Petr Kadlec <petr.kadlec@gmail.com>

Searching Session NTK 2010

NTK, 5. 10. 2010

(META)DATA VE WIKIPEDII

A JAK JE DOSTAT DOVNITŘ A VEN

Obsah

- Projekty Wikimedia Foundation
- (Meta)data ve Wikipedii
- Jak je dostat z Wikipedie
- Konkrétní projekty a nástroje
- Jak dostat (meta)data *do* Wikipedie

Projekty nadace Wikimedia

- Wikipedie – encyklopedie (2001)
- Wikislovník – slovník (2002)
- Wikicitáty – sbírka citátů (2003)
- Wikiknihy – manuály apod. (2003)
- Wikizdroje – původní texty (2003)
- Wikimedia Commons – soubory (2004)
- Wikizprávy – zpravodajství (2004)
- Wikiverzita – výukové materiály (2006)

MediaWiki

- Všechny projekty běží na MediaWiki
- „wiki-engine“ v PHP, MySQL
- Všechno open-source
  - http://www.mediawiki.org/wiki/MediaWiki/cs
- „Wikisyntaxe“ –formátovací jazyk
  - '''Tučné''', [http://example.cz Link], [[Odkaz]]
- Technika vespod ovlivňuje, co a jak snadno se dá z Wikipedie dostat

Autorskoprávní vložka

- Chcete využívat data z Wikipedie?
- Můžete!
  - Veškerý textový obsah pod CC-BY-SA 3.0
- Jen uvádějte odkaz na původní článek na Wikipedii a zachovejte licenci.
- Obrázky můžete šířit taky, konkrétní svobodná licence uvedena na stránce obrázku.

(Meta)data ve Wikipedii

- Jak všichni víme, ve Wikipedii je spousta potenciálně zajímavých věcí

Dejvice

Souřadnice: 50°06'19" s. š., 14°22'51" v. d.

Dejvice jsou městská čtvrť a katastrální území v pražské městské části Praha 6, rozkládající se severně od Pražského hradu.

Obsah [skrýt]

Historie

[editovat]

Historie moderních Dejvic jakožto městské čtvrti začíná ve 20. letech minulého století. V této době byla celá čtvrť i s jejím centrem, Vítězným náměstím, postavena najednou podle moderního urbanistického plánu architekta Antonína Engela. Zavedena byla tramvajová doprava, těsně před válkou i trolejbusová. Další výstavba se z Vítězného náměstí postupně posouvala na sever. V těchto místech byl také po válce postaven hotel International reprezentující socialistický realismus.

V roce 1978 sem bylo zavedeno metro (stanice Leninova, dnes Dejvická), vznikla také dnešní moderní silnice Evropská (dříve také Leninova) spojující letiště s centrem Prahy, a to přestavbou několika původních ulic (Kladenská).

Charakter čtvrti

[editovat]

Dejvice jsou relativně luxusní rezidenční čtvrtí, jíž sice dominují velkolepé budovy vysokých škol a armády, ale obsahuje i mnoho klidných míst s vilami movitých měšťanů a rezidencemi zastupitelských úřadů. Ceněno je těsné sousedství jak s centrální oblasti Prahy, tak s velkými plochami zeleně (Stromovka a Divoká Šárka).

Dejvice

část obce a katastrální území hl. města Prahy

Dejvická ulice

kód katastrálního území:	729272
připojení k Praze:	1922
městská část:	Praha 6
správní obvod (pověřený úřad):	Praha 6
městský obvod:	Praha 6
počet územně tech. jednotek:	1
základní sídelní jednotky:	14
katastrální výměra:	7,39 km²
obyvatel:	23 721 (16. 10. 2006)
hustota zalidnění:	3 210 obyv./km²

(Meta)data ve Wikipedii

- Encyklopedický text
- „Infoboxy“ – přehledové tabulky se základními údaji
- Zeměpisné souřadnice
- Bibliografické citace
- Odkazy na cizojazyčné ekvivalenty
- Atd. atd.

