facto jedinou možnou) strategií zapojit se do co nejvíce společných projektů a implementovat ty formátové strategie, které jsou co možná nejrozšířenější. Z výše uvedeného vyplývá, že tato problematika je stále ve vývoji a jedině pravidelným monitorováním situace a soustavným přijímáním nových opatření lze udržovat stav na optimální úrovni.

Orientační přehled o vhodnosti jednotlivých formátů z hlediska dlouhodobé ochrany nabízí například doporučení Florida Digital Archive. [1]
Pro ilustraci uvádíme hodnocení formátů rastrových obrazů.

Vhodné pro dlouhodobou archivaci	Částečně vhodné pro dlouhodobou archivaci	Nevhodné pro dlouhodobou archivaci
TIFF (nekomprimovaný) JPEG2000 (bezztrátový) (*.jp2) PNG (*.png)	BMP (*.bmp) JPEG/JFIF (*.jpg) JPEG2000 (ztrátový) (*.jp2) TIFF (komprimovaný) GIF (*.gif) Digital Negative DNG (*.dng)	MrSID (*.sid) TIFF (ve formátu Planar) FlashPix (*.fpx) PhotoShop (*.psd) RAW JPEG 2000 Part 2 (*.jpf, *.jpx) všechny další formáty

Systém dlouhodobé ochrany NDK by měl být schopen přijmout jakýkoliv typ vstupních formátů. Jak již bylo zmíněno, hlavní objem dat budou tvořit dokumenty z digitalizace (v současnosti probíhají diskuse o volbě formátů); u těchto bude muset repozitář zajistit plnou dlouhodobou ochranu) nebo dokumenty z archivace webu. V případě dat z WebArchivu je rozsah používaných formátů na tvorbu webových stránek velmi široký, repozitář tedy bude muset data přijímat ve formátu ARC či nověji WARC, ve kterém jsou produkována při procesu archivace nástrojem Heritrix.

Hlavními strategiemi, jak předcházet zastarávání formátů, budou:
- kontrola validity formátů při vstupu dat do repozitáře;
- extrakce technických metadat;
- specifikace klíčových vlastností dokumentů při plánování dlouhodobé ochrany
- udržování databáze vložených formátů a jejich technických a administrativních popisů;
- sledování vývoje v oblasti dlouhodobé ochrany;
- sledování nových formátů;
- sledování očekávání uživatelů a testování a případná migrace do nových formátů.

V případě, že systém za součinnosti technického analytika dojde k závěru, že určitý formát je zastaralý, musí dojít k naplánování ochranné akce a následně k aktivitě, která zajistí další použitelnost digitálního objektu. Nejužívanější metodou je migrace (z formátu do nového formátu), další je emulace, tj. simulace prostředí hardware a software, ve kterém byl původní digitální objekt použitelný. V současnosti se s emulací počítá pro dokumenty vzešlé z činnosti WebArchivu.

Důvěryhodnost a spolehlivost jako cíl dlouhodobé ochrany digitálních dat v repozitáři NDK

Digitální repozitář NDK si klade za cíl dostát nárokům důvěryhodného digitálního repozitáře, tedy zachovat dnes uložená data v použitelné podobě i pro vzdálenou budoucnost. Základním rámcem pro vybudování důvěryhodného repozitáře je již zmíněný model OAIS. Repozitář bude pracovat s digitálními dokumenty ve formě archivních balíčků. Každý archivní balíček (AIP) bude, kromě vlastního uchovávaného dokumentu, obsahovat také popisné informace pro účely dlouhodobé ochrany, informace o balíčku samotném (např. informace, jak je zabalen) a popisné informace o celém balíčku. Popisná informace pro účely dlouhodobé ochrany pak obsahuje medatata, která informují o integritě, kontextualitě, historii a jednoznačné identifikaci dokumentu (fixity, context, provenance, reference information). Popisná informace o celém balíčku pak obvykle opakuje informace o kontextualitě a jednoznačné identifikaci balíčku, aby balíček byl nejen vhodně archivovaný, ale také aby byl vůbec vyhledatelný.

Vedle toho musí repozitář poskytnout uživatelům v určité míře další informace, které jim umožní archivované informaci porozumět (např. dokumentace popisující, že balíčky pocházejí z těch a těch digitalizačních projektů, obsahují určité sbírky s určitým typem dokumentů pocházejících z nějaké konkrétní doby apod.).

