

PŘEDNÁŠKA 2.

Metadata



Martina Husáková

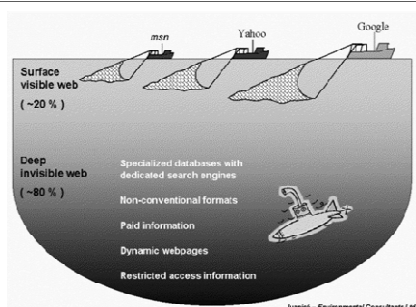
20. 2. 2014

Obsah přednášky

- Charakteristika webového prostoru
- Data, informace a znalosti
- Role a účel metadat
- Typy metadat
- Metadatové frameworky
- Tvorba metadat
- Praktické problémy a budoucnost

2

Struktura webového prostoru

[Zdroj URL: <http://calhoun.globa2.vic.edu.au/2010/05/08/deep-web-vs-surface-web/>]

3

Webový prostor

- „**Viditelný web**“ („Povrchový web“): web tvořen zejména
 - statickými, veřejně dostupnými HTML stránkami
 - PDF soubory, soubory typu Word, Excel, Corel Draw
 - obrázky, video sekvencemi, hudebními soubory, které jsou indexované současnými tradičními vyhledávači (Google, Bing)
- Indexováno jen zhruba 20% celkového webového prostoru
- „**Skrytý web**“ („Neviditelný web“): web tvořen zejména
 - daty uloženými v databázích
 - dynamickými webovými stránkami
 - stránkami vyžadující zadání identifikačních údajů od uživatele,
 - daty prezentovanými na webu v nekonvenčních formátech
- Většinová část webového prostoru neindexovaná současnými vyhledávači - zhruba 55 – 80%

4

Požadavky na obsah webu

- Čitelnost webových dat a informací pro člověka i stroj (webového agenta)
- Pro zpracování dat stroji je nezbytný sebe-popisný charakter webového obsahu
- Specifikace standardizovaných přístupů pro „porozumění“ obsahu webu stroji
- **Metadata** jsou řešením pro poskytnutí významu datům, která jsou na webu

5

Metadata

- **Metadata**: klíčový aspekt sémantického webu
Obsah sémantického webu = data + metadata
- Smysl: obohacení dat na webu o významově bohaté popisy (výroky) pro dosažení automatizovaného zpracování webových dokumentů webovými agenty
- Přejít od webu informačního k webu znalostnímu

6

Metadata: definice

- **Metadata**
 - řecky *meta* (navíc, po, s, změna)
 - latinsky *datum* (co je dáno, definice, popis)
- Různý pohled na metadata různými profesionály => různé definice => příčina možných nedorozumění
- Tradiční (zjednodušená) definice:
 - „**Metadata jsou data o datech.**“
- Další definice:
 - „Metadata jsou informace o datech.“
 - „Metadata jsou informace o informacích.“
 - „Metadata obsahují informace o datech.“

[Definice metadat: Zdroj URL: <http://www.websters-online-dictionary.org/definitions/Metadata>]

7

Data, informace a znalosti

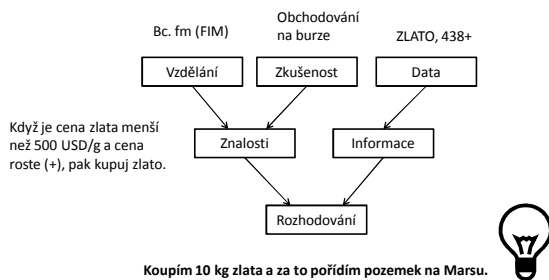
- **Data:** získané a zachycené údaje bez významu
 - forma: text, kód, obraz, zvuk, symbol
 - základ pro vznik informací nebo znalostí
- **Informace:** data s významem
 - odpovědi na otázky: kdo?, co?, kde?, kdy?
 - Cíl: snížit entropii
- **Znalosti:** informace nebo data zacláčená do souvislosti s cílem porozumět nebo řešit problém
 - odpověď na otázku jak?
 - forma: postup, návod, soubor pravidel, vztahy
- **Moudrost:** znalost vedoucí k porozumění



[Zdroj] kniha: J. T. Pollock: *Semantic Web For Dummies*. Wiley, 2009. str. 118.]

