ProArc– open source řešení pro produkci a archivaci digitálních dokumentů

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 ProArc je možné rozšířit o volitelnou komponentu pro sledování průběhu digitalizacem – RDflow. ProArc je založený na Fedora Commons repository, podporuje standardy Národní knihovny ČR pro digitalizaci a je kompatibilní se systémem Kramerius.

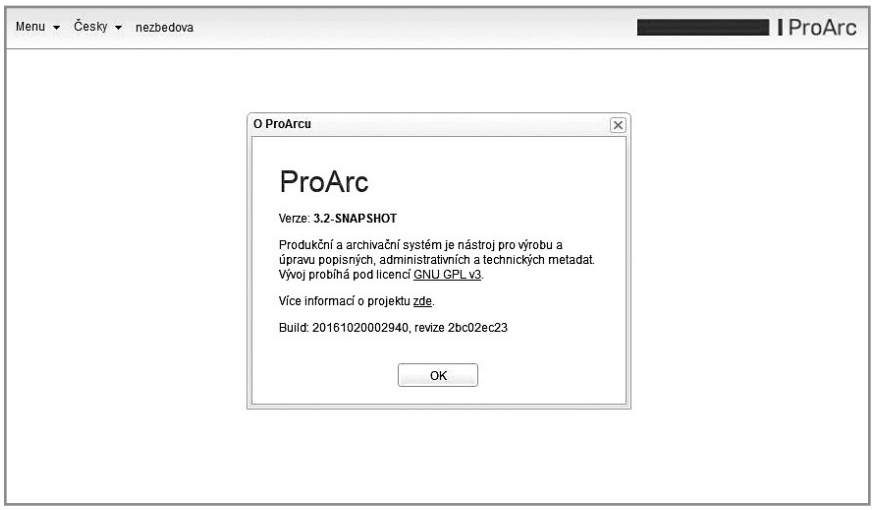
Digitální dokumenty se stávají samozřejmou součástí našich životů. Při masivním rozšíření digitalizace v knihovnách je kladen důraz nejen na samotné obrazové soubory, ale i na výrobu, úpravu a archivaci metadat digitálních dokumentů. Nelze zapomínat ani na potřebu sledování celého pracovního procesu digitalizace. Pro usnadnění těchto činností spojených s digitalizací je vyvíjen systém ProArc.

Produkční a archivační systém ProArc je volně dostupný nástroj na výrobu a editaci popisných, technických a administrativních metadat k digitalizovaným i born digital dokumentům. ProArc je založený na Fedora Commons repository, podporuje standardy Národní knihovny ČR pro digitalizaci a je kompatibilní se systémem Kramerius.

Systém ProArc byl vyvíjen jako součást projektu „Česká digitální knihovna a nástroje pro zajištění komplexních digitalizačních procesů", jenž byl financován z Programu aplikovaného výzkumu a vývoje národní a kulturní identity (NAKI) Ministerstva kultury ČR. Po skončení tohoto projektu je ProArc i nadále vyvíjen pod záštitou Knihovny AV ČR v. v. i. v úzké spolupráci se Studijní a vědeckou knihovnou v Hradci Králové, Městskou knihovnou v Praze, knihovnou Fakulty sociálních věd Univerzity Karlovy. Analytické a programátorské práce jsou zajišťovány firmou INCAD (pobočkou Search Technologies).

Systém ProArc také obsahuje volitelnou komponentu pro tvorbu pracovních úkolů a sledování jednotlivých kroků na digitální lince – RDflow.

Popis systému ProArc, dokumentace, instalační balíček, informace o aktuálním stavu vývoje a řešených issues jsou umístěny na adrese https://github.com/proarc/proarc/wiki.



Obr. 1 Úvodní obrazovka

Systém ProArc je open source, který je vystavěn na volně dostupných řešeních. V nejnovější verzi 3. 2 to je úložiště Fedora Commons 3.8.1 , Java Oracle JDK 1.8 a PostgreSQL 9. 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.

