ProArc

– open source řešení pro produkci a archivaci digitálních dokumentů

Martina NEZBEDOVÁ
Knihovna AV ČR, v. v. i., Praha
nezbedova@knav.cz

INFORUM 2015: 21. ročník konference o profesionálních informačních zdrojích
Praha, 26. – 27. 5. 2015

Abstrakt

Produkční a archivační systém ProArc je volně dostupný nástroj na výrobu a editaci popisných, technických a administrativních metadat k digitalizovaným i born digital dokumentům. Systém umožňuje vytvoření importních dat pro systém Kramerius. Jedním z výstupů jsou PSP balíčky vytvořené podle NDK standardů Národní knihovny. Na vývoji systému ProArc spolupracuje Knihovna AV ČR, v. v. i. s firmou INCAD, která zajišťuje analytické a programátorské práce. Od roku 2013, kdy byl tento systém plně nasazen v Digitalizačním centru KNAV, v něm bylo zpracováno více než 400 tisíc stran různých typů dokumentů. Systém začínají v současné době využívat i další instituce. Ve vývoji je archivační část, která umožní propojení na systém Archivematica. Příspěvek představí ProArc z pohledu zpracovatele.

Systém ProArc je jedním z výstupů projektu Česká digitální knihovna. Cílem celého projektu je vytvoření jednotného rozhraní nad digitálními knihovnami v ČR pro koncové uživatele a poskytování dat pro mezinárodní projekty. Projekt je financován z programu NAKI Ministerstva kultury ČR.

Produkční a archivační systém ProArc je volně dostupný nástroj na výrobu a editaci popisných, technických a administrativních metadat k digitalizovaným i born digital dokumentům.

Na vývoji systému ProArc úzce spolupracuje Knihovna AV ČR, v. v. i., hlavní koordinátor vývoje, s firmou INCAD, která zajišťuje analytické a programátorské práce. Systém ProArc je jedním z výstupů projektu Česká digitální knihovna. Cílem celého projektu je vytvoření jednotného rozhraní nad digitálními knihovnami v ČR pro koncové uživatele a poskytování dat pro mezinárodní projekty. Projekt je financován z programu NAKI Ministerstva kultury ČR.

Od roku 2013, kdy byl tento systém plně nasazen v Digitalizačním centru KNAV, v něm bylo zpracováno více než 400 tisíc stran různých typů dokumentů. Systém začínají v současné době využívat i další instituce např. Studijní a vědecká knihovna v Hradci Králové.

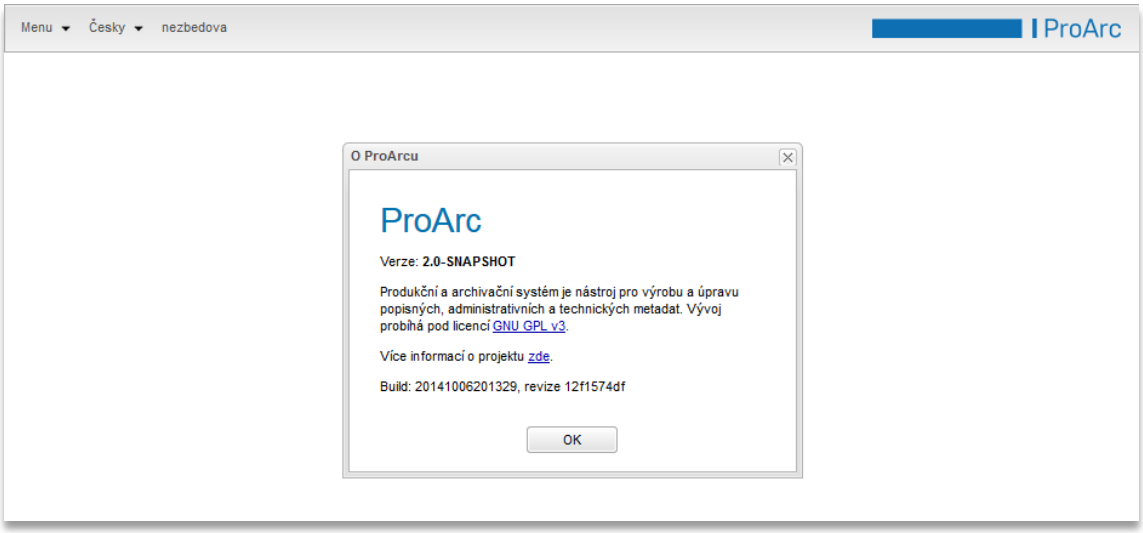
Produkční část systému ProArc je nástroj na výrobu a úpravu popisných, administrativních, technických metadat a jejich editaci. V tomto systému je možné nejen zakládat zcela nové objekty, ale i používat metadata z externích systémů (např. Aleph, Registr digitalizace). Systém ProArc podporuje dávkové a hromadné úpravy a také částečně automatizovanou produkci digitálních dokumentů. Systém ProArc automaticky generuje UUID pro jednotlivé objekty a umožňuje vyhledávání v metadatech. Rovněž je možné následné rozšíření o nové datové modely.