Na důvěryhodnosti repozitáře se nepodílí jen vhodný informační model, softwarová architektura nebo technická infrastruktura.

Deset základních principů důvěryhodného repozitáře vytváří rámec pro širší kontext důvěryhodného repozitáře [2]:

Repozitář:
- věnuje se trvalé správě digitálních objektů pro definovanou komunitu/definované komunity;
- musí prokázat organizační způsobilost pro tento úkol (tzn. vhodně financovaná, personálně zajištěná a řízená instituce);
- dostojí smluvním a právním požadavkům a splní povinnosti z nich vyplývající;
- má vypracovanou účelnou a účinnou metodiku;
- získává a zpracovává digitální objekty podle stanovených kritérií, která odpovídají jeho cílům a schopnostem;
- udržuje a zajišťuje dlouhodobou integritu, autenticitu a použitelnost spravovaných digitálních objektů;
- archivuje potřebná metadata o všech akcích, které byly s digitálními objekty v průběhu jejich uložení provedeny; zároveň shromažďuje související informace o vzniku, podpoře dostupnosti a využívání objektů před jejich vstupem do repozitáře;

- naplňuje potřebná kritéria pro zpřístupňování;
- má strategický program pro plánování ochrany;
- má odpovídající technickou infrastrukturu potřebnou k trvalému udržování a zabezpečení spravovaných digitálních objektů.

V praxi to znamená, že důvěryhodnost je určena fungováním repozitáře ve vztahu k těm, kdo ho financují, kdo do něj dokumenty vkládají a ve vztahu k jeho uživatelům. Důvěryhodnost je tedy činností prokázaná vhodnost k danému účelu. Ukazatelem důvěryhodnosti mohou být certifikace repozitáře, ale také trvající zájem a důvěra uživatelů. Repozitář musí být schopen uživatelům prokázat původ digitálních dokumentů a poskytovat jim je ve srozumitelné a použitelné podobě. Podstatný vliv na důvěryhodnost repozitáře má pak také způsob jeho organizace, management, plánování, financování, dokumentace, pracovní postupy.

Autenticita

Důležitou vlastností důvěryhodného repozitáře je podpora zachování autenticity. Autenticitou je myšleno prokázání původu dokumentu a toho, že kopie, kterou vidí uživatel, souhlasí s originálním dokumentem, tj. nic v něm nebylo během doby změněno. Systém musí uchovávat metadata o všech procesech, akcích a změnách, které se na určitém objektu odehrály v průběhu jeho uložení tak, aby bylo možné tyto změny sledovat až k původnímu dokumentu (např. po několika migracích apod.).

Metadata

Metadata jsou v jakémkoliv systému, který se snaží o dlouhodobou ochranu digitálních objektů, klíčová. Nejen metadata popisná potřebná k vyhledávání, daleko důležitější z hlediska funkcionality jsou tzv. metadata ochranná. Ta podporují a dokumentují proces dlouhodobé ochrany digitálních dat, zachycují všechny informace potřebné pro ochranné akce, informace o těchto akcích, celkový kontext a okolnosti vzniku digitálního objektu apod.

Popisná metadata (údaje o entitě, tj. autor, název, rozsah, popis apod.) budou muset reflektovat vývoj, který v této oblasti v NK ČR, MZK i v dalších knihovnách v ČR pro různé projekty za dobu jejich existence proběhl. Archivační systém musí být schopen si poradit v první řadě s formátem Dublin Core, MODS (Metadata Object Description Schema), MARCXML, a musí být flexibilní v případě jakékoliv další změny v oblasti popisných metadat. Je nutné si uvědomit, že typů dokumentů, které budou v systému uchovávány, bude více a ne pro všechny je vhodný jeden typ popisných metadat (jednotlivé obrazy, entity složené z více obrazů, audio, video apod.).