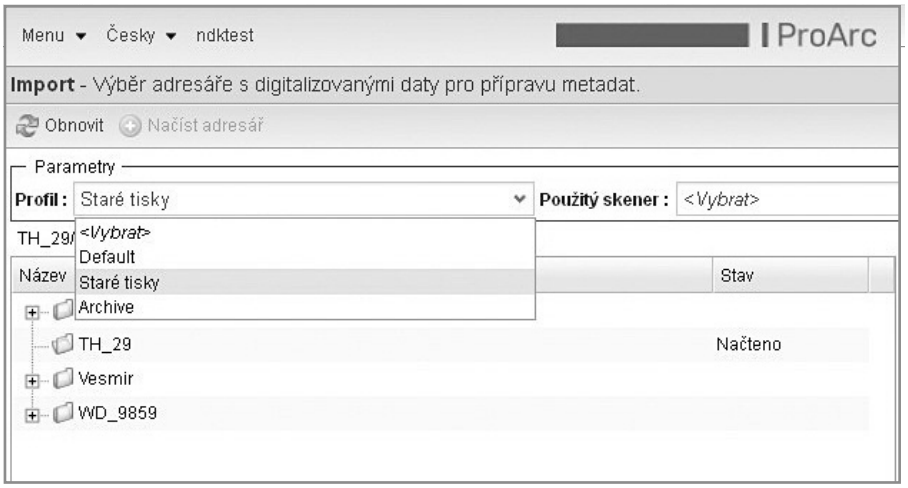
Systém ProArc obsahuje produkční a archivační část a volitelnou komponentu RDflow.

Produkční část

Produkční část systému ProArc je nástrojem na výrobu a úpravu popisných, administrativních a technických metadat. Lze v něm nejen zakládat zcela nové objekty, ale i používat metadata z externích systémů (např. Aleph, Registr digitalizace, Souborný katalog Národní knihovny). Metadata jsou editovatelná jak v připravených formulářích, které odpovídají definici metadatových formátů pro digitalizaci periodik a monografií, tak i v editovatelném xml.

Systém ProArc automaticky generuje UUID pro jednotlivé objekty. Pro zpracování jednotlivých předloh si lze vybrat z několika profilů a modelů podle typu zpracovávaných předloh.

Výběrem profilu se zvolí možnosti popisných metadat u typu strany. Základní profil *Default* plně odpovídá standardům NDK. Na základě požadavků badatelů, pracujících se starými tisky, vznikl volitelný profil *Staré tisky*, ve kterém se nachází některé specifické typy stran usnadňující badatelskou činnost (např. Dedikace). Zatím posledním profilem je *Archive*, který slouží ke znovunačtení souboru exportovaného z ProArcu typem exportu *Archivace.*



Obr. 2 Výběr profilů pro zpracování metadat.

Volba modelu se řídí typem předlohy zakládaného objektu. Ve formuláři vybraného modelu lze tvořit metadata k jednotlivým nadřazeným objektům. S typem zvoleného modelu souvisí i možnosti exportu. K dispozici jsou tyto modely:

- 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
- STT Svazek monografie
- STT Příloha monografie
- STT Strana
- Strana
- eČlánek