Jak je dostat z Wikipedie

- Některá lépe, některá hůře.
- Základ wiki tvoří nestrukturovaný chaos.
- Možnosti však jsou…

- Dvě otázky:
  - Přístup k datům
  - Formát dat

Přístup k datům

- Toolserver
  - Replikovaná SQL databáze
- XML dumpy
  - Stažitelný mirror Wikipedie
- SQL dumpy
  - Některé zajímavé tabulky
- MW API
- … a HTML screenscraping

Toolserver

- Serverová farma provozovaná WM DE
  - Přístup víceméně komukoli na žádost
- MySQL s replikovanými daty WM
- Tudíž přímý SQL přístup
- Ideální pro agregační dotazy, statistiky, při potřebě co nejaktuálnějších dat
- `https://wiki.toolserver.org/view/Hlavní_strana`

XML dumpy

- Projekty WMF jsou pod svobodnou licencí, právo na fork ⇒ data k dispozici
- Úplné XML dumpy obsahu lze stáhnout
- XML soubory obsahující texty článků
- Stále ve wikisyntaxi, bez vazeb atd.
- Celý dump včetně historie je *gigantický*
  - cswiki ~0,5 GB@7z, dewiki ~7,7 GB@7z
  - (bez obrázků)

XML dumpy (2)

- Ideální, pokud chcete spustit mirror
- Analýza je nad tím trochu složitější
  - Jediným parserem wikitextu je MediaWiki
- Ale je to *kompletní* Wikipedie včetně celé historie
  - Kromě smazaných článků a soukromých dat
- Ideální pro zkoumání vývoje v čase
- `http://download.wikimedia.org`

SQL dumpy

- Doplněk k XML dumpům
- V podstatě historický pozůstatek
- Zjednodušuje některou práci
- Metadata z databázových tabulek
  - Odkazy, kategorizace atp.
- Struktura vázána na MediaWiki
- `http://download.wikimedia.org`

MediaWiki API

- REST (HTTP)
- Různé formáty (JSON, XML, …)
- Dotazy na metadata
- Zajímavá schopnost: render
- Výkon? Přetěžování serverů WMF?
- `http://cs.wikipedia.org/w/api.php`