8

Data, informace, znalosti – příklad 1



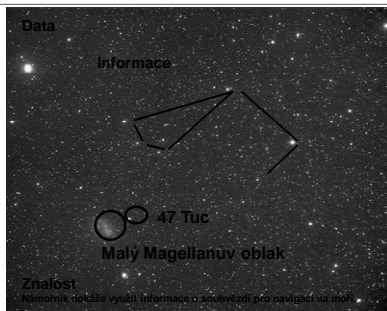
9

Data, informace, znalosti – příklad 2

- [illegible]

10

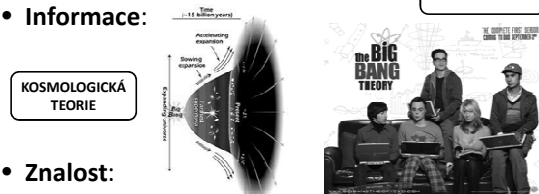
Data, informace, znalosti – příklad 3



11

Data, informace, znalosti – příklad 4

- **Data:** Теория Большого Взрыва
- **Informace:** 



- **Znalost:**
- Plánování a rozhodování:
„Nebudu mít čas učit se s Tebou po večerech na zkoušku ze ZT2, protože od pondělí do čtvrtka od 21:20 dávají Teorii veľkého ťiesku na Prima Cool!“

[Zdroj URL: http://www.world-science.net/exclusives/060929_energycons.htm]

[Zdroj URL: <http://img543.imageshack.us/img543/621/69616826.jpg>]

12

Příklad metadat: Dublin Core

- Pracovní setkání v městě Dublin r. 1995 realizováno za účelem vytvoření univerzálního způsobu pro popis informačních zdrojů a jejich snadné vyhledávání
- **RDF slovník**
(angl. Dublin Metadata Core Element Set (DC))
- Nabízí soubor metadatových prvků pro popis síťových zdrojů dle standardů knihovnické a informační vědy

19

Příklad metadat: Dublin Core

- Důraz kladen na:
 - jednoduchost (metadatové záznamy tvořené i nezaškolenými autory zdrojů)
 - sémantickou interoperabilitu mezi složitějšími formáty
 - mezinárodní konsensus: efektivní infrastruktura pro zpřístupňování zdrojů na webu
 - rozšiřitelnost: v kódování struktury a komplikovanější sémantice pro popis zdrojů
 - modifikovatelnost

20

Příklad metadat: Dublin Core

Metadatové prvky - Dublin Core Metadata Element Set

Prvek	Název (cs)	URI (RDF popis prvku)
Title	Název	http://purl.org/dc/elements/1.1/title
Creator	Tvůrce	http://purl.org/dc/elements/1.1/creator
Subject	Předmět	http://purl.org/dc/elements/1.1/subject
Description	Popis	http://purl.org/dc/elements/1.1/description
Publisher	Vydavatel	http://purl.org/dc/elements/1.1/publisher
Contributor	Přispěvatel	http://purl.org/dc/elements/1.1/contributor
Date	Datum	http://purl.org/dc/elements/1.1/date
Type	Typ	http://purl.org/dc/elements/1.1/type
Format	Formát	http://purl.org/dc/elements/1.1/format
Identifier	Identifikátor	http://purl.org/dc/elements/1.1/identifier
Source	Zdroj	http://purl.org/dc/elements/1.1/source
Language	Jazyk	http://purl.org/dc/elements/1.1/language
Relation	Vztah	http://purl.org/dc/elements/1.1/relation
Coverage	Pokrytí	http://purl.org/dc/elements/1.1/coverage
Rights	Práva	http://purl.org/dc/elements/1.1/rights

[Zdroj URL: Soubor metadatových prvků: http://www.lcs.muni.cz/dublin_core/elems.html
[Zdroj URL: http://cs.wikiversity.org/wiki/Dublin_Core]

