AiP Beroun s.r.o., http://www.aipberoun.cz, tel.: +420 311 611 237, fax: +420 311 611 238
Talichova 807, Beroun, 266 01; IČO: 25778943, DIČ: 25778943

ZDOKONALENÍ VIRTUÁLNÍHO BADATELSKÉHO PROSTŘEDÍ MANUSCRIPTORIA ROZŠÍŘENÍ SLUŽEB MANUSCRIPTORIA ROZŠÍŘENÍM NORMALIZOVANÉ INFORMACE O TYPU OBSAHU OBRAZŮ

Zpráva

AiP Beroun, autor: Ing. Tomáš Psohlavec

Obsah

1 Úvod o dokumentu

AiP Beroun uzavřela s Národní knihovnou České republiky Smlouvu o spolupráci ve výzkumu a vývoji: Zdokonalení virtuálního badatelského prostředí Manuscriptoria - rozšíření služeb Manuscriptoria rozšířením normalizované informace o typu obsahu obrazů.

1.1 Účel

Tento dokument tvoří zprávu k rozšíření aktuálních služeb Manuscriptoria začleněním normalizované informace o typu obsahu obrazů (text, iluminace, hudební notace, bordura, tabulka, diagram) a doplněním hledání dle této položky dle poptávky zadavatele č.j. 2295/KGŘ/11 ze dne 23.8.2011.

1.2 Předpokládaný čtenář

Tento dokument je určen především pro Zadavatele (NK ČR) a pro Řešitele úkolu (AiP Beroun) jako popis pilotního řešení. Dále je tento dokument určen všem, kteří se podílejí na rozvoji projektu Manuscriptorium jako uživatelé.

1.3 Termíny a konvence

Termíny a konvence použité v tomto dokumentu, pokud zde nejsou přímo vysvětleny, jsou popsány a definovány v dokumentu [2].

- MnS Manuscriptorium
- IIR Rozpoznání informací v obraze (Image Infromation Recognition)
- IIR Aplikace Aplikace pro rozpoznávání informací v obraze

1.4 Reference

V dokumentu se odkazujeme na následující zdroje:

[1] Guidelines for Electronic Text Encoding and Interchange, Representation of Primary Sources
http://www.tei-c.org/release/doc/tei-p5-doc/en/html/PH.html#PHFAX
[2] Manuscriptorium v.2.0 – analýza systému, AiP Beroun 2004

2 Úvod

NK ČR (Zadavatel) vyvíjí se svým subdodavatelem desktopovou aplikaci pro rozpoznávání informací v obrazech (IIR Aplikace). Zadavatel IIR Aplikaci poskytl Řešiteli. Řešitel zanalyzoval možnosti využití výstupů této aplikace v Manuscriptoriu a připravil pilotní řešení, na jehož základě lze posuzovat účelnost rozšíření služeb, které Manuscriptorium poskytuje koncovým uživatelům, případně Content Providerům.

3 Analytická fáze

V souvislosti se zamýšlenou implementací IIR do Manuscriptoria bylo nejprve nutno zanalyzovat možnosti zachycení požadovaných informací v metadatech. Vzhledem k tomu, že MnS pracuje s formátem TEI P5, byly přirozeně prozkoumány možnosti zápisu informací v TEI P5 přímo v elementu `<facsimile>`: využity byly podelementy `<zone>` způsobem, který popisuje následující příklad.

```
<surface xml:id="ID0011R">
  <desc>
    <label>11r</label>
  </desc>
  <zone xml:id="ID0001v__graphic" ulx="0" uly="0" lrx="1082" lry="1555">
    <graphic url="I/MVCHK_HR_13_II_A_13__2FMI/N1/HR_13_II_A_13__2FMIN10011R.JPG"/>
  </zone>
  <zone xml:id="ID0011R__pic_1" ulx="63" uly="51" lrx="681" lry="681">
    <desc>Libovolný volitelný popis (červená oblast).</desc>
  </zone>
  <zone xml:id="ID0011R__pic_2" points="21,1044 822,1044 822,783 1059,783 1059,1536
21,1536">
    <desc> Libovolný volitelný popis (zelená oblast).</desc>
  </zone>
</surface>
```

Postup při generaci atributu xml:id pro element `<zone>` byl navržen tak, že

```
@xml:id = ../surface/@xml:id+"__"+pic+"_"+N
```

Kde `pic` je zkratka typu rozeznaného grafického elementu a `N` je pořadové číslo výskytu daného typu grafického elementu v daném obraze.