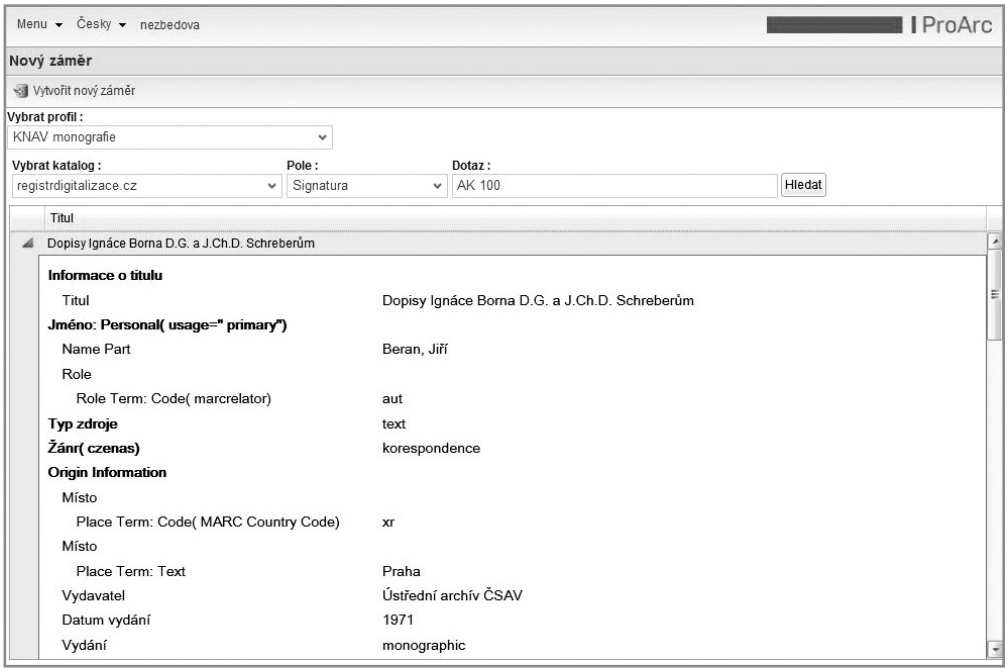
Formuláře modelů NDK mají položky odpovídající NDK standardům a nad povinně vyplnitelnými poli (Mandatory) je prováděna validace vyplnění těchto polí. Tyto objekty jsou exportovány jako PSP balíčky, ale je u nich možný i export pro K4 v xml.

Formuláře modelů K4 obsahují vybraná pole MODS. I zde probíhá validace povinných polí. Exportem je pouze xml.

Formuláře modelů pro staré tisky (STT) mají položky shodné s NDK formuláři, ale protože neobsahují OCR, je u nich možný jen K4 export, obsahující xml. Do některé z příštích verzí je připravován model STT Vícedílná monografie, který umožní, stejně jako tomu je v případě NDK Vícedílné monografie, spojit jednotlivé díly starých tisků pod jeden titul.

Formulář pro eČlánek má možnost volby recenzovaného/nerecenzovaného článku a obsahuje položky s povinně vyplnitelnými poli, nad kterými je prováděna validace vyplnění těchto polí. Po založení e-článku se připojí již hotová metadata z připojené databáze Knihovna AV ČR Analytika nebo je lze vyplnit ručně. 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 a Crossref Export.

