NTK

50°6'14.083"N, 14°23'26.365"E
Národní technická knihovna
National Library of Technology

210 mm

Linked data
(nejen) v knihovnách

Milan Janíček
milan.janicek at techlib.cz

odd. rozvoje elektronických služeb
Národní technická knihovna
Praha

Searching Session 2013, 8.10.2013, NTK Praha

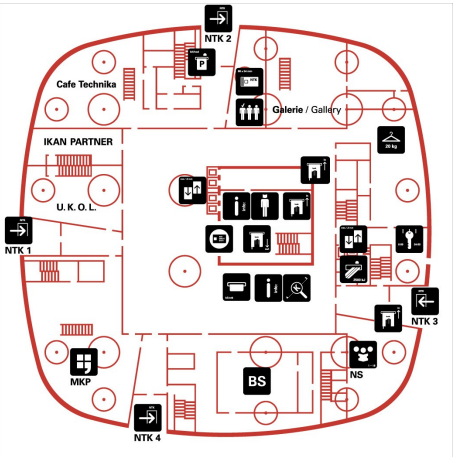
NTK

50°6'14.083"N, 14°23'26.365"E
Národní technická knihovna
National Library of Technology

210 mm

Plán

1) Web a sémantický web
2) Technologie sémantického webu
3) Tvorba dat v knihovně
4) Sbohem, a díky za ryby
5) Linked data – slovníky a datasety
6) Linked data a knihovny
7) Služby postavené na linked data
8) Shrnutí



210 mm

Sémantický web

termín pochází od Tima Berners-Lee

vynálezce World Wide Webu

vynález (Webu) měl 4 části:

- Hypertext Transfer Protokol (HTTP) - komunikace
- URL / URI adresy pro stránky
  - UR = Uniform Resource ...
  - URL – location – umístění
  - URI - identifier – identifikátor - bude hrát roli později ;-)
- HTML - značkovací jazyk pro tvorbu dokumentů
- web server, web browser

210 mm

World Wide Web

- byl postaven na propojení dokumentů - stránek
- byl určen pro lidi
- syntaxe HTML se začala brzo používat pro vizuální prvky
  - význam obsahu byl nejasný...
  - .. a strojově nerozpoznatelný
  - postupně docházelo ke snaze (opět) oddělit význam od vzhledu
    - např. s využitím kaskádových stylů (CSS)

210 mm

Sémantický web

- větší důraz na popis obsahu (sémantika)
- web nejen k propojení dokumentů, ale i k propojení dat

- důležitý prvek: umožnit strojům zpracovávat vztahy

- zapojení ontologií a technologie Resource Description Framework (RDF)

210 mm

RDF – Resource Description Framework

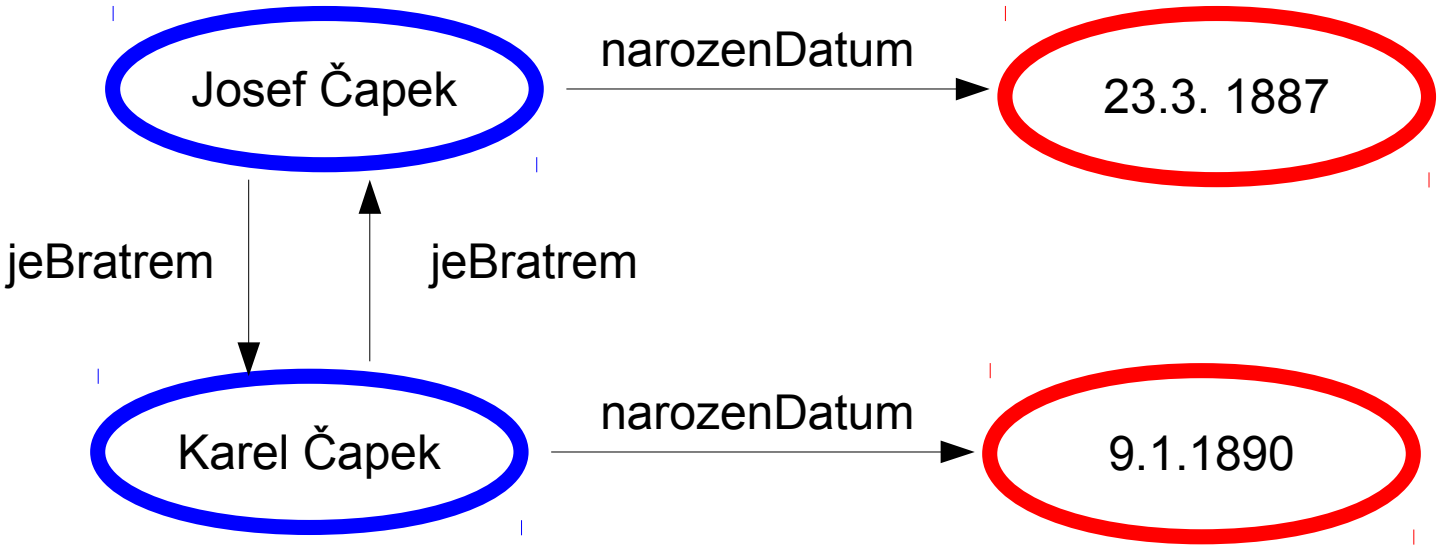
- mechanismus umožňující popsat vztahy
  - ve formě triplu (trojice): subjekt, predikát, objekt

