ÚČAST NK ČR V PROJEKTU TELPLUS – VYTVÁŘENÍ OCR SOUBORŮ

Tomáš Foltýn, Národní knihovna ČR

Základní informace o projektu TELplus
Zakotvení mezi dalšími evropskými projekty

TELplus je jedním z mnoha evropských projektů, které si kladou za cíl rozšířit množství nabízených služeb a obsah Evropské knihovny ("The European Library").[1] Ta nabízí rychlý přístup ke sbírkám více než 45 národních knihoven Evropy. Její portál uživatelům jedinečným způsobem umožňuje nahlédnout nejen do bibliografického popisu jednotlivých dokumentů, ale zároveň i do digitální kopie, pokud jí daná instituce nechrání z důvodu autorských práv. V současné době je zde k dispozici více než 150 miliónů záznamů, přičemž tento počet stále vzrůstá, což její funkcionalitu udrží na vysoké úrovni i nadále. Ta už nebude spojena pouze s dalším budováním Evropské knihovny, nýbrž i s mezioborovými projekty typu portálu Europeana.[2]

V rámci naplňování zmíněných základních idejí Evropské knihovny bylo spuštěno poměrně velké množství projektů, které měly služby této knihovny zkvalitnit a zatraktivnit. Některé z nich již byly ukončeny (např. TEL-ME-MOR[3] nebo EDL[4]), jiné právě probíhají (např. FUMAGABA[5] nebo TELplus[6]) a další budou jistě brzy následovat, aby se co nejdříve podařilo naplnit ideál Evropské knihovny, který shrnuli její tvůrci následující větou: „*Smyslem Evropské knihovny je otevřít všem zájemcům ucelený sou-*

[1] Pro podrobnější informace navštivte webovou stránku "The European Library" http://search.theeuropeanlibrary.org/portal/en/index.html.

[2] Beta verze tohoto portálu, kde jsou k dispozici i detailní informace o projektu, se nachází na http://www.europeana.eu/portal/index.html.

[3] Detailnější informace na http://www.theeuropeanlibrary.org/portal/organisation/cooperation/archive/telmemor/index.php.

[4] Podrobné informace na http://www.theeuropeanlibrary.org/portal/organisation/cooperation/archive/edlproject/index.php.

[5] Projektová stránka http://www.theeuropeanlibrary.org/portal/organisation/cooperation/fumagaba/.

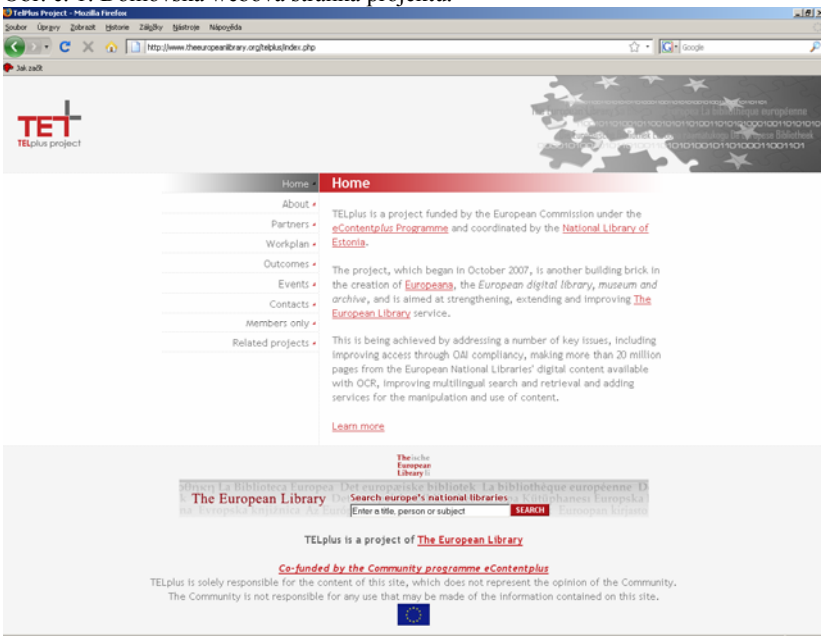
[6] Webová stránka projektu TELplus http://www.theeuropeanlibrary.org/telplus/index.php.