21

Příklad metadat: Dublin Core

Příklad reprezentace v DC

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dc="http://purl.org/dc/elements/1.1/">
  <rdf:description rdf:about="http://www.kosmas.cz/knihy/157186/straka-v-risi-entropie/">
    <dc:creator>Markéta Baňková</dc:creator>
    <dc:title>Straka v říši entropie</dc:title>
    <dc:description>
      Straka v říši entropie jsou bajky pro dospělé i děti. Zvířata v nich luští záhady existence a fungování světa. Zjistíme, proč se hroch udržel na vodní hladině i proč někteří samci nepohodnou samičkou z igelitu. Jak lišky procházejí krizí středního věku a kdo všechno chce být ideálem krásy. Proč nemá smysl bojovat proti nepořádku a jak zatočit s jezvcem notorikem. Jak může notorik pád ze schodů omluvit zakřivením časoprostoru a že i myši mají určitý názor na kvantovou neurčitost.
    </dc:description>
    <dc:date>2010-11-20</dc:date>
  </rdf:description>
</rdf:RDF>
```

22

Příklad metadat: EXIF

EXIF (Exchangeable Image File Format)
Většina formátů digitálních fotografií obsahuje přidružené informace (metadata) k fotografii.

Skupiny metadat:

- informace o fotoaparátu
- nastavení fotoaparátu pro snímek
- poznámky výrobce
- GPS údaje
- audio poznámka
- ostatní (ver. EXIFU)

Využití: studium nastavení parametrů fotografie může vést ke zlepšování kvality fotek do budoucna

Klíčový rys: úprava EXIF informací nelze měnit výslednou fotografii



[Zdroj URL: http://www.fotografarovi.cz/art/fotekh_dif/exif.html]

23

Příklady metadat: ID3 tag

- ID3 tag: formát audio souborů umožňující připojit metadata ke skladbě
- Zejména určen pro MP3 soubory
- Verze: ID3 ver. 1 a ver. 2
- ID3 ver. 1:
 - informace o umělci, názvu audia souboru, albu, roku publikování a žánru
 - možnost přidání komentářů
 - délka 128 bytů
 - lokace: na konci audio souboru

[Zdroj URL: <http://www.id3.org/>]

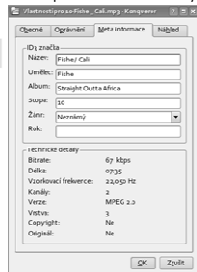
24

Příklady metadat: ID3 tag ver. 1

Specifikace ID3 tagu ver. 1

Sign	Length	Position	Description
(bytes)	(bytes)	(bytes)	
A	3	(0-2)	Tag identification. 'Not contain TAG' if tag exists and is correct.
B	30	(3-32)	Title
C	30	(33-62)	Artist
D	30	(63-92)	Album
E	4	(93-96)	Year
F	30	(97-126)	Comment
G	1	(127)	Genre

Příklad mp3 souboru s metadaty

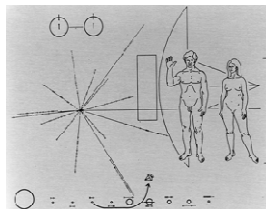


[Zdroj URL: http://mpegedit.org/mpegedit/mpeg_format/mpeghdr.html#MPEGTAG]
[Zdroj URL: http://upload.wikimedia.org/wikipedia/commons/8/80/ID3_in_KDE3_Icon.png]

25

Příklady metadat: Plaketa pro Pioneer

- Plaketa z pozlaceného hliníku umístěná na sondě Pioneer 10 a 11
- Nese poselství případným mimozemským nálezcům
- Autoři: koncept návrhu: C. E. Sagan, F. Drake, umělecká podoba: L. Salzmanová-Saganová

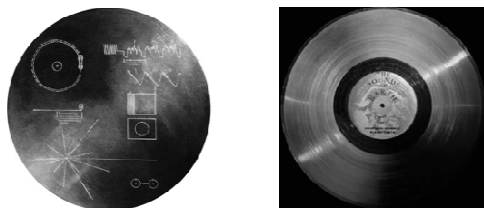


[Zdroj Obrázek: Plaketa: <http://en.wikipedia.org/wiki/File:Pioneer10-plaque.jpg>]
[Zdroj Obrázek: Pioneer: http://historicspacecraft.com/Probes_Outter_Planets.html]

26

Příklady metadat: Plaketa pro Voyager

- Deska z pozlaceného hliníku nesená sondami Voyager 1 a 2 s poselstvím případným mimozemským nálezcům ve formě gramofonové desky
- Účel: deska popisuje návod na použití gramofonové desky obsahující multimediální soubory (obrázky, zvuky, hudební skladby, pozdravy v 55-ti jazycích matematické a fyzikální definice, popis sluneční soustavy, a další.)