Termín **ochranná metadata** zahrnuje několik kategorií obyčejně užívaných k rozlišení typů metadat. Jsou to:

- *administrativní (včetně práv a povolení, historie provedených akcí apod.)*
- *technická*
- *strukturální*

Konkrétněji, pracovní skupina PREMIS nahlíží na ochranná metadata z hlediska podpory funkcí udržování životnosti, výkonnosti, dlouhodobé srozumitelnosti a „zobrazovatelnosti“, autentičnosti a identity digitálního objektu v kontextu ochrany. Životností je míněn bitstream archivovaných digitálních objektů neporušený a čitelný z média, na kterém je uložen. Zobrazovatelnost odkazuje na schopnost převedení bitstreamu do formy, kterou může člověk bez problémů číst, nebo s ní může pracovat počítač. Srozumitelnost označuje poskytování dostatečného množství informací, tj. zobrazovaný obsah je pochopitel-

ný cílovému uživateli. Ochranná metadata mohou sloužit jako vstup do procesů ochrany a zároveň zachycují výstup z těch samých procesů.
Zvláštní pozornost je věnována zdokumentování původu, tj. historii digitálního objektu a zdokumentování vztahů a souvislostí, hlavně vztahů mezi různými objekty v rámci úložiště. Ochranná metadata musí podporovat především plánování ochrany a provádění ochranných akcí.

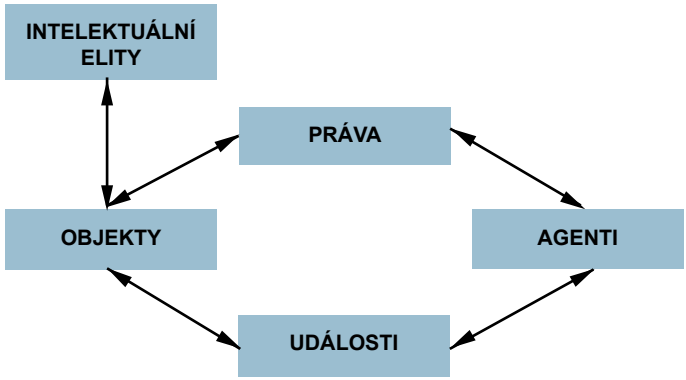
Je těžké vytyčit jasné hranice v tom, jaké typy informací spadají do sféry ochranných metadat. Konsensus platí pro pět oblastí informací relevantních pro ochranná metadata[3]:

Provenience: ochranná metadata by měla zaznamenávat informaci o historii digitálního objektu (ideálně dohledatelnou až k okamžiku vzniku objektu) a zároveň udržovat informace o jakýchkoliv změnách provedených na objektu v budoucnu (vlastnictví, držení objektu apod.).
Autenticita: ochranná metadata by měla obsahovat informaci dostatečnou k ověření toho, že archivovaný digitální objekt je tím, za co se vydává, nebyl změněn, zaměněn, ať už záměrně nebo nechtěně, aniž by o tom byl záznam v metadatech.
Ochranné aktivity: ochranná metadata by měla dokumentovat události provedené na objektech v rámci jejich ochrany včetně jejich dopadu na vzhled, vlastnosti a funkcionalitu objektu.
Technické prostředí: ochranná metadata by měla popisovat technické nároky jako je hardware, operační systém a softwarové aplikace potřebné k zobrazení a použití digitálního objektu ve stavu, v jakém je aktuálně uložen v archivačním systému.
Management práv: ochranná metadata by měla zaznamenat jakákoliv omezení vyplývající z autorského zákona, která by mohla ohrozit schopnost systému provést ochranná opatření na digitálním objektu a poskytnout objekt současným a budoucím uživatelům.

Ve workflow NDK bude většina metadat vznikat již v první fázi během digitalizace (nebo web harvestu) nebo bezprostředně po nich – metadata popisná, technická a administrativní. Toto se bude dít „mimo“ archivační systém. V další fázi, již v rámci archivačního systému, budou automaticky přidávána další metadata, např. identifikátory (vnitřní i vnější); metadata o formátech (pomocí registrů jako je PRONOM) a další metadata technická (nástroj JHOVE). Budou vytvořena další ochranná metadata dokumentující operace provedené při kompletaci definitivního informačního balíčku SIP (záznam o kontrolních operacích apod.). Před vstupem SIP balíčku do archivu budou metadata normalizována a validována, následně v archivním modulu budou podle situace připojována další metadata dokumentující životní cyklus balíčku AIP v archivu (použití, exporty, ochranné operace, atd.). Finální podoba metadat bude uložena v systému společně s digitálními objekty v rámci balíčku AIP. V průběhu času může vzniknout potřeba metadata v rámci AIP změnit (opravy, doplnění apod.), což zajistí archivační systém.