bor znalostí, informací a kulturních artefaktů, které nabízejí evropské národní knihovny.“[7]

Projekt TELplus

TELplus je projekt financovaný Evropskou komisí v rámci podprogramu eContentplus. Hlavními koordinátory jsou Eremo s.r.l. a Národní knihovna Estonska. Projekt TELplus se rozeběhl v říjnu roku 2007 a skončí v prosinci roku 2009. Celkový rozpočet projektu je 6 501 714 EUR, z čehož polovinu tvoří dotace z prostředků Evropské komise a druhou polovinu spoluúčast jednotlivých partnerů.

Obr. č. 1: Domovská webová stránka projektu.[8]



Projekt TELplus je tematicky velice rozvrstvený. Mezi jeho hlavní cíle patří vytvoření více než 20 milionů OCR textových souborů napříč evropskými národními knihovnami, což by mělo uživatelům usnadnit vyhledávání

[7] V originále: “*The European Library exists to open up the universe of knowledge, information and cultures of all Europe’s national libraries.”* Srov. http://search.theeuropeanlibrary.org/portal/en/index.html.

[8] Viz http://www.theeuropeanlibrary.org/telplus/index.php.

v digitálních dokumentech. Další klíčovou náplní projektu TELplus je zlepšení propojení jednotlivých digitálních knihoven pomocí protokolu OAI-PMH, což má usnadnit sklízení dat určených pro Evropskou knihovnu. Třetím hlavním cílem je vylepšit fulltextové vyhledávání a další služby pro manipulaci a používání digitálního obsahu. Projekt TELplus dále řeší i dílčí úkoly, mezi něž patří například budování skupin uživatelů, které by měly dále zkvalitnit služby portálů spojených s evropskými projekty, nebo připojení národních knihoven Bulharska a Rumunska mezi plnohodnotné členy Evropské knihovny. Průběžné výsledky plnění těchto snah jsou k dispozici na projektovém webu.

Z uvedených důvodů je řešitelský tým velice různorodý. Největší část tvoří jednotlivé národní knihovny, zastoupeny jsou však i univerzitní knihovny nebo vývojová a výzkumná centra. Aby byla zaručena maximální efektivita při řešení mnoha různých aktivit, byli jednotliví partneři rozděleni do osmi tématických pracovních skupin, které řeší jednotlivé úkoly. Někteří z níže uvedených partnerů se účastní na aktivitách většího množství pracovních skupin.

Tabulka č. 1: Výčet řešitelů projektu

Řešitelé projektu
Estonská národní knihovna
Rakouská národní knihovna
Rakouské výzkumné centrum GmbH
Italská národní knihovna Florencie
Eremo s.r.l.
Francouzská národní knihovna
Německá národní knihovna
Institut Superior Técnico, Portugalsko
Univerzita Kapodistrian Atény
Bulharská národní knihovna
Národní knihovna České Republiky
Maďarská národní knihovna
Národní a univerzitní knihovna Island
Lotyšská národní knihovna
Litevská národní knihovna
Národní knihovna Nizozemí
Norská národní knihovna

Polská národní knihovna
Portugalská národní knihovna
Národní knihovna Rumunska
Národní a univerzitní knihovna Lublaň, Slovinsko
Národní knihovna Španělska
Národní knihovna Švédska
Slovenská národní knihovna
Univerzita Padova
Univerzita Vrije, Slovinsko

„Workpackage 1" a účast Národní knihovny České republiky na řešení cílů projektu TELplus

Úkoly první pracovní skupiny

První pracovní skupina se podle svého názvu „OCR dříve digitalizovaných materiálů" zabývá vytvářením textových souborů vlastních digitalizovaných dokumentů. Partneři sdružení do této skupiny se v průběhu projektu zavázali vytvořit více než 20 milionů OCR textových stran a díky nim obohatit fulltextové vyhledávání Evropské knihovny. Na plnění tohoto úkolu se podílí 14 evropských národních knihoven, z nichž každá se do projektu zapojí mimo vlastních textových souborů i svými znalosti a zkušenostmi s tvorbou tohoto konkrétního typu dat a digitalizace vůbec. Zajímavé a důležité informace si knihovny průběžně vyměňují během projektových jednání a workshopů. Všechny zpracované sbírky budou následně sklizeny pomocí nástrojů využívajících otevřený protokol OAI-PMH, které jsou vyvíjeny v dalších pracovních skupinách. V rámci plnění projektu by měly být formulovány i tipy pro nejlepší zpracování OCR, které by bylo možné využít i po skončení projektu.