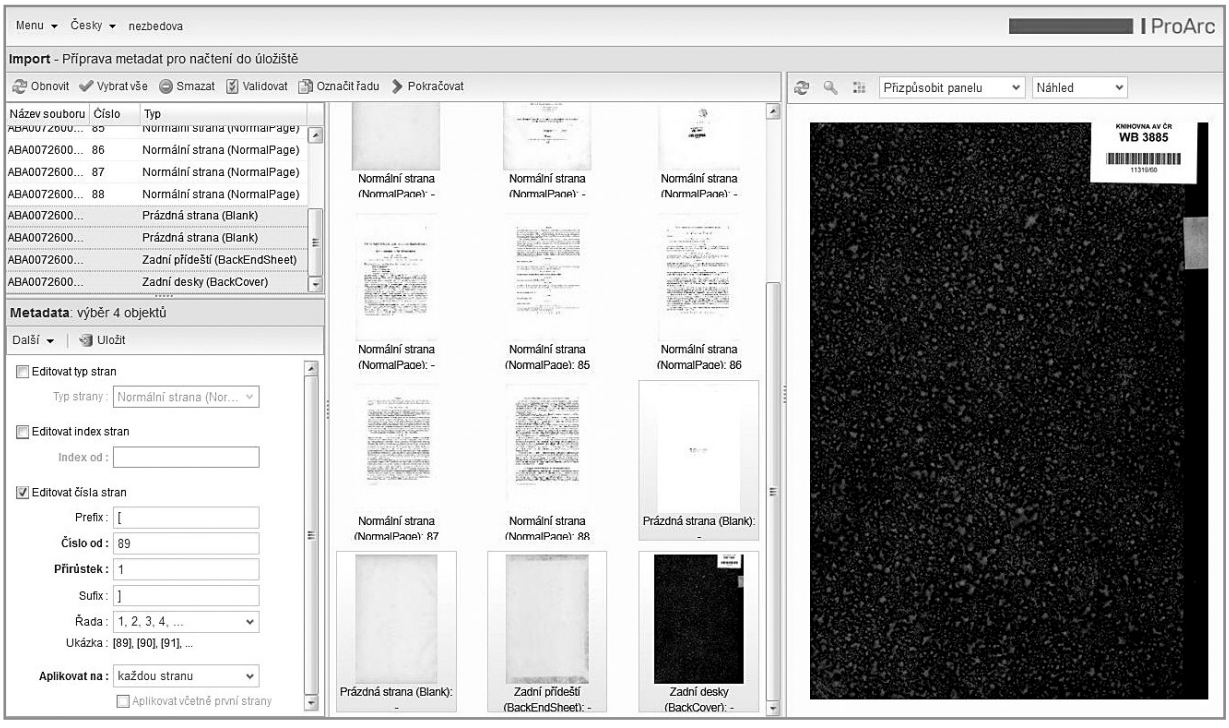
Produkce digitálních dokumentů v systému ProArc probíhá částečně automatizovaně, předností je možnost dávkových a hromadných úprav. Rychlou paginaci stran umožňuje hromadné přiřazování arabských i římských číslic, kombinací číslic a písmen nebo použití hranatých závorek. V případě foliace lze využít hromadného popisu ob stranu – recto/verso (1r, 1v). ProArc tímto



Obr. 3 Možnosti exportů

způsobem umožňuje popis stran až po osmerkách (přiřadí stejné údaje první a sedmnácté straně). Hromadně lze také měnit typy za sebou jdoucích stran (např. *Obsah*) i vybraných stran z celého dokumentu (např. *Prázdná strana*). Pracovní prostředí pro tvorbu metadat si lze přizpůsobit podle typu zpracovávaných dokumentů nebo pracovních zvyklostí. Uživatelsky příjemná je i možnost volby podbarvení jednotlivých stran.

Jednotlivé strany lze přesouvat. Kontrolu umožňuje během popisu stran náhled aktuálně zpracovávané strany. Před připojením stran s popisnými metadaty k nadřazenému objektu probíhá validace vyplnění všech čísel stran.



Obr. 4 Popis stran

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

Před exportem je lze přidělit URN:NBN nejen vlastním digitalizovaným titulům, ale i předlohám zpracovávaných pro jiné knihovny.

Z ProArcu lze provádět export NDK PSP (balíček plně odpovídající standardům NDK), export pro Kramerius 4 (xml), CEJSH export (e-články), export původních skenů a export pro archivaci.

Archivační část

U všech NDK modelů vytvořených v produkční části ProArcu lze použít export *Archivace*. Tím dojde k vytvoření exportního balíčku určeného pro archivaci. Výsledný formát tohoto balíčku vychází z NDK PSP, ale není totožný. Je popsán souborem mets.xml ve formátu METS. Rozpoznatelnost je možná na základě identifikátorů UUID. Fedora datastreamy jsou kopírovány do samostatných adresářů pojmenovaných podle DS ID a referencovány ze sekce mets:fileSec. Popisná metadata (DC, MODS) jsou v sekci mets:dmdSec. Hierarchie digitálních objektů je rozdělena do samostatných balíků podle pravidel NDK. Vazby mezi objekty jsou popsány v sekci mets:structMap.