  - vztah: Máma mele maso.
    - subjekt (kdo) = máma
    - predikát (popis vztahu) = mele
    - objekt (co) = maso

210 mm

RDF – Resource Description Framework

- Karel Čapek se narodil 9.1. 1890
- Josef Čapek se narodil 23. 3.1887
- Karel Čapek je bratrem Josefa Čapka



210 mm

Ontologie

- „explicitní specifikace konceptualizace“
- popisuje množinu konceptů v nějaké doméně a jejich vztah
  - na základě popsaných zákonitostí lze odvozovat nové informace

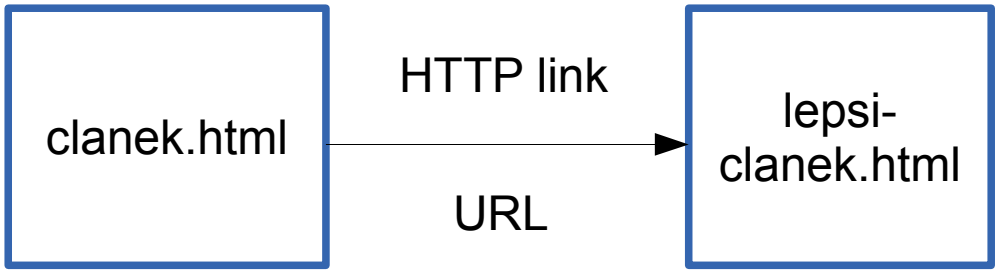
- *osoba* se narodil *datum*
- *muž* je bratrem *osoba*

- ontologie mohou být velmi komplexní
- … někdy je ovšem užitečná i jednoduchá ontologie ;-)

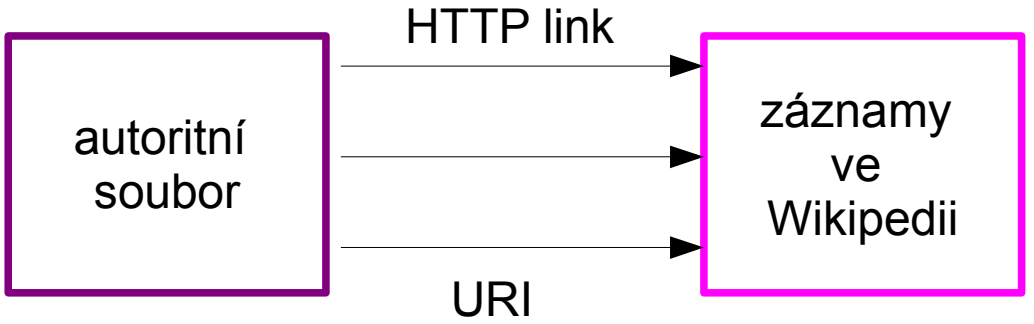
210 mm

Linked data

- WWW propojil dokumenty



- Sémantický web chce propojovat data



210 mm

Požadavky na linked data

- je potřeba jasná struktura
  - RDF, slovníky – umožňují nalezení společného jazyka
- možnost k datům přistupovat a využívat je
  - licence – ideálně s co nejmenšími omezeními (CC0)
- využití unikátních identifikátorů
  - URI – (najít společné body ;-)
- zveřejnění a přístupnost
  - HTTP – standardní způsob přístupu