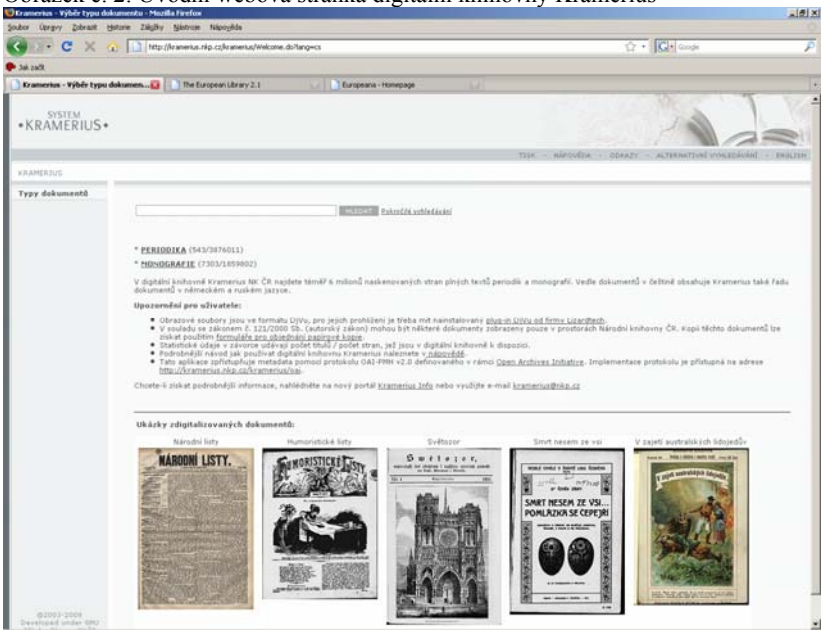
Účast Národní knihovny České republiky

Národní knihovna České Republiky se zavázala, že do projektu TELplus přispěje vytvořením 3 400 000 textových souborů. Takto vysokým počtem se Národní knihovna ČR zařadila za Francouzskou národní knihovnu a Národní knihovnu Španělska na třetí místo v počtu dodaných souborů. Celkový počet textových souborů je rozdělen na dvě části:

2 400 000 stran monografií a zhruba 1 000 000 stran periodik. Kritériem pro výběr uvedených dokumentů byla návaznost na další projekt, který Národní knihovna ČR řeší – tzv. „Norské fondy". Jejich cílem je zachování bohemikálních monografií, které jsou ohroženy degradací papíru. Pomocí projektu TELplus pak budou k takto zdigitalizovaným dokumentům vytvořeny OCR soubory. Jednotlivé tituly periodik pak byly zvoleny v součinnosti

s programem Ministerstva kultury ČR Veřejné informační služby knihoven, na kterém se Národní knihovna ČR též podílí. Zvolené dokumenty pocházejí z devatenáctého a počátku dvacátého století a jsou převážně napsány v českém a německém jazyce. Jejich digitální kopie jsou uživatelům k dispozici v souladu s autorským právem prostřednictvím digitální knihovny Kramerius.[9] Rozpočet Národní knihovny na tvorbu OCR byl po několika úpravách stanoven na 177 000 EUR určených na tvorbu OCR (z toho polovinu tvoří dotace Evropské komise) a 14 000 EUR na cestovné (polovinu opět tvoří dotace).

Obrázek č. 2: Úvodní webová stránka digitální knihovny Kramerius



Produkce textových souborů stojí v řadě operací digitalizace originálních dokumentů téměř až na konci procesu, neboť následuje až po fázích výběru dokumentů, mikrofilmování, vlastní digitalizace a ořezů. Téměř veškeré OCR probíhají jako většina operací v rámci digitalizace dat ve spolupráci s dodavateli, kteří textové soubory získávají pomocí softwaru FineReader od společnosti Abbyy. Stejný software využívá i většina dalších partnerů.[10]

[9] Viz http://kramerius.nkp.cz/kramerius/Welcome.do.