Aby bylo možné vztáhnout umístění grafického elementu relativně k danému obrazu i zrcadlu stránky, je vždy součástí záznamu o daném `<surface>` i informace o rozměrech obrazu a o zrcadle, například:

```
<zone xml:id="ID0001v__graphic" ulx="0" uly="0" lrx="995" lry="1120"/>
<zone xml:id="ID0001v__mirror" ulx="164" uly="32" lrx="849" lry="1031"/>
```

Tak je zajištěno, že mění-li se rozlišení obrazu (například použije-li se jiná uživatelská kvalita), budou informace zanesené v metadatech i nadále platné.

Poznámka: povšimněte si v příkladu uvedeného zapouzdření elementu `<graphic>` s odkazem na odpovídající datový soubor do elementu `<zone>` - tento přístup lze aplikovat i u ostatních `<zone>` elementů v případě existence doplňkových obrazových dat (například výřez iniciály ve vyšším rozlišení dokumentů). Bližší informace o aparátu k zachycení nově generovaných metadat uvádí [1].

4 Implementační fáze

4.1 Cíle implementace

Cílem řešení je obohatit uživatelské rozhraní Manuscriptoria o funkce založené na existenci nových metadat generovaných IIR Aplikací. Za tím účelem AIP vytvořila pilotní řešení, umožňující Zadavateli zhodnotit kvalitu generovaných informací a zvážit možnosti jejího nasazení do ostrého provozu MnS.

V součinnosti s odpovědnými pracovníky Zadavatele byly popsány tyto dvě nové funkce:

1. Vyhledávání dokumentů dle obsahu faksimile:
   vyhledávací formulář bude obohacen o checkboxy, které omezí hledání například jen na dokumenty s hudební notací, či dokumenty s iniciálami.

2. Rychlý přístup na relevantní stránky:
   z detailu záznamu bude možno otevřít faksimile na konkrétní stránce, například na stránce s iluminací.

Seznam rozeznávaných grafických elementů, které budou využity k dosažení výše uvedeného:

- Ilustrace
- Iniciála
- Notace
- Marginélie
- Tabulka

4.2 Příprava implementace

Součástí přípravných fází řešení byla formulace požadavků na IIR Aplikaci dodávané Zadavatelem, které se týkaly požadovaných vstupů a výstupů.

Cílem úkolu bylo umožnit efektivní hromadné zpracování obrazů. Proto bylo požadováno, aby IIR Aplikace mohla pracovat dvěma způsoby, které se liší v závislosti na použitém druhu vstupů:

1. Neexistují metadata, vstupem je adresářová struktura s obrazy:
   IIR Aplikace předepsaným způsobem generuje nová metadata ve formátu TEI P5 a ukládá jako samostatné soubory (jeden pro každý adresář s obrazy)

2. Existují metadata, vstupem je adresářová struktura s XML soubory:
   IIR Aplikace zpracovává obrazové soubory z umístění odkazovaného v metadatech a doplňuje do existujících metadat nové informace.

Kromě výstupu ve formátu TEI P5 XML bylo též požadováno, aby aplikace generovala log soubory s informací o míře jistoty, s jakou byly jednotlivé výskyty informací ve stránce rozeznány. Tento log soubor může být mimo jiné využit při analýze úspěšnosti.

4.3 Workflow zpracování metadat

IIR Aplikace byla začleněna do workflow pilotního prostředí MnS. Vstupem ke zpracování jsou existující XML soubory. Data ke zpracování musí být IIR Aplikaci

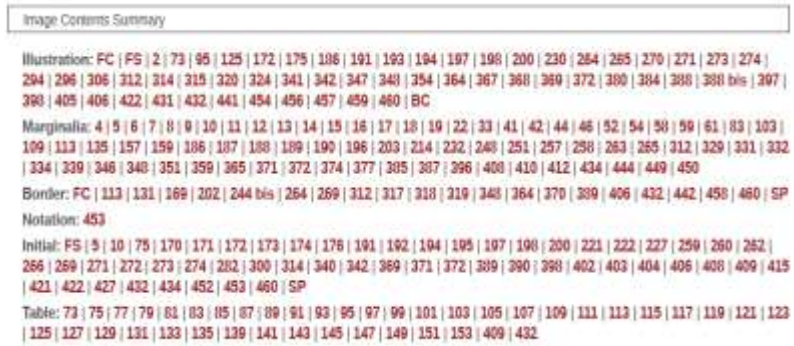
dostupná lokálně, čili je nutno centralizovat zpracovávané obrazové soubory i odpovídající metadata (pracovní kopie) a ta aktualizovat podle aktuálního místa uložení dat. Toto řešení obsahuje výrazný podíl ruční práce při přípravě a zpracování dat a metadat, což je při realizaci pilotního řešení akceptovatelné. Pro případné ostré nasazení bude nutné jednotlivé procesy automatizovat, aby bylo možno bezpečně provádět dávkové hromadné zpracování.