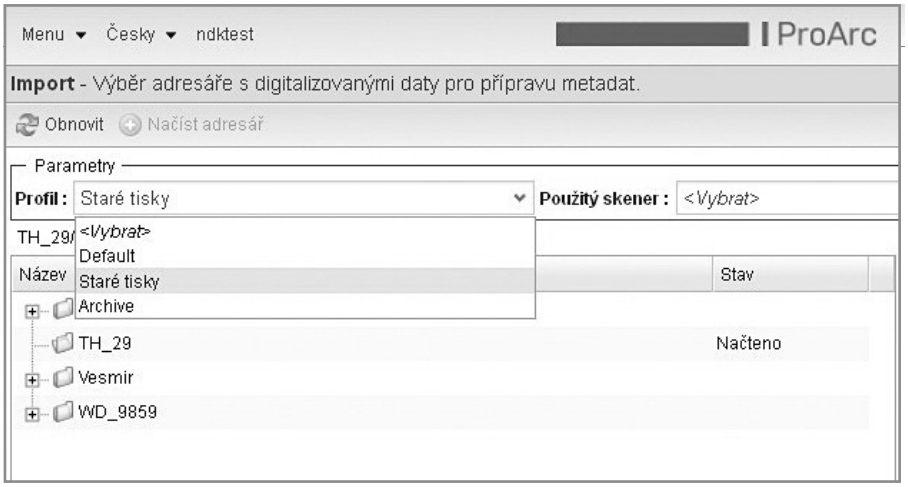
Adresáře obsahují data:

ALTO (METS ALTO soubory)
AUDIT (Historie změn)
FOXML (Kopie FOXML z úložiště Fedora)
FULL (Img)
NDK_ARCHIVAL (Img)
NDK_USER (Img)
PREVIEW (Img)
RAW (Img – původní skeny)
RELS-EXT (popis vazeb RDF)
TEXT_OCR (Txt)
THUMBNAIL (Img)
mets.xml (metadatový popis objektu)

Archivační balíček lze opětovně nahrát do produkční části ProArcu pomocí modelu *Archive.* Jde o plnohodnotné objekty uvnitř úložiště. Po importu archivačního balíčku lze upravovat metadata, která je třeba znovu archivovat exportem, protože soubory, ze kterých byla tato data načtena, zůstanou beze změn.

Volitelná komponenta RDflow

Ve verzi 3.1 došlo k rozšíření systému ProArc o volitelnou komponentu RDflow, která je pracovním prostředím pro sledování průběhu digitalizace. Umožňuje nejen plánování a následné sledování digitalizace zvolené předlohy, ale také sledování jednotlivých úkolů napříč všemi digitalizovanými předlohami. Výhodou této komponenty je její velká variabilnost. Toto workflow lze nakonfigurovat podle specifických potřeb jednotlivých digitalizačních linek pomocí xml. Ukázkové workflow.xml je umístěno na adrese *https://github.com/proarc/proarc/wiki/Popis-workflow.xml.*



Obr. 5 Vytváření nového záměru, který se bude sledovat

V RDflow lze vytvářet jednotlivé Záměry. Záměr je souhrn všech akcí, které mají proběhnout na předloze. Jako předloha je brána jednotka nesoucí identifikátor shodný pro katalog, Registr digitalizace a produkční část ProArcu.

K jednotlivým záměrům lze vybrat jeden z profilů. Profil je přednastavený seznam všech úkonů, které mají proběhnout na předloze při digitalizaci. Jednotlivým úkonům jsou vytvořeny předem definované úkoly. Profily lze upravovat, nově sestavovat, přidávat nebo ubírat podle potřeb jednotlivých digitalizačních linek v xml dokumentu wokrflow.xml v adresáři $PROARC_HOME.