[10] Rozšíření funkcí ABBYY FineReaderu a vylepšení výsledků rozeznávání textů je cílem dalšího z mezinárodních evropských projektů Impactu. Podrobnější informace na projektové stránce http://www.impact-project.eu.

Pro testování výsledků OCR se zatím nepoužívá žádný automatický nástroj, ale pouze namátková manuální kontrola. Brzy by se však měly objevit i nekomerční systémy, které kontrolu kvality usnadní. Alternativní cestou může být i spolupráce s uživateli.[11]

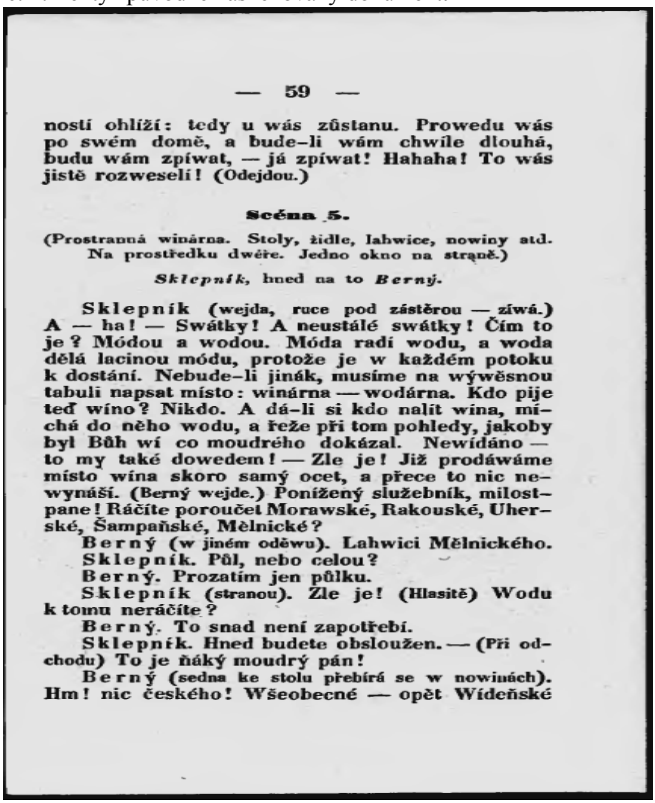
Obrázek č. 3: Textový soubor knihy „Mašinkář“ Jana Nepomuka Nürnbergera.[12]

```
— 59 —
noslí ohlíží: tedy u vás zůstanu. Prowedu wás
po swém domě, a bude-li wám chwíle dlouhá,
budu wám zpíwat, —já zpíwat.' Hahaha! To wás
jistě rozweselí! (Odejdou.)
Scéna .5.
(Prostranná winárna. Stoly, židle, lahwice, nowiny
ald.
Na prostředku dwéře. Jedno okno na straně.)
Sklepník, hned na to Berný.
Sklepník (wejda, ruce pod zástěrou — zíwá.)
A — ha! — Swátky! A neustálé swátky! čím to
je ? Módou a wodou. Móda radí wodu, a woda
dělá lacinou módu, protože je w každém potoku
k dostání. Nebude-li jinak, musíme na wýwěsnou
tabuli napsat místo: winárna — wodárna. Kdo pije
teď wíno? Nikdo. A dá-li si kdo nalít wína, mí-
chá do něho wodu, a řeže při tom pohledy, jakoby
byl Bůh wí co moudrého dokázal. Newídáno —
to my také dowedem ! — Zle je! Již prodáwáme
místo wína skoro samý ocet, a přece to nic ne-
wynáší. (Berný wejde.) Ponížený služebník, milost-
pane! Ráčíte poroučet Morawské, Rakouské, Uher-
ské, Šampaňské, Mělnické?
Berný (w jiném oděvní). Lahwici Mělnického.
Sklepník. Půl, nebo celou?
Berný. Prozatím jen půlku.
Sklepník (stranou). Zle je! (Hlasitě) Wodu
k tomu neráčíte?
Berný. To snad není zapotřebí.
Sklepník. Hned budete obsloužen. — (Při od-
chodu) To je ňáký moudrý pán!
Berný (šedna ke stolu přebírá se w novinách).
Hm! nic českého! Wšeobecné — opět Wídeňské
```

[11] Zajímavý projekt, který využívá uživatelských oprav, probíhá v Austrálii. Srov. http://ndpbeta.nla.gov.au/ndp/del/home.
[12] Srov. http://kramerius.nkp.cz/kramerius/document/ABA001_2301800059.txt.

Kvalita textových souborů je na vysoké úrovni zejména u českých dokumentů, jenž jsou psány latinkou. Naopak problémové jsou zejména švabachové německé texty či případně kombinace více jazyků či fontů v jednom dokumentu. Jednou z cest pro vylepšení výsledků OCR takto psaných materiálů je vytváření znalostních bází termínů psaných těmito typy písma.

Obrázek č. 4: Tentýž původně naskenovaný dokument.[13]



— 59 —

nosti ohlíží: tedy u wás zůstanu. Prowedu wás po swém domě, a bude-li wám chwíle dlouhá, budu wám zpíwat, — já zpíwat! Hahaha! To wás jistě rozweselí! (Odejdou.)

Scéna 5.

(Prostranná winárna. Stoly, židle, lahwice, nowiny atd. Na prostředku dwéře. Jedno okno na stranč.)

Sklepník, hned na to *Berný*.

Sklepník (wejda, ruce pod zástěrou — zíwá.) A — ha! — Swátky! A neustálé swátky! Čím to je? Módou a wodou. Móda radí wodu, a woda dělá lacinou módu, protože je w každém potoku k dostání. Nebude-li jinák, musíme na wýwěsnou tabuli napsat místo: winárna — wodárna. Kdo pije teď wíno? Nikdo. A dá-li si kdo nalít wína, míchá do něho wodu, a řeže při tom pohledy, jakoby byl Bůh wí co moudrého dokázal. Newídáno — to my také dowedem! — Zle je! Již prodáwáme místo wína skoro samý ocet, a přece to nic newynáší. (Berný wejde.) Ponížený služebník, milostpane! Ráčíte poroučet Morawské, Rakouské, Uherské, Šampaňské, Mělnické?

Berný (w jiném oděwu). Lahwici Mělnického.

Sklepník. Půl, nebo celou?

Berný. Prozatím jen půlku.

Sklepník (stranou). Zle je! (Hlasitě) Wodu k tomu neráčíte?

Berný. To snad není zapotřebí.

Sklepník. Hned budete obsloužen. — (Při odchodu) To je ňáký moudrý pán!

Berný (sedna ke stolu přebírá se w nowinách). Hm! nic českého! Wšeobecné — opět Wídeňské

V současné době nejsou textové soubory pro běžné uživatele přímo k dispozici. Využívají se pro fulltextové vyhledávání dokumentů, které probíhá pomocí vyhledávače Lucene. Digitální knihovna Kramerius svým uživatelům nabízí dvě úrovně vyhledávání – základní fulltextové a pokročilé.

[13] Srov. http://kramerius.nkp.cz/kramerius/MShowPageDoc.do?id=469920&mcp=&idpi=12541049&author=.

Výstup z produkce OCR je k dispozici v běžném textovém souboru a formátu METS a je ho možné sklízet pomocí protokolu OAI-PMH a pomocí http serverů. Popisná metadata každého dokumentu jsou rovněž dostupná s pomocí metadat formátu METS, podle popisu MARC21 XML, tak i dle jednoduchého Dublin Core.

Závěr

Projekt TELplus je pouze jedním z mnoha projektů, kterých se Národní knihovna ČR účastní jako plnohodnotný partner. Tyto projekty knihovnu obohacují nejen po stránce tvorby dat a jejich sdílení s nadnárodními portály, ale zároveň jí přináší nové poznatky jak po stránce teoretické, tak i praktické. Zároveň ukazují Národní knihovnu jako dobře fungující instituci, za kterou mluví výsledky její práce.