- video: http://commons.wikimedia.org/w/index.php?title=File%3ALinked-open-data-Europeana-video.ogv

210 mm

Knihovna

- stará, tradiční, zavedená instituce

- klasifikace dokumentů
  - už v Královské knihovně - 900 př. n.l. Nineveh (Asýrie)

- cílem je v množství dokumentů najít to co hledáme
  - => tvorba (meta)dat
- 1876 - Charles Ammi Cutter – bibliografický systém má umožnit
  - najít knihu u níž zná autora / název / kategorii / téma
  - najít co je v knihovně k dispozici od autora / k tématu / v daném druhu literatury
  - pomoci najít knihu podle edice / jejího typu

NTK

50°6'14.083"N, 14°23'26.365"E
Národní technická knihovna
National Library of Technology

210 mm

Knihovna

- knihovníci mají rádi pořádek
  - klasifikační schémata
    - Dewey Decimal Classification
    - Library of Congress classification
    - a další ;-)
  - autority
    - jmenné autority
      - osoby
    - věcné autority
      - témata

- => knihovna produkuje DATA

210 mm

Knihovna a data

- záznamy o knihách (bibliografické)
- záznamy o lidech (jmenné autority)
  - příklad:
    - http://aleph.nkp.cz/F/?func=find-b&local_base=AUT&find_code=WRD&request=nov%C3%A1k
- řízené slovníky (věcné autority)
  - PSH - polytematický strukturovaný heslář
    - http://psh.ntkcz.cz/skos/
  - MeSH-CZ
    - http://www.medvik.cz/medvik/
- => „zdroje tu jsou“
- využívají se obvykle v rámci ‘k tomu určených systémů’
- (často jsou někde zavřená)

210 mm

Knihovna a … MARC

- jak data propojit výše zmiňovaným způsobem?
- je tu jeden zádrhel: formáty MARC

- MAchine-Readable Cataloging
  - formát ze 60.-70. let
  - mezinárodní standard 1973 (ISO 2709)
  - formát umožňující výměnu bibliografických záznamů prostřednictvím počítačů

- význam “machine readable” se za 40 let poněkud posunul

- viz příští slide..

01496nam a2200397 a
4500001001000000030009000100050017000190080041000 36FMT000500077020
0018000820400016001000800018001161000026001342450226001602500020003 8
6260005700406300001500463504002400478650002900502650003500531653001 0
0056670000220057 6KPW001000598OWN001100608CAT002600619CAT00400064
5CAT004100685CAT003800726CAT004100764CAT004100805CAT004100846CAT0
04100887CAT004100928CAT00410096991000880101 0^^000000003^^CZ
PrSTK^^20110107124245.0^^990607s1995 xr f| |0|1|0|cze d^^ BK^^
^_a80-7050-228-2^^ ^_aABA013^_bcze^^ ^_a025.31^_2undef^^1 ^_aMaxwell,
Margarett F.^^10^_aPříručka k AACR2 :^_brevize 1988 : výklad a příklady k
Anglo-americkým katalogizačním pravidlům /^_cMargarett F. Maxwell ; s novou
kapitolou Judith A. Carter ; český překlad Národní knihovna České republiky^^ ^_a1.
české vyd.^^ ^_aPraha :^_bNárodní knihovna České republiky,^_c1995^^ ^_aix,
435 s.^^ ^_aObsahuje rejstřík^^07^_apravidla
popisu^_xin^_2psh^^07^_aidentifikační popis^_xin^_2psh^^0 ^_aAACR2^^1
^_aCarter, Judith A.^^ kpw8146^^ ^_aPUBLIC^^ ^_c20070102^_lSTK01^_h0648^^
^_aSEBKOVAE^_b10^_c20070605^_lSTK01^_h0958^^
^_aHOLECKOVA^_b10^_c20070615^_lSTK01^_h0917^^
^_aSMUTNY^_b10^_c20070717^_lSTK01^_h1447^^
^_aHOLECKOVA^_b10^_c20090105^_lSTK01^_h1516^^
^_aJANECKOVA^_b10^_c20090302^_lSTK01^_h1124^^
^_aHOLECKOVA^_b10^_c20090309^_lSTK01^_h1511^^
^_aKOZUCHOVA^_b10^_c20090717^_lSTK01^_h0922^^
^_aKOZUCHOVA^_b10^_c20090717^_lSTK01
^_h0923^^ ^_aJANECKOVA^_b10^_c20110107^_lSTK01^_h1242^^
^_aABA013^_bE 15882^_bSF 182^_bSF 31/96^_bSF 26/96^_bSF 27/96^_bSF
29/96
^_bSF 30/96^_bSF 01088^^^]