Systém ProArc podporuje standardy Národní knihovny ČR pro digitalizaci (plná podpora MODS, administrativních a technických metadat) a přidělování URN:NBN. Dodržování těchto standardů umožní spolupracovat s Národní knihovnou při sdílení dat. Plnění těchto standardů je požadováno v dotačním programu VISK 7 a v krajských digitalizačních projektech. Typy těchto dokumentů jsou zakládány ve formulářích označených jako NDK (Národní digitální knihovna). Zároveň je pro ně připraven export v podobě PSP balíčku dle standardů pro NDK. Systém ProArc je plně kompatibilní se systémem Kramerius.

Systém ProArc je určen i pro zpracování born digital dokumentů. Tyto dokumenty je po zpracování možné exportovat jak pro systém Kramerius, tak i jako metadata do společné bibliografické databáze akademií věd Visegradské čtyřky CEJSH.

Systém ProArc je open source, který je vystavěn na volně dostupných řešeních Fedora Commons repozitory, Java a PostgreSQL. Systém ProArc je webová aplikace, která využívá lokální server. Pro generování grafického formátu JPEG2000 využívá v rámci standardů NDK program Kakadu, ale podporuje i užití jiných programů. Pro potřeby OCR je využit komerční ABBYY Recognition Server, který umožňuje generovat formát ALTO XML. Technickou podporu zajišťuje firma INCAD.

Projektová dokumentace, informace o aktuálním stavu vývoje a řešených issues jsou spolu s instalačním balíčkem verze ProArc 2.0 NDK umístěny na adrese https://code.google.com/p/archivacni-system/wiki/ProArc.



V úvodní obrazovce se v horní liště nachází rozbalovací roletka Menu, která skrývá základní složky Import a Úložiště s jednotlivými pořadači a složku Zařízení obsahující formuláře odpovídající NDK standardům pro výrobu technických metadat (informace o skenerech a parametrech skenování).

Tvorba metadat začíná vytvořením databáze možností používaných technických metadat vyplněním různých typů používaných skenerů a aktuálních parametrů skenování. Tyto údaje lze doplňovat nebo měnit podle potřeby digitalizace. Nastavení je společné pro všechny pracovníky na jedné instalaci ProArcu.

Dalším krokem je založení objektu, ke kterému budou následně připojeny soubory s hotovými popisnými metadaty. Jednotlivé objekty, které lze založit jsou:

NDK Periodikum

NDK Ročník

NDK Číslo

NDK Příloha periodika

NDK Článek

NDK Obrázek/Mapa – vnitřní část

NDK Vícedílná monografie

NDK Svazek monografie

NDK Příloha monografie

NDK Kapitola

NDK Kartografický dokument

NDK Hudebnina

K4 Periodikum

K4 Ročník

K4 Výtisk

K4 Monografie

K4 Monografie – volná část

Strana

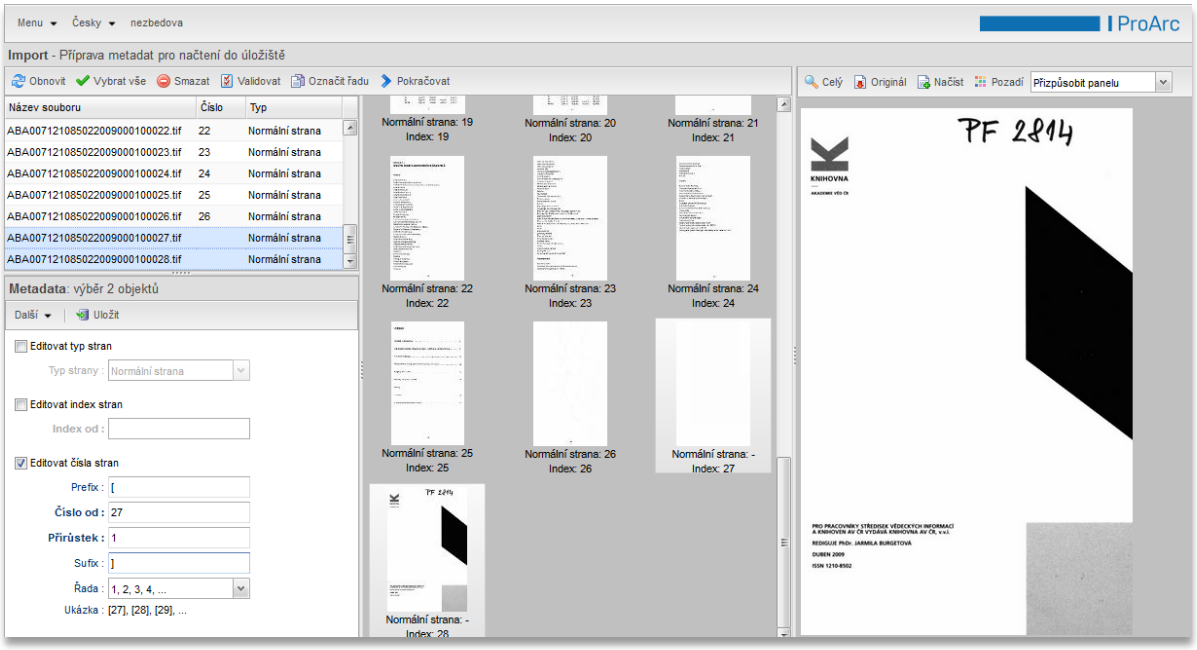
E-článek

Formuláře s označením NDK obsahují položky pro vyplnění odpovídající NDK standardům a nad poli s povinností vyplnění (Mandatory) je prováděna validace vyplnění těchto polí. Tyto objekty jsou exportovány jako PSP balíčky, ale je u nich možný i export pro K4 v xml. Formuláře s označením K4 obsahují vybraná pole MODS (lze si přepnout i na plný MODS). I zde probíhá validace povinných polí. Exportem je pouze xml.

Založit lze jak prázdný formulář pro následné ruční vyplnění, tak i formulář s poli vyplněnými ve vybrané databázi (např. Registr digitalizace, Aleph). Pro výběr slouží jeden z identifikátorů (např. čárový kód, ISSN, číslo ČNB).

Dalším krokem je výběr adresáře s digitalizovanými daty pro přípravu metadat z importního adresáře na serveru, ke kterému se připojí technická metadata a vybrané soubory jsou načteny do sestavy určené pro tvorbu popisných metadat.

Obrazovka s názvem Tvorba metadat je velice variabilní a lze si ji přizpůsobit podle typu zpracovávaných dokumentů nebo pracovních zvyklostí. Popisná metadata lze tvořit po jednotlivých stranách nebo výběrem souvislé řady stran nebo výběrem různých jednotlivých stran. Popisná metadata je možné přiřazovat po řádcích obsahujících název souboru, v malých náhledech, ve velkých náhledech nebo jejich kombinací.