Repozitář NDK použije pro zabalení metadat do jednotlivých informačních balíčků formát METS, pro vyjádření ochranných metadat v rámci METS bude vyžadován formát PREMIS (The Preservation Metadata : Implementation Strategies). PREMIS je metadatový formát pro dlouhodobou archivaci založený na modelu OAIS a jako takový je široce využívaný v komunitě zabývající se dlouhodobou ochranou digitálních dat. Soupisem elementů formátu a jejich využití je PREMIS Data dictionary. Formát, díky svému datovému modelu (viz obrázek), dokáže tak říkajíc naplnit části METSu, určené pro administrativní metadata. Jde o část METS nazvanou amdSec (tj. administrativní metadata).

Sekce administrativních metadat „amdSec“ má sama další čtyři části:
a) techMD – technická metadata
b) rightsMD – administrativní, případně legislativní práva k objektům
c) sourceMD – popis původce údajů obsažených v METS dokumentu
d) digiprovMD – metadata spojená s digitálními zdroji



Samotná implementace PREMISu do METSu bude následující, běžně využívaná v mnoha knihovnách a zároveň přímo doporučená radou pro formát PREMIS.

METS část amdSec	části formátu PREMIS
techMD – technická metadata	PREMISobject + případné další (MIX)
rightsMD – administrativní práva legislativní práva	PREMISrights METSrights + PREMISagent
digiprovMD – metadata o událostech	PREMISevents + PREMISagent

Systém dlouhodobé ochrany implementovaný v repozitáři NDK musí ovšem také podporovat použití jiných formátů metadat pro různé typy archivovaných digitálních objektů, což spojení METS a PREMIS formátu velmi dobře umožňuje.

Legislativa

NDK se pohybuje v mantinelech platné české legislativy pro danou oblast, kterou představují především:

Knihovní zákon, který definuje postavení a funkce NK ČR i MZK v rámci systému knihoven v ČR.

Právo úplného povinného výtisku a z něho plynoucí povinnost dlouhodobé ochrany a zpřístupnění neperiodických publikací pro NK ČR i MZK definuje **Zákon o neperiodických publikacích** [4], totéž pro periodika definuje **Tiskový zákon** [5]. Ani jeden z uvedených zákonů neřeší problematiku povinného odevzdávání síťových publikací, které je nutným předpokladem jejich dlouhodobé ochrany a zpřístupnění. Návrh samostatného zákona připravený podle vzoru zahraniční legislativy pro danou oblast je momentálně ve stadiu schvalovacího řízení na MK ČR.

Autorský zákon [6] ve svém novelizovaném znění sice umožňuje lokální zpřístupnění většiny digitalizovaných i born digital dokumentů v prostorách NK ČR, MZK a dalších knihoven na území ČR, stále však představuje značnou bariéru pro široké zpřístupnění národního kulturního dědictví v digitální formě. Nejedná se o náš specifický problém, ale o problém mezinárodní.
Širší národní kontext vytváří **Státní informační a komunikační politika** [7], platná i připravovaná mezinárodní (především evropská) legislativa a různé strategie i doporučení pro danou oblast.

Závěr: Aktuální stav a plánovaný postup

NK ČR spolu s MZK připravují v rámci IOP (Smart Administration) projekt se dvěma hlavními cílovými liniemi:

- urychlení digitalizace (dvě digitalizační centra v Praze a v Brně, nasazení masové digitalizace);
- dlouhodobá ochrana digitálních objektů a přístup k nim (důvěryhodný digitální repozitář umístěný ve dvou geograficky odlišných lokalitách: Praha a Brno).

Pro konkrétnější představu o obsahu i rozsahu zmíněného projektu uvádíme několik čísel: jádro českého národního kulturního dědictví (dokumenty publikované na našem území od roku 1801 včetně + historické dokumenty do roku 1800 uložené v českých knihovnách) tvoří přibližně 1,2 milionu dokumentů, což představuje 350 milionů stránek. Digitalizace tohoto množství současným tempem by trvala zhruba 300 let, během nichž by se řada dokumentů vytištěných na kyselém papíře a/nebo často využívaných dostala do stavu, kdy by je nebylo možné vůbec digitalizovat, a náš stát by tak nenávratně ztratil velmi důležitou část svého kulturního dědictví. Projekt umožní digitalizovat těchto 350 milionů stránek během 20 let. Nejohroženější a nejvyužívanější dokumenty (většinou noviny) by měly být digitalizovány během prvních pěti let projektu v letech 2009 - 2014.