210 mm

MARC … must die!

- problémy s formátem MARC
- http://marc-must-die.info/index.php/MARC_issues
- význam závisí na obsahu jiných polí
  - význam pole **245$b** je určen obsahem pole **245$a**
    - **245** 1 0 **$a** Beginning JSP, JSF and Tomcat : **$b** Java web development / **$c** Giulio Zambon
    - **245** 1 0 **$a** Java servlet and JSP cookbook : **$b** [practical solutions to real-world problems] / **$c** Bruce W. Perry
    - **245** 0 0 **$a** National Technical Library = **$b** Národní technická knihovna = Bibliotheque technique nationale = [Guo li ke xüe ji shu tu shu guan] = Biblioteca Técnica Nacional : 50°6'14.376"N, 14°23'26.613"E / **$c** [texts Roman Brychta ... et al. ; photography Andrea Lhotáková]

210 mm

MARC … must die!

- hodnoty jsou smíchané s dalšími informacemi
  - pole 020 obsahuje ISBN a další informaci
    - **020 $a** 80-85282-70-4 (brož.)
    - **020 $a** (nev.)
    - **020 $a** (váz.)
- hodnota se používá jako identifikátor
    - **100** 1 **$a** Satrapa, Pavel, **$d** 1964- **$7** mzk2002148247
    - **100** 1 **$a** Satrapa, Pavel **$4** aut
- https://vufind.techlib.cz/vufind/Search/Results?lookfor=perl+pro+zelen%C3%A1%C4%8De&type=AllFields&submit=+Hledat

NTK
50°6'14.083"N, 14°23'26.365"E
Národní technická knihovna
National Library of Technology

perl pro zelenáče | Všechna pole | Hledat

Pokročilé

Jednotné přihlášení NTK

Česky

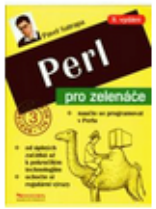
Domovská stránka › Vyhledávání: perl pro zelenáče

Doporučená témata mezi výsledky.
objektově orientované programování (1) programovací jazyk Perl (1)

Zobrazuji **1** - **2** z **2** pro vyhledávání: **'perl pro zelenáče'**, doba hledání: 0.12s

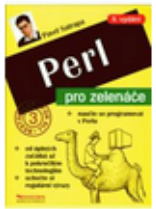
Seřadit podle: Relevance

Alternativní vyhledávání:
perl pro » perlu pro , pedal pro



Perl pro zelenáče : naučte se programovat v Perlu /
Autor: Satrapa, Pavel **Vydáno:** 2001
Umístění: Sklad
Dostupný
Kniha

Přidat k oblíbeným



Perl pro zelenáče /
Autor: Satrapa, Pavel, 1964- **Vydáno:** 2001
Umístění: regál 6D026
Vypůjčeno
Kniha

Přidat k oblíbeným

Vyhledávací nástroje: Vytvořit RSS | Poslat e-mailem | Uložit vyhledávání

Zúžit vyhledávání

Sbírka
NTK (1)
VŠCHT (1)

Formát
Kniha (2)

Call Number
Q - Věda (1)

Autor
Satrapa, Pavel (1)
Satrapa, Pavel, 1964- (1)

Jazyk
čeština (2)

Žánr
příručky (2)

PSH
objektově orientované programování (1)
programovací jazyk Perl (1)

210 mm

MARC … must die!