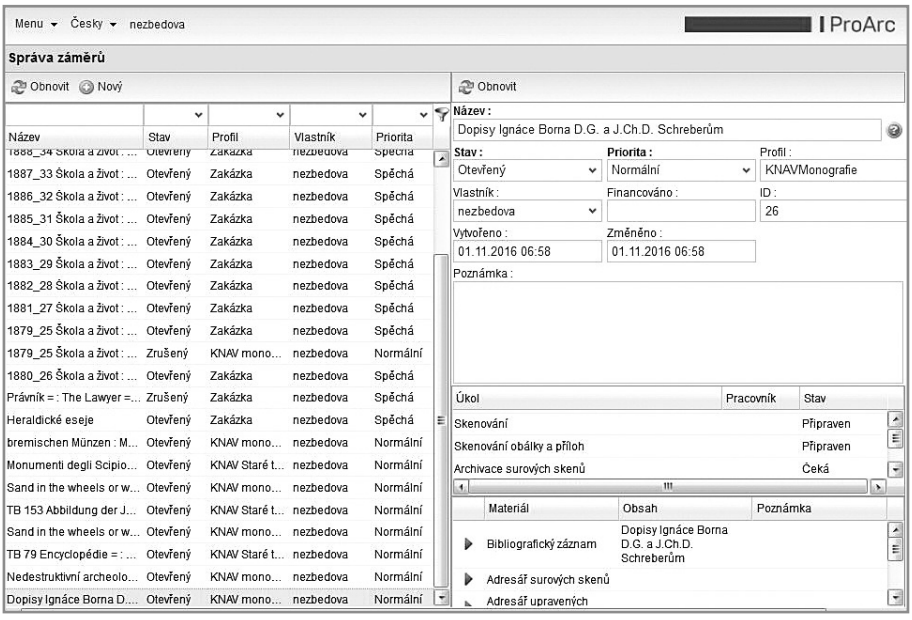
```
<!--  ******************** JOB: KNAV Monografie
********************************   -->
- <job name="KNAVMonografie" priority="2">
  <!--  *** Lokalizace ***   -->
  <title lang="cs">KNAV monografie</title>
  <title lang="en">KNAV Monograph</title>
  <hint lang="cs">Zpracování Monografie KNAV</hint>
  <hint lang="en">KNAV Monograph workflow</hint>
  <!--  *** Seznam úkolů ***   -->
  <step taskRef="task.RdImportProgress" />
  <step taskRef="task.correction" optional="true" />
  <step taskRef="task.pageCutting" optional="true" />
- <step taskRef="task.scan">
    <blocker taskRef="task.pageCutting" />
    <setParam paramRef="param.scan.dpi">400</setParam>
    <setParam paramRef="param.scan.imageColor">8 bit</setParam>
    <setParam paramRef="param.scan.fileFormat">tiff</setParam>
  </step>
- <step taskRef="task.scanOfCover">
    <blocker taskRef="task.pageCutting" />
    <setParam paramRef="param.scanOfCover.dpi">400</setParam>
    <setParam paramRef="param.scanOfCover.imageColor">8
      bit</setParam>
    <setParam paramRef="param.scanOfCover.fileFormat">tiff</setParam>
  </step>
```

Obr. 6 Ukázka části XML pro profil KNAV Monografie

Postup pro práci s komponentou RDflow

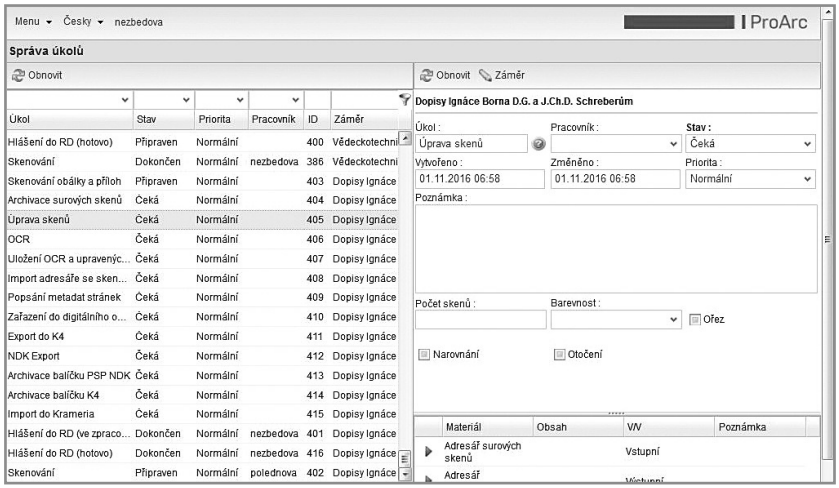
Protože se jedná o volitelnou komponentu, v základní instalaci ProArcu není a je třeba ji dodat. Přidává se xml, které obsahuje přehled všech úkolů a jejich sestavení do jednotlivých profilů.

