



europæana
newspapers



Projekt Europeana Newspapers

Evropská online brána k digitálním novinovým sbírkám

Aleš Pekárek, Asociace evropských vědeckých
knihoven (LIBER)

13. Konference „Archívy, knihovny a muzea v
digitálním světě, Praha, 28.11.2012

Obsah

- Cíle a záměry projektu
- Partnerské instituce
- Oblasti projektových aktivit
- Komunikace a diseminace projektu
- Přidružení partneři projektu

Proč zrovna noviny?

„Noviny jsou bazarem dějin“ (Die Zeitungen sind der Sekundenzeiger der Geschichte)

Arthur Schopenhauer

„Noviny jsou jedním z nejpopulárnějších materiálů v naší knihovně, protože přímo přetékaají informacemi a mají co říci širokému spektru našich návštěvníků. V rámci projektu dodáme do Europeany 2 miliony našich plně digitalizovaných stránek, čímž je zpřístupníme mnohem většímu okruhu uživatelů a v mnohem širším kontextu než doposud.“

Lily Knibbeler,

Ředitelka Marketingu a Služeb Královské (Národní) knihovny Nizozemí

V novinách si zkrátka každý něco najde!

Cíle & Záměry

1) Výběr, Rafinace („refinement“) a Agregace obsahu

- Učinit z Europeany největšího poskytovatele pan-evropské digitální sbírky novin
- Dodání více než **18 million novinových stránek** do Europeany, mnoho z nich společně s vyextrahovaným fulltextem (OCR, OLR, NER)

2) Analýza existujících digitálních novinových sbírek

- Celoevropský průzkum

3) Kvalita & Doporučení nejlepší praxe

- Vytvoření optimalizovaného workflow a infrastruktur pro agregaci dat
- Doporučení nejlepších praxí pro pročištění, workflow, metadata a evaluačních nástrojů v oblasti digitalizace novinových sbírek

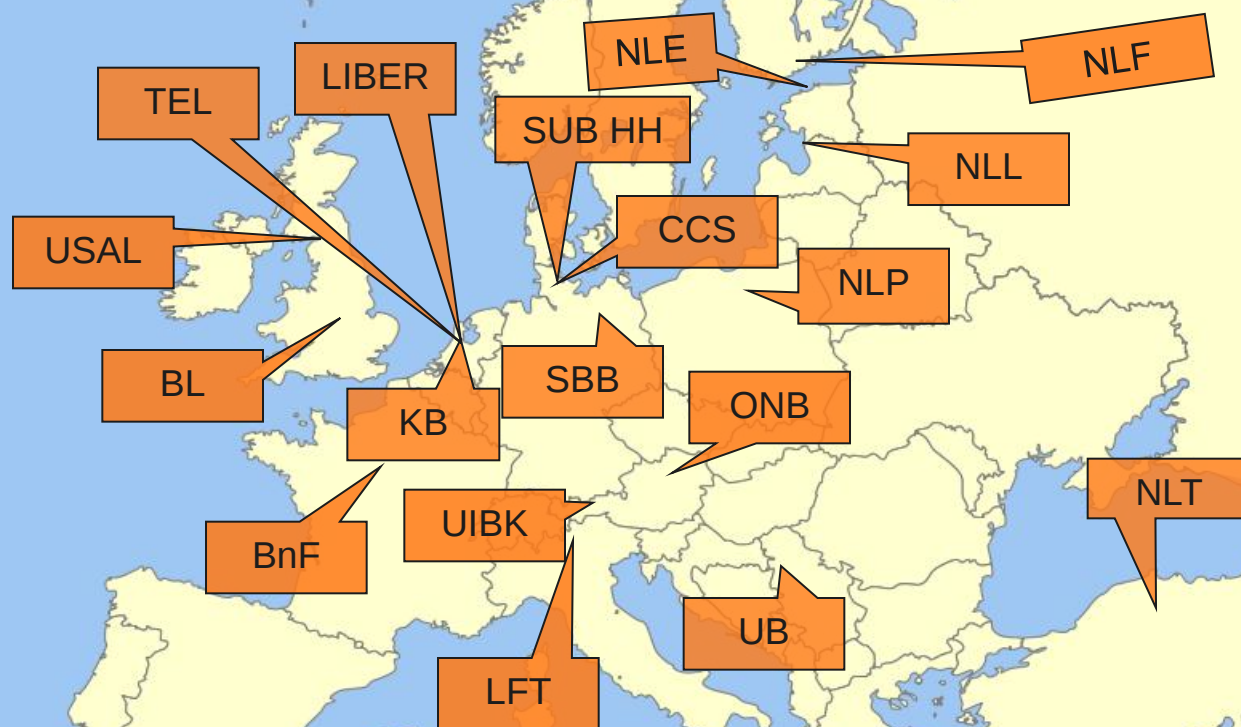
4) Prezence obsahu a fulltextové vyhledávání