[Zdroj URL: Obrázek plakety: <http://electronicbreakfast.blogspot.com/2010/12/voyager-1-another-step-closer-to.html>]
[Zdroj URL: Obrázek gramofon. Desky: <http://www.explainthatstuff.com/record-players.html>]

27

Příklady metadat

- Videoukázka z youtube.com



<http://www.youtube.com/watch?v=ulArB9DAnW4>

28

Metadatový framework (1)

- Sestává z množiny specifikací, které jsou zaměřeny např. na reprezentaci, manipulaci, uložení, údržbu a dotazování se na data a metadata
- Komponenty:
 - Datový model: soubor datových typů a funkcí, pomocí kterých se tvoří abstraktní pohled na webový dokument s možností přístupu k hodnotám uložených v proměnných datových typech

29

Metadatový framework (2)

- Sémantika: význam komponent jazyka pro tvorbu metadat
- Serializační formát: poskytuje meta-jazyk a syntaktické konstrukce pro kódování metadat
 - Příklad: RDF i OWL mají různé datové modely, přičemž jsou serializovány (zapsány) s použitím XML syntaxe
- Dotazovací jazyk: existence různých přístupů pro dotazování (např. RDF – SPARQL, OWL – SQWRL)

30

Metadatový framework (3)

- **Metadatový framework na bázi XML (W3C)**
- Jeden z nejstarších vyvinut s cílem oddělení obsahu od způsobu prezentace webového obsahu
- **Komponenty:**
 - XLink, XPointer: reprezentace hypertextových odkazů
 - XSL specifikace s XSLT (transformace XML dokumentu do formy HTML stránky zobrazitelné na webu)
 - Specifikace Xquery a Xpath: extrakce XML informací z webových dokumentů
 - Specifikace XML Schématu: definice struktury XML dokumentů pro rozšíření jejich sémantiky

31

Metadatový framework (4)

- **Metadatový framework na bázi RDF (W3C)**
- Primárně určen pro reprezentaci metadat přidělených různým webovým zdrojům
- **Komponenty:**
 - Specifikace RDF založená na XML (RDF/XML)
 - Specifikace RDF Schématu: tvorba hierarchií
 - Specifikace SPARQL: protokol a dotazovací jazyk pro zpřístupnění RDF dat

32

Metadatový framework (5)

- **Metadatový framework založený na OWL (W3C)**
- OWL jazyk nabízí konstrukce, které umožňují vyjádřit různé typy omezení a axiomů na úrovni schémat a dat
 - Na úrovni dat: specifikace vztahů mezi jedincem (instancí) a třídou spolu se vztahy different-from a some-as
 - Na úrovni schématu: specifikace vztahů mezi třídami (is-a), disjunktnost a ekvivalenci tříd

33

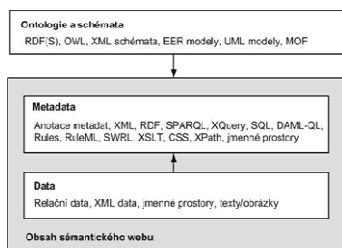
Metadatový framework (6)

- Ontologie OWL lze tvořit v různých dialekttech:
 - OWL Full
 - OWL DL
 - OWL Lite
- Součástí frameworku jsou i dotazovací jazyky:
 - OWL-QL
 - SPARQL
 - SQWRL

34

Jazyky pro tvorbu metadat

- SGML
- (X)HTML
- XML, XML(S)
- RDF (RDFa, eRDF)
- RDF(S)
- Mikroformáty
- DAML+OIL
- OWL
- XTM, LTM



[Zdroj obrázek: Kashyap, V. a kol. The Semantic Web: Semantics for Data and Services on the Web. Springer, 2010. ISBN 978-3-642-09530-6]

35

Standardizace metadat

- Nutné používat standardy pro zajištění jednotného způsobu reprezentace metadat
- Pokud toto nebude zajištěno => snižování hodnoty a významu metadat
- Potřeba standardizace byla řešena již před nástupem služby WWW
- Příklad:
 - Odlišné pojmenování metadatového pole:
 - „Autor“, „Tvůrce“, „Sestavovatel“
 - Odlišné pojmenování hodnoty metadatového pole