Začlenění výstupů IIR Aplikace do workflow zpracování metadat se samozřejmě neobešlo bez změny rutin pro generování FRT/FDM souborů – datových souborů katalogu MnS tak, aby bylo možno vyhledávat dle typu obrazu.

4.4 Implementace do koncového rozhraní

V souladu s cíllem pilotního řešení byly provedeny změny v pokročilém formuláři, jež umožňují žádoucím způsobem omezit výsledky vyhledávání.

Dále bylo upraveno zobrazení detailu záznamu tak, že při výpisu informací se generují seznamy stránek s odkazy na stránky obsahující dané typy informací.



4.5 Pilotní řešení

Pro pilotní řešení byla zvolena data dokumentů české provenience. Z části se jedná o náhodný výběr a z části o sadu zvláštních dokumentů různých typů a vlastností, jež AIP vybrala nad rámec smlouvy a které považuje za důležité pro testování.

Sadu dokumentů je možno změnit dle požadavků Zadavatele v závislosti na aktuálních potřebách a výsledcích testování.

V době kompletace této zprávy se jednalo o 88 dokumenů náhodně vybraných a 25 dokumentů cíleně vybraných [viz dokumentace].

Pilotní řešení je volně dostupné na adrese

http://www.manuscriptorium.com/apps/pilot/iir/

a aby bylo možno testovat jej v mezinárodním měřítku, je k dispozici i anglická verze

http://www.manuscriptorium.com/apps/pilot/iir/en

Některé pro testování nedůležité koncovo-uživatelské funkce byly v pilotním řešení zakázány.

5 Závěr

Pilotní řešení je nyní plně k dispozici na výše uvedené adrese. Pracuje s aktuálními výsledky poskytovanými IIR Aplikací.

Manuscriptorium aktuálně zpřístupňuje cca 5 000 000 obrazů. Měření rychlosti při zpracování vzorků dat ukazuje, že lze IIR Aplikaci využívat i v takto masovém měřítku. Pokud tedy Zadavatel dojde po testování k rozhodnutí, že je účelné IIR Aplikaci a jí generovaná metadata využívat v MnS, lze dále řešit její nasazení v ostrém provozu, mimo jiné:

- řešit nasazení do workflow ostrého MnS,
- vyřešit otázku aplikace nad externě uloženými daty (například zvážit možnost centralizace obrazů a spojení IIR s generováním standardizované Normal sady dat),
- zapojení technologie do aplikace M-Tool a usnadnit popis oblastí obrazu,
- dále rozšířit koncovo-uživatelské funkce při prohlížení faksimile atp.

Lze také uvažovat o tom, že kvalitu rozeznaných informací je možné dále vylepšovat v poloautomatickém provozu. S vhodně navrženým nástrojem pro hromadnou poloruční práci (například rychlou vizuální kontrolu a odmazávání nesprávně rozeznaných oblastí – bordury, marginálie). Rychlost práce by mohlo navýšit využití informací z logu IIR Aplikace.

6 Příloha 1: Seznam dokumentů cíleně vybraných do testovacího vzorku

Seznam je tvořen FyzId jednotlivých digitalizovaných exemplářů. Hledání v pilotu lze provést podle signatury (např. CO.X.13 pro KKPS__CO_X_13_____332I). Prvních 6 znaků FyzId je interní kód knihovny dle seznamu MnS.

KKPS__CO_X_13_____332I
KNMP__MS_F_1_____181I
MZM___A_7077_3_____1FAH
NKCR__XIII_A_6_____1IWS
NKCR__XXIII_C_124____1MZN
NMP___25_A_13_____2INJ
NMP___25_A_5_____18TJ
NMP___26_B_8_____2TGS
PNP___TR_I_27_____24NN
SK____BE_VIII_66_____0QEO
SK____DE_III_22_____1POI
SOAZ__FOND_VELKOSTATE1O1P
VKO___M_I_306_____05WR
VMO___K_14905_____01RS
ZK____20_H_3_14131___03WM
ZK____20_K_21_14261_H0YIH
ZK____27_C_18_1_1875516VH
ZK____7_D_8_5093_____16PJ
ZMP___502_F_13_____27MI
ZMP___510_A_011_____0R9N
ZMP___510_A_016_____2C4O
ZMP___510_B_003_____0QLJ
ZMP___510_B_024_____0OSH
ZMP___510_CH_1_____2SXL
ZMP___513_A_4_____0UBQ