- Vylepšený přístup k novinovým sbírkám v rámci Europeany (Content Browser)

Projektové konsorcium & Stakeholderi

- 17 partnerů z 12 zemí
 - Národní knihovny – obsahoví a techničtí partneři
 - Univerzitní knihovny – převážně techničtí partneři
 - Malý a střední podnik (CCS – Content Conversion Specialists)
- Externí partneři a stakeholderi:
 - Zapojení knihoven vně projektového konsorcia
- Celkový rámec:
 - Financováno jako Best Practice Network v programu ICT-PSP Evropské komise
 - Trvání projektu: Únor 2012 – Leden 2015

Partneři projektu Europeana Newspapers



Partneři projektu

1. Staatsbibliothek zu Berlin
(koordinátor projektu)
2. National Library of the
Netherlands
3. National Library of Estonia
4. Österreichische
Nationalbibliothek
5. National Library of Finland
6. Staats- und
Universitätsbibliothek Hamburg
7. Bibliothèque nationale de
France
8. National Library of Poland
9. University of Salford
10. CCS Content Conversion
Specialists GmbH
11. Stichting LIBER
12. National Library of Latvia
13. National Library of Turkey
14. University Library of Belgrade
15. University of Innsbruck
16. Landesbibliothek Dr. Friedrich
Tessmann
17. The British Library

Struktura projektových aktivit (+ koordinátoři)

- **Work Package 1: Koordinace a řízení projektu**
 - Berlin State Library (SBB)
- **Work Package 2: Výběr a rafinace digitalizovaných novin**
 - National Library of the Netherlands (KB)
- **Work Package 3: Evaluace a hodnocení kvality výběru**
 - University of Salford (USAL)
- **Work Package 4: Agregace a příprava novin pro Europeanu**
 - The European Library (TEL)
- **Work Package 5: Doporučení „Best Practice“ v oblasti novinových metadat**
 - University of Innsbruck (UIBK)
- **Work Package 6: Diseminace a zužitkování projektu**
 - Association of European Research Libraries (LIBER)

WP 1: Koordinace a řízení projektu

- **Administrace projektu**
 - Řízení všech finančních a organizačních záležitostí
- **Finanční kontrola**
- **Projektová komunikace**
 - Sharepoint
- **Zajištění kvality projektu**
 - monitoring, evaluace a reporting výsledků
- **Risk management**
 - Předcházení konfliktů v rámci konsorcia 😊

WP 2: Rafinace digitalizovaných novin

- Analýza obsahu a výběr vhodných souborů z digitálních novinových sbírek k “zušlechtění” (rafinaci)
- Definice požadavků na minimální kvalitu digitalizovaných novin
- Koordinace zušlechtování vybraného novinového obsahu – novin poskytnutých partnerskými institucemi
- Doporučení nejlepších praxí v oblasti rafinace digitálních novinových sbírek

WP2: Rafinace – OCR and OLR

- 8 milionů stránek bude do Europeany naagregováno „tak jak jsou“

=====

- 10 milionů zrafinovaných stránek:
OCR (UIBK, Rakousko)

Z toho:

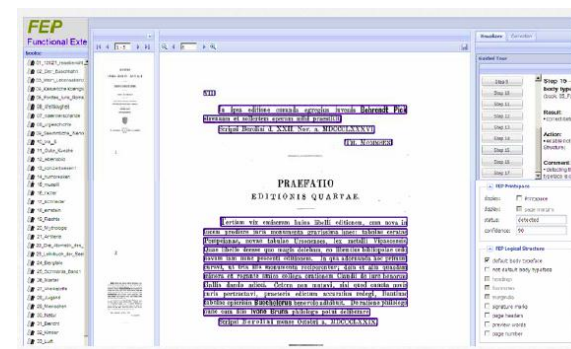
- 2 miliony:
OCR/OLR (segmentace článků)
(CCS, Německo)