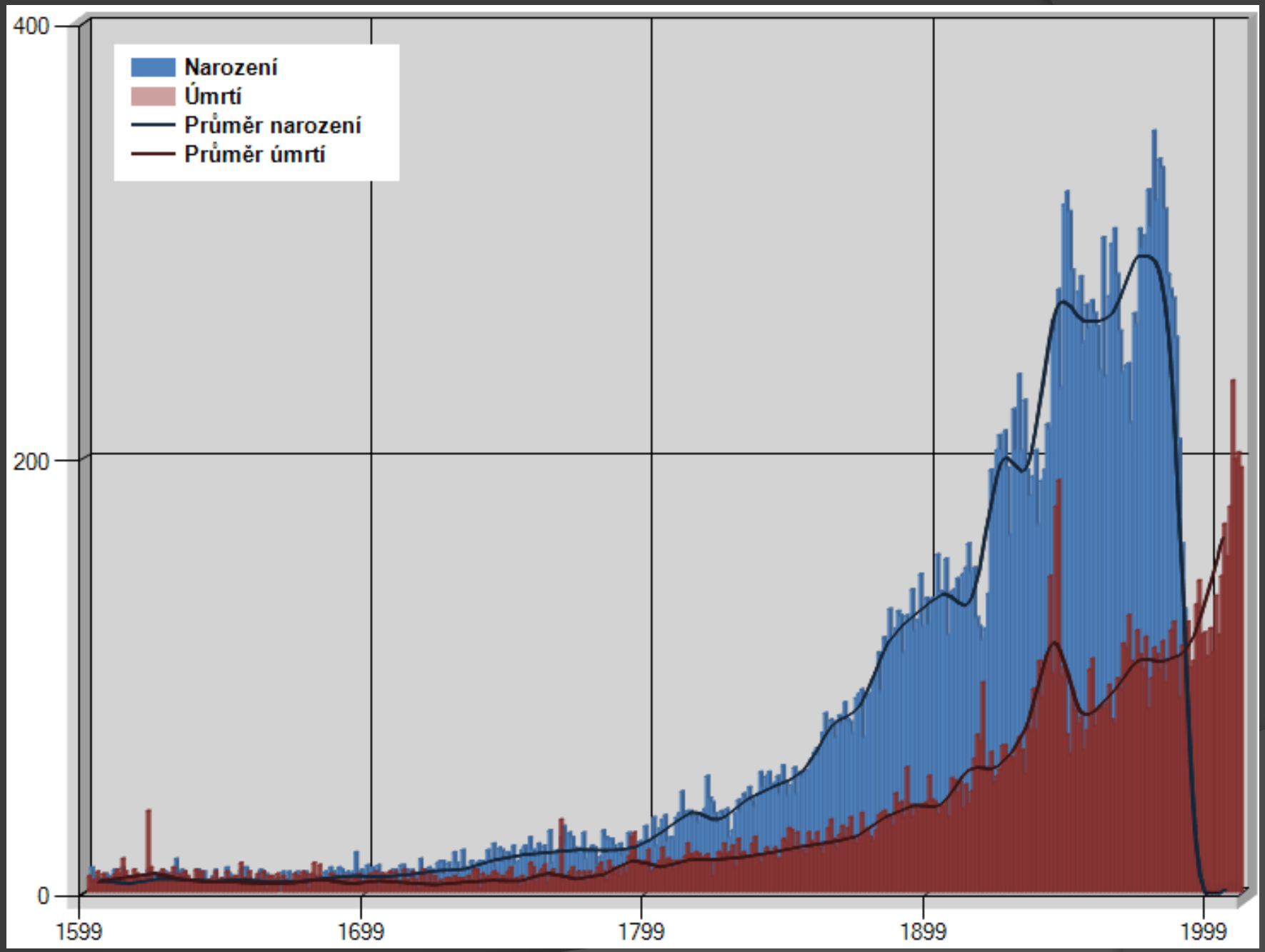
HTML screenscraping

- Fůůůj…
- Nespolehlivé, nikdo nezaručí stabilitu
- Pro spoustu věcí jediná možnost

- Alespoň mikroformáty?
  - …ještě se k tomu vrátíme

Konkrétní příklady

- Základní biografická data
  - Řekněme… data narození a úmrtí


400
200
0
Narození
Úmrtí
Průměr narození
Průměr úmrtí
1599
1699
1799
1899
1999

Data narození a úmrtí

- PHP skript na toolserveru
- Data zjišťuje pomocí kategorií
  - Kategorie:Narození 1234
  - Kategorie:Úmrtí 1234

- Co kdybychom chtěli dny v roce?
  - …bylo by to těžší.

Externí odkazy

- Tabulka externích odkazů
  - Dostupná na toolserveru
  - Dostupná přes API
  - Dostupná v SQL dumpech
- `http://wpcz.org/Special:Linksearch/*.techlib.cz`
- `http://cs.wikipedia.org/w/api.php?action=query&list=exturlusage&euquery=*.techlib.cz`

Mezijazykové odkazy

- Použití jako slovník?
  - Základ slovníku? → Wikislovník!
- Opět: tabulka v MediaWiki
  - Dostupné přes API, na toolserveru, z dumpů
- `http://cs.wikipedia.org/w/api.php?action=query&prop=langlinks&titles=Pes&lllimit=200&redirects`

A co infoboxy?