- vyskytuje se několikanásobné zadávání jednoho údaje
  - informace o vydání
    - 008 000316s2001 xr f 001 0 cze d
    - 260 $c c2001

- ⇒ všechna pravidla pro zápis do MARCu se musí brát v potaz a silně komplikují strojové zpracování

210 mm

v MARCu

- existuje velký objem dat
  - ...která by se dala využít
- používají se komplexní pravidla
  - AACR2 (-> RDA)
- formát je ale technicky zastaralý
- formát se těžko využívá v současných aplikacích
  - (ne že by byl špatně, ale byl určen k něčemu jinému)

- LoC v roce 2011 zahájila práci na frameworku BIBFRAME - měl by v budoucnu nahradit MARC
- více informací například v prezentaci Thomas Meehan: Beyond MARC: MARC, linked data, and Bibframe
  - http://www.slideshare.net/orangeaurochs/marcld2013

210 mm

Linked data

- opustíme na chvíli knihovnu... necháme ji vytvářet záznamy
- jak se tvoří linked data?
- důležitá je struktura
  - určují ji “slovníky” (ontologie)
- důležitý je obsah

210 mm

Linked data - slovníky

- ontologie určující podobu triplů
  - Linked Open Vocabularies: http://lov.okfn.org/dataset/lov/

- FOAF
  - popisuje lidi a jejich vztahy
  - umožňuje popisovat osoby (jméno, mail, obrázek..) a vztahy mezi nimi
- Dublin Core
  - jednoduchý (15 prvků) i rozšířený Dublin Core
  - Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage, Rights
- umožňuje popsat dokument

210 mm

Linked data - slovníky

- SKOS
  - http://www.w3.org/2009/08/skos-reference/skos.html
  - simple knowledge organization systém
  - pro tvorbu řízených heslářů, klasifikačních schémat, tezaurů
  - umožňuje vytvořit vztahy mezi koncepty

```
<skos:Concept rdf:about="http://psh.ntkcz.cz/skos/PSH13629">
        <skos:inScheme rdf:resource="http://psh.ntkcz.cz/skos/"/>
         <dc:identifier>PSH13629</dc:identifier>
         <skos:prefLabel xml:lang="cs">sémantický web</skos:prefLabel>
        <skos:prefLabel xml:lang="en">semantic web</skos:prefLabel>
</skos:Concept>
```

- využití slovníků umožňuje interoperabilitu dat (přes nějaké společné prvky - třeba ~~jméno~~ identifikátor člověka)

210 mm

Linked data – datové zdroje

- datové zdroje
  - http://www.w3.org/wiki/TaskForces/CommunityProjects/LinkingOpenData/DataSets (starší stránka)
  - http://datahub.io/dataset (novější stránka)
- samotný obsah

- DBpedia
  - využívá strukturovaného textu z Wikipedie
  - strukturovaný text lze extrahovat, převést do triplů a dotazovat se nad ním
  - příklady dotazů (a odpovědi):
    - http://dbpedia.org/OnlineAccess#h28-6

210 mm

Linked data – datové zdroje

- Geonames
  - informace o 6 milionech míst
- data.gov
  - data americké vlády - 1000 datasetů
- data.gov.uk
  - data britské vlády - 3600 datasetů
- Eurostat
  - statistická data Evropské unie
- BBC
  - webové stránky BBC využívají linked data - snadněji se vytváří kontextové propojení
- … a další

NTK

50°6'14.083"N, 14°23'26.365"E
Národní technická knihovna
National Library of Technology

210 mm

Data knihoven, archivů a muzeí

- co mohou nabídnout data paměťových institucí?

- bibliografické záznamy
- autority
- řízené slovníky

- metadata o archivních materiálech a sbírkách
- digitalizační projekty
- obsahy knih
- statistiky využívanosti
- informace o lokacích,
- seznamy literatury
- ...