Každá digitalizační linka má své postupy digitalizace. Nejprve je třeba stanovit úkoly, jejichž plnění u jednotlivých záměrů chceme sledovat. Základem je skenování, zpracování a zveřejnění. Po vytvoření nového záměru lze sledovat jednotlivé kroky na tomto záměru. Zda je předloha již naskenována, zpracována nebo zveřejněna. Jednotlivé kroky obsahují blockery, které zabraňují vyplnění úkolů mimo stanovený postup prací.



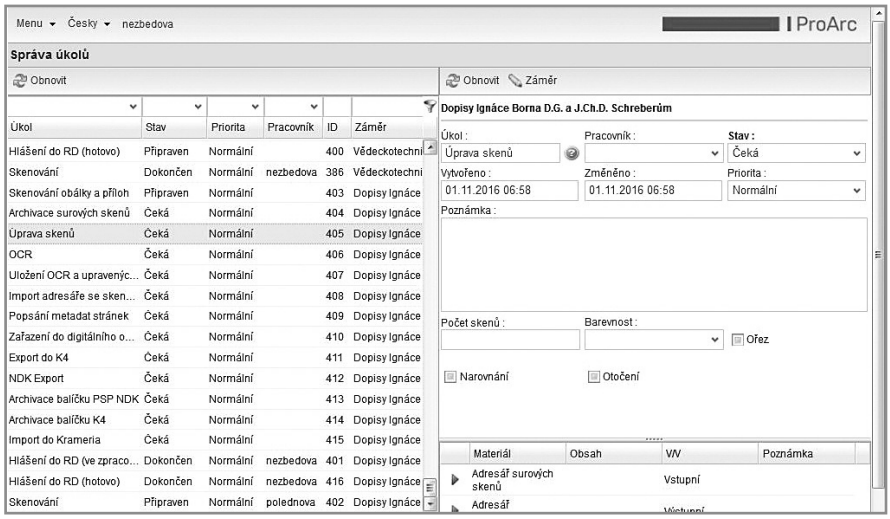
Obr. 7 Detail Správy záměrů. V levé části je seznam všech záměrů, v pravé jsou informace o vybraném záměru

Další možností je sledování jednotlivých úkolů na všech záměrech. Zde je možné zjistit, kolik předloh je již naskenovaných, zpracovaných, případně zveřejněných.



Obr. 8 Správa úkolů. V levé části je seznam všech úkolů, v pravé jsou informace o vybraném záměru.

Protože u jednotlivých úkolů mohou být přednastaveni pracovníci, je takto možné i plánování práce jednotlivců a sledování plnění úkolů. Vygenerovat lze tedy vše, co má pracovník již hotovo i počet zbývajících úkolů.



Obr. 9 Správa úkolů. V levé části je seznam všech jednoho pracovníka. Vpravo je vybraný záměr.

Již se připravuje další verze, ve které bude upravena a rozšířena volitelná komponenta RDflow o možnost sledování digitalizace periodik a vícedílných monografiív dvouúrovňové struktuře. Bude tak možné sledovat nejen práci na celém nadřazeném titulu, ale i na jednotlivých předlohách. Zároveň tak bude možné přehlednější sledování práce na doskenech a při postupném doplňování ucelených řad periodik.

Produkční část systému ProArc je velice vhodným a uživatelsky příjemným nástrojem pro výrobu metadat a jejich následného využití jak v Krameriovi, 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://github.com/proarc/proarc*.

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

Zdroj:

Dostupné na internete: <https://github.com/proarc/proarc>.

Dostupné na internete: <http://www.inforum.cz/sbornik/2015/17/>.

Mgr. Martina Nezbedová
nezbedova@knav.cz
Knihovna Akademie věd ČR