- Nejzajímavější, nejtěžší.
- Momentálně dvě možnosti:
  - Parsovat wikitext.
  - Parsovat HTML.
- Wikitext
  - Sice šablony, ale syntaxe stejně neexistuje
- HTML
  - Nestrukturovaný binec

Je potřeba cílené úsilí

- Díky šablonám občas jde nějaká užitečná data připravit ve strojově čitelném formátu
- Buď nějak využít datového modelu MediaWiki
  - Kategorie, externí odkazy, …
- Nebo do výstupního HTML
  - Mikroformáty

Cílené úsilí

- Autority NK ČR
- Vkládání identifikátoru autoritního záznamu do článků

aut.nkp.cz

Autoritní záznam: Národní technická knihovna (Praha, Česko) Zpět

Pro otevření rejstříku **"Předmět. hesla"** klepněte na podtržené **návěští pole**, pro vyhledání připojených záznamů v bázi klepněte na podtržené **heslo**.

Ident. číslo	kn20050320002
Záhlaví	Národní technická knihovna (Praha, Česko)
Odkaz. forma	NTK
Odkaz. forma	Národní technická knihovna v Praze
Viz též	Státní technická knihovna (Praha, Česko)
Zdroj	Skolková, Linda: V Praze vzniknou dvě nové budovy velkých knihoven . (Ikaros [online], Roč. 9, č. 2 (2005))
	www(Národní technická knihovna)
Další informace	Oficiální stránka korporace: http://www.techlib.cz/cs/
	Wikipedie (Národní technická knihovna)

Další informace:

WIKIPEDIE Otevřená encyklopedie

© 2009 Ex Libris, NK ČR

cs.wikipedia.org

Reference [editovat]

1. ↑ V Praze byla otevřena nová Národní technická knihovna
2. ↑ http://www.techlib.cz/cs/o-nas/narodni-technicka-knihovna/
3. ↑ *Grand Biblio*, červen 2009, roč. 3, čís. 7-8, s. 23-33. Dostupné online.

Seznam děl
v databázi Národní knihovny ČR, jejichž tématem je *Národní technická knihovna*

Externí odkazy [editovat]

- Národní technická knihovna
- Vizualizace vítězného projektu
- NTK z hlediska energetické náročnosti

Wikimedia Commons nabízí obrázky, zvuky či videa k tématu *Národní technická knihovna*

Kategorie: ČVUT | Dejvice | Knihovny v Praze

Vazby na autority NK ČR

- Šablona vkládaná (ručně) do článků
- Skrytá kategorie
- Dotaz přes MediaWiki API
- Vlepeno JavaScriptem do Alephu
  - (Chcete taky?)

- Momentálně svázáno přes 6 700 článků

Cílené úsilí (2)

- Zeměpisné souřadnice
- Do článků patří tak jako tak
- Dají se z nich nějak dostat?
- Momentálně těžko, ale jde to
  - Na anglické Wikipedii mikroformáty, u nás momentálně ne
- `http://maps.google.com/maps?lci=org.wikipedia.cs`

Cílené úsilí (3)

- Bibliografické citace
- Vložení COinS
- Úprava citačních šablon

Literatura

- HEUSELER, Holger. *Mars: Pathfinder, Sojourner a dobývání rudé planety*. Praha : Mladá fronta, 1999. ISBN 80-204-0794-4. Find in a Library
- CARR, Michael H. *The surface of Mars*. New York : Cambridge University Press, 2006. ISBN 0-521-87201-4. Find in a Library
- ČEMAN, Róbert; PITTICH, Eduard. *Vesmír - 1 Slunečni soustava*. Bratislava : Mapa Slovakia, 2002. ISBN 80-8067-072-2. S. 192-227. Find in a Library

Cílené úsilí – a co vy?

- Máte nějaká data?
- Chcete nějaká data?
- Nápady na užitečné mikroformáty?
- Pokud to bude možné a užitečné, dá se to zařídit!

- Ozvěte se!

Díky za pozornost!

- petr.kadlec@gmail.com
- http://cs.wikipedia.org/wiki/User:Mormegil
- http://wikimedia.cz/