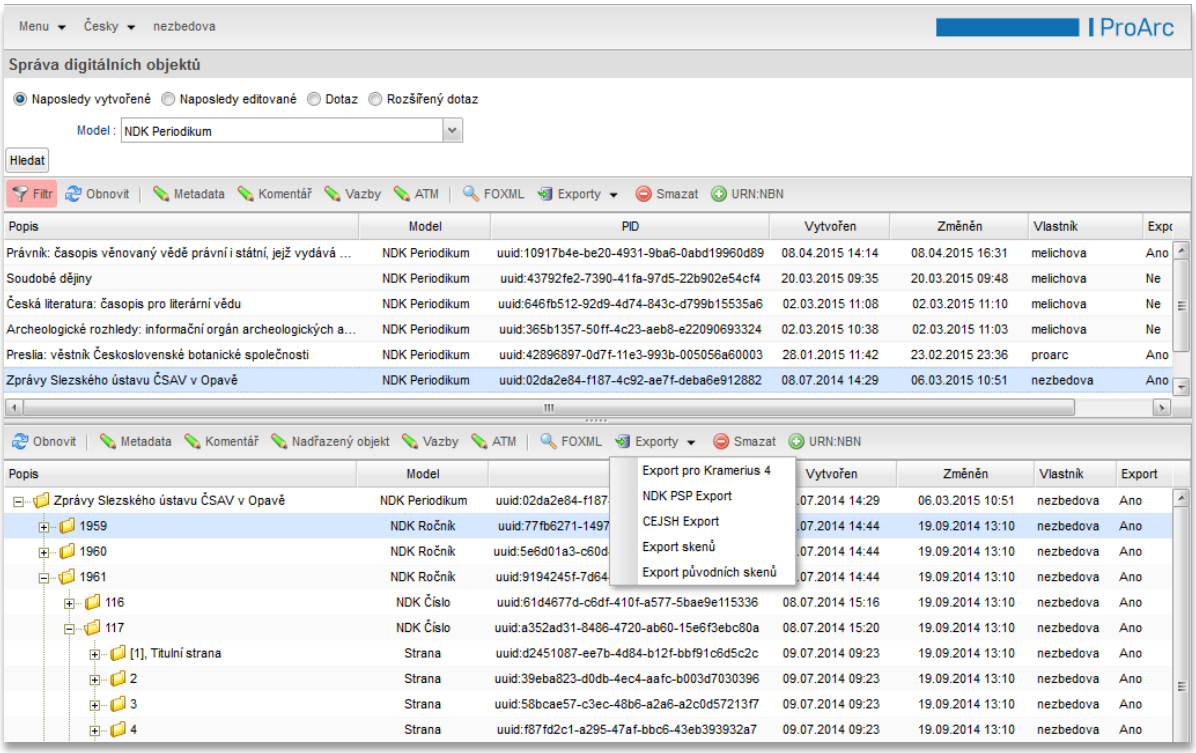


Pro výrobu popisných metadat jsou připraveny typy stran odpovídající standardům NDK. Pro usnadnění tohoto popisu je automaticky přednastaven typ strany – normální strana. K doplnění ostatních typů stran slouží roletka s předepsanými typy stran. Pro doplňování čísel jednotlivých stran jsou připravené číselné řady arabských i římských číslic a písmenná řada, doplněné možností prefixu a sufixu. Před připojením souborů k nadřazenému objektu prochází všechny soubory validací na vyplnění čísla strany. I po připojení k nadřazenému objektu lze typy i čísla stran editovat, prohlížet si náhledy jednotlivých stran nebo tyto strany přesouvat.

Užitečnou funkcí je možnost Označit řadu, tj. řadu po sobě jdoucích stran s vyplněnými popisnými metadaty a tyto následně aplikovat na další řadu objektů. Při popisu výtisků, které mají vždy stejné typy stran a jsou stejně číslované, to velice zjednoduší a zrychlí jejich popis, protože jednou vytvořená popisná metadata se jen zkopírují na další výtisky.

Důležitou součástí systému ProArc je správa digitálních objektů, která obsahuje všechny založené typy objektů, ve kterých lze vyhledávat. U jednotlivých objektů jsou informace o datu založení, vlastníku záznamů a případném exportu. V této databázi lze vyhledávat podle různých kritérií. Rovněž je zde možná editace ať už jednotlivých objektů nebo hromadná přepnutím do obrazovky vazby.

Ze systému ProArc lze po přidělení URN:NBN exportovat NDK PSP balíčky plně odpovídající standardům NDK. Dalšími možnostmi jsou export pro K4, export skenů, export původních skenů a pro born digital dokumenty export pro CEJSH.



Prohlížení a editace již hotových objektů umožňuje obrazovka Vazby, ve které je také možné přesouvat jednotlivé soubory i celé řady souborů jak v jednotlivých objektech, tak i mezi objekty navzájem. Tím je zajištěna náprava možných chyb.

V systému ProArc je také možné zpracovávat born digital dokumenty. Po založení e-článku se připojí již hotová metadata z připojené databáze Knihovna AV ČR Analytika a v dalším kroku se přidá plný text v pdf formátu. Exportovat lze jak plný text s přidanými metadaty do Krameria, tak jen metadata formou Export CEJSH.

Produkční část systému ProArc je velice vhodným a uživatelsky příjemným nástrojem pro výrobu metadat a jejich následného využití jak v systému Kramerius, tak i ke vzájemnému sdílení s ostatními institucemi, které dodržují standardy NDK. Systém se stále vyvíjí a vylepšuje pro širší využití v praxi (např. popis starých tisků).

Informace o systému ProArc jsou na https://code.google.com/p/archivacni-system/

K dispozici je i diskusní skupina na http://groups.google.com/group/proarc-users