UIBK obohatí vyextrahovaný obsah strukturálními informacemi z její „Platformy porozumění dokumentům“,

CCS pak informacemi o sloupcích, článcích; zoning; segmentace článků a jejich úrovní

CCS zároveň poskytuje knihovnám technologii pro manuální korekci výsledných strukturálních informací

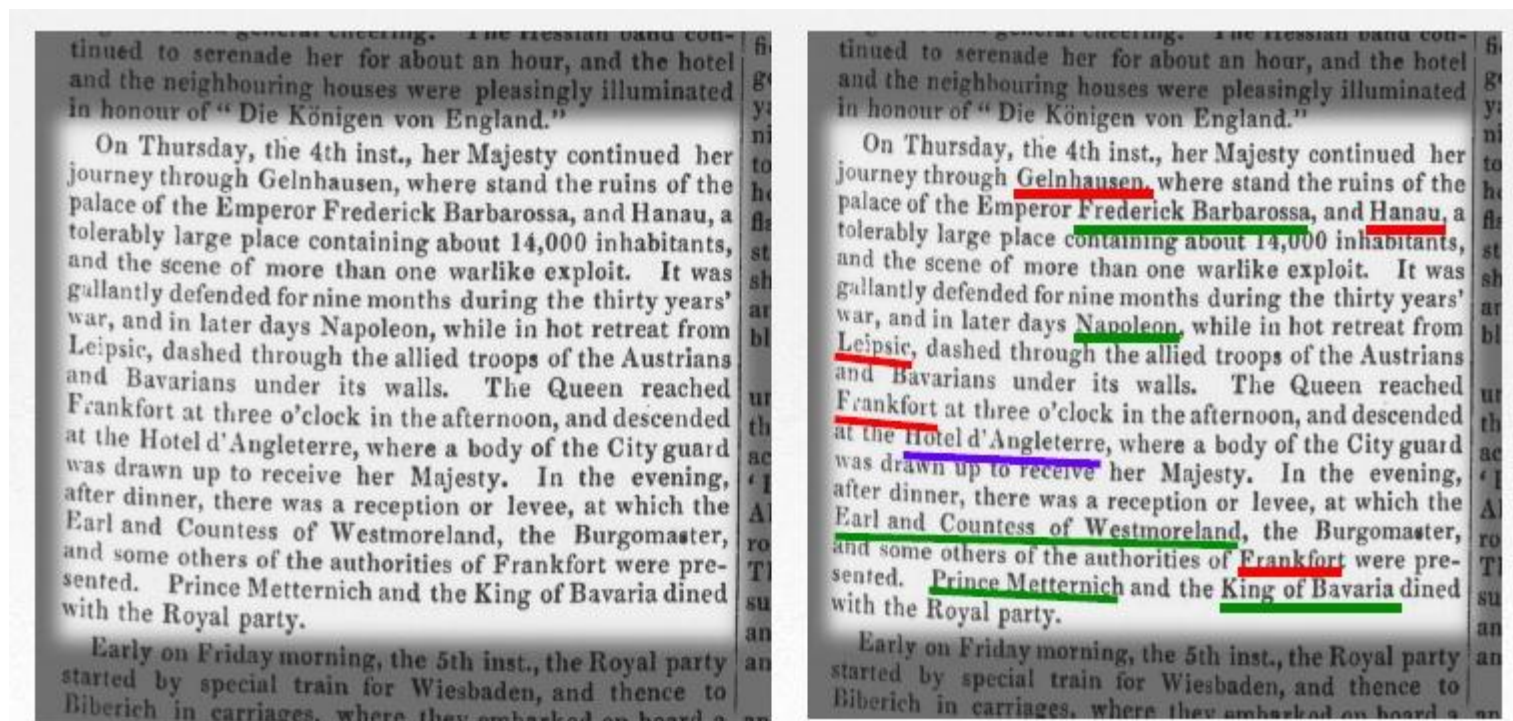
CCS: Automatické rozpoznávání a segmentace novinových sloupců a článků



UIBK: Detekce titulků a poznámek;
Automatické generování obsahu

WP 2: Rafinace – Rozpoznávání jmenných entit

- KB dodala systém rozpoznávání jmenných entit (NER) (pro holandštinu, angličtinu a němčinu)