Výsledky projektu budou následující:

- Digitalizace dokumentů vydaných v a po roce 1801: 540 000 dokumentů, 137 milionů stran;
- Digitalizace historických dokumentů vydaných do roku 1800: 20 000 dokumentů, 9 milionů stran;
- WebArchiv: sklízení a archivace 5 miliard souborů;
- Důvěryhodný digitální repozitář (certifikovaný interním i externím auditem);
- Uživatelsky příjemný a „customizovatelný“ přístup k digitálnímu obsahu pro různé uživatele.

Celkový rozpočet celého projektu by měl být 706 milionů Kč (85 % podpora, 15 % spoluúčast), na úrovni IOP však existuje převis poptávky po financích na straně projektů nad objemem disponibilních prostředků, proto se již hovoří o redukcích.

V současné době připravuje tým NDK podklady pro Studii proveditelnosti, která je povinnou součástí projektů předkládaných v rámci Smart Administration. Dle posledních informací sdělených při zasedání Výboru pro koordinaci Smart administration by měla být výzva k podání projektů obsahující i linii 1.1.d), do níž tematicky spadá i náš projekt Vytvoření Národní digitální knihovny, vyhlášena na přelomu května a června. Pokud se tak skutečně stane, navzdory časovým skluzům ve vyhlašování výzev, stále ještě existuje naděje na dodržení vytčeného harmonogramu:

Číslo	Název etapy	Začátek	Konec
1	**Přípravná fáze projektu**	1.1.2007	31.12.2009
2	**Investiční fáze č. 1** (vybudování digitalizačních pracovišť a centrálního digitálního repozitáře – stavební část + výběr dodavatelů technologií)	1.1.2010	31.12.2010
3	**Investiční fáze č. 2** (vybavení digitalizačních pracovišť a centrálního repozitáře, uživatelsky vlídné a diferencované zpřístupnění –- technologická část)	1.1.2011	30.6.2011
4	**Provozní fáze projektu** (provoz digitalizačních pracovišť a centrálního repozitáře, uživatelsky vlídné a diferencované zpřístupnění)	1.7.2011	31.7.2014
5	**Ukončení projektu**	31.8.2014	31.12.2014

Literatura:

[1] *Florida Digital Archive* [online]. Gainesville, FL : Florida Digital Archive, c2003 [cit. 2009-05-09]. Recommended Data Formats for Preservation Purposes in the Florida Digital Archive. Dostupný z WWW: <http://www.fcla.edu/digitalArchive/pdfs/recFormats.pdf>.

[2] *OCLC* [online]. Dublin, Ohio : OCLC, 2008 [cit. 2009-05-09]. PREMIS (PREservation Metadata : Implementation Strategies) Working Group. Dostupný z WWW: <http://www.oclc.org/research/projects/pmwg>.

[3] LAVOIE, Brian; GARTNER, Richard. *Preservation Metadata* [online]. Heslington (York, GB) : Digital Preservation Coalition, 2005 [cit. 2009-05-11]. Dostupný z WWW: <http://www.dpconline.org/docs/reports/dpctw05-01.pdf>.

[4] Česko. Zákon č. 46/2000 ze dne 22. února 2000 o právech a povinnostech při vydávání periodického tisku a o změně některých dalších zákonů (tiskový zákon). In *Sbírka zákonů České republiky*. 2000, částka 17, s. 586-593. Dostupné též z WWW: <http://aplikace.mvcr.cz/archiv2008/sbirka/2000/sb017-00.pdf>.

[5] Česko. Zákon č. 37/1995 ze dne 8. února 1995 o neperiodických publikacích. In *Sbírka zákonů České republiky*. 1995, částka 8, s. 459-460. Dostupné též z WWW: <http://www.lexdata.cz/lexdata/sb_free.nsf/c12571d20046a0b20000000000000000/c12571d20046a0b2c12566d4007447-f9?OpenDocument>.

[6] Česko. Zákon č. 216/2006 ze dne 25. dubna 2006, kterým se mění zákon č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, a některé další zákony. In *Sbírka zákonů České republiky*. 2006, částka 72, s. 2707-2726. Dostupné též z WWW: <http://aplikace.mvcr.cz/archiv2008/sbirka/2006/sb072-06.pdf>.

[7] Česko. Vláda. Státní informační a komunikační politika [Usnesení vlády ČR č. 265 ze dne 24. března 2004 ke Státní informační a komunikační politice]. Text usnesení dostupný také z WWW: <http://knihovnam.nkp.cz/docs/SIKP_def.pdf>.