NTK

50°6'14.083"N, 14°23'26.365"E
Národní technická knihovna
National Library of Technology

210 mm

Data knihoven, archivů a muzeí

- co nabízejí?
- OCLC video: http://www.youtube.com/watch?v=fWfEYcnk8Z8

- open bibliographic data - http://datahub.io/group/bibliographic

- British National Bibliography
- LIBRIS
  - švédský souborný katalog
- Harvard
  - téměř 12 milionů záznamů z Harvard University
- Europeana
  - výstupy digitalizace (včetně NDK)

NTK

50°6'14.083"N, 14°23'26.365"E
Národní technická knihovna
National Library of Technology

210 mm

Data knihoven, archivů a muzeí

- VIAF
  - Virtual International Authority File
  - autoritní záznamy OCLC získané od národních partnerů
- Project Gutenberg
- University of Huddersfield
  - data o výpůjčkách a doporučeních
- Nature Publising Group
  - data o 900.000 článcích
- Zeitschriftendatenbank
  - údaje ze německého souborného katalogu časopisů
- Library of Congress Subject Headings
- Polythematic Structured Subject Heading System
  - polytematický strukturovaný heslář

210 mm

Služby postavené na linked data

- LIBRIS
  - http://libris.kb.se/
  - švédský souborný katalog
  - postavený na linked data
  - využívá SKOS, VIAF, BIBO (bibliographic ontology), Dublin Core
  - odkazuje na DBpedii

210 mm

Služby postavené na linked data

- GoPubMed
  - http://www.gopubmed.org/
  - Technische Universität Dresden
  - vyhledávání v PubMedu pomocí hesláře
  - data: Medline
  - slovníky: Gene Ontology, MESH

210 mm

Služby postavené na linked data

- Trenches to triples
  - http://data.aim25.ac.uk/about_t3.php
  - experimentální projekt - rozšíření metadatových záznamů z archivu King's College o sémantické prvky
  - vlastní koncepty (týkající se 1.sv. války)
  - data: Library of Congress, geonames
  - rozšířený záznam: http://www.kingscollections.org/catalogues/lhcma/collection/m/ma76-001

210 mm

Služby postavené na linked data

- Linked Jazz
  - http://linkedjazz.org
  - popis vztahu mezi jazzovými hudebníky na základě přepisů archivních rozhovorů s využitím open linked data a crowdsourcingu
  - data: DBpedia, VIAF + vlastní přepisy rozhovorů
  - slovníky: FOAF, Relationship Ontology, Music ontology

210 mm

Proč linked data?

- umožňují lepší využití znalostí a dat vytvářených knihovnami
  - potenciálně může zvýšit význam knihoven
- tvoří nové propojení s webem
- je to přirozené pokračování toho o co knihovnám jde - zpracování a zpřístupnění znalostí

210 mm

Problémy zavádění linked data

- je nutná určitá změna uvažování
  - menší kontrola nad daty
    - minimální kontrola nad cizími datasety
    - větší závislost na cizí práci
  - MARC21 bude muset být nahrazen
  - není úplně jasný výsledek
    - většinou se data vystaví a pak se ukáže co se s nimi stane ;-)
      - například využití PSH na Univerzitě Pardubice
        - http://www.upce.cz/vvr/lide.html
- problém s licencováním dat
  - ideální licence pro další zpracování je CC0 - Public Domain *)

*) CC0 ovšem není kompatibilní s českým právem, vhodnější je licence Open Data Commons Public Domain Dedication and License
více se problematice věnuje dokument http://www.techlib.cz/files/download/id/3157/open-bibliographic-data-ntk-studie-2012.pdf

210 mm

...

- velcí hráči se novému trendu přizpůsobují / se přizpůsobili
  - Library of Congress
  - Deutsche Nationalbibliothek
  - LIBRIS
  - Harvard University
  - Medline
  - British Library
  - ...

  - a co v ČR???

NTK

50°6'14.083"N, 14°23'26.365"E
Národní technická knihovna
National Library of Technology

210 mm

Linked data
(nejen) v knihovnách

Děkuji za pozornost!

Milan Janíček
milan.janicek at techlib.cz