WP 3: Evaluace a hodnocení kvality

- Vytvoření uživatelských scénářů se vzorovými profily, datovými sadami, „ground truth“ a nástroji hodnocení kvality výstupů
- Přehled faktorů využitelnosti, limitů a reálného potenciálu existujících vstupních materiálů
- Identifikace kritických míst procesu společně s návrhem řešení
- Evaluace jednotlivých aktivit procesu
- Doporučení nejlepších praxí pro podobné post-digitalizační projekty

WP 4: Agregace a příprava pro Europeanu

- Identifikace a analýza veřejných i soukromých zdigitalizovaných novinových sbírek napříč Evropou
- Vytvoření reálného časového rámce agregace novinového obsahu
- Vytvoření celoevropského registru digitalizovaných novinových sbírek
- Doporučení, jak napasovat existující novinová metadata na EDM (standard Europeany)
- Agregace novinových metadat partnerů projektu
- Vytvoření fulltextového indexu novinového obsahu
- Vývoj a spuštění prohlížeče obsahu (full-text + metadata + originální skeny či náhledy)

WP 4: Agregace obsahu

- Agregace 18 milionů digitalizovaných novinových stránek do služby *Europeana* prostřednictvím *Evropské knihovny* (jednotný portál evropských národních knihoven)
- Transformace vstupních metadat na bázi EDM
- Distribuce dat do Europeany (po jejich finální transformaci v rámci TEL)



www.europeana.eu



www.theeuropeanlibrary.org

WP 4: Prezentace & Přístup k plným textům

Funkce obsahového browseru:

- **Fulltextové vyhledávání**, např.
 - podle klíčového slova,
 - podle jmenných entit
 - podle novinových sbírek
 - podle data vydání novin...
- **Prohlížení naskenovaných novinových stránek**
- **Propojení s příbuzným obsahem nenovinového typu (fotky, hudba, pohlednice...)**

Explore Europe's cultural collections

Lewis and Clarke

CAPTAIN LEWIS'S EXPEDITION.

THE following particulars of the expedition of captain Lewis, from the mouth of the Missouri, which empties into the Mississippi at St. Louis, to the Pacific ocean, transmitted to general Clarke, of Kentucky, by his brother, who accompanied the expedition, will be found interesting.

St. Louis, 23d Sept. 1806.

DEAR BROTHER,

We arrived at this place at 12 o'clock to-day, from the Pacific ocean, where we remained during the last winter, near the entrance of the Columbia river. This station we left on the 27th of March last, and should have reached St. Louis early in August, had we not been

WP 5: Metadata best practice recommendations

- Analysis of metadata formats in use by libraries
- Align metadata models with the METS/ALTO standard and release best practise recommendations
- Usability of the recommendation will be tested through an evaluation cycle
- Provide recommendations on best practices for refinement of digitized newspaper collections for Europeana

WP 6: Komunikace a diseminace projektu

- Cíle
 - Vytvoření celoevropské publicity
 - Větší používanost Europeany
 - Zajištění informovanosti cílových skupin
- Úkoly
 1. Komunikace v mediích
 2. Workshopy a účast na konferencích
 - Tři hlavní diseminační workshopy (rafinace, agregace, závěrečný)
 - Národní informační dny (minimálně 10)
 - Rozšíření sítě o další partnery
 3. Zužitkování výsledků projektu



Hlavní komunikační kanály projektu

WWW: <http://www.europeana-newspapers.eu/>

Twitter: @eurnews

Facebook: <https://www.facebook.com/EuropeanaNewspapers>

Newsletter: <http://bit.ly/TsoMpY> (najdete na webové stránce projektu)

Slideshare: http://www.slideshare.net/Europeana_Newspapers



Přidružení partneři projektu

- Rozšíření sítě o tzv. Přidružené partnery projektu (nefinancované, nepřímé benefity z partnerství)
 - Možnost poskytnutí metadat digitálních novinových sbírek do Europeany v rámci procesů projektu → rapidní nárůst návštěvníků
 - Využití a přístup k výsledkům projektu (doporučení, dokumentace, SW nástroje, metodika workflow)
 - Hlavně pro držitele obsahu
 - Většina identifikována v rámci průzkumu (40 vyplněných dotazníků, 35 potenciálních přidružených partnerů – o metodě výběru se stále diskutuje)
 - Za ČR zájem projevila NK ČR a MZK (i ostatní instituce s digitálními novinovými sbírkami jsou vítány)

Děkuji za pozornost

pekarek.ales@gmail.com

www.europeana-newspapers.eu

www.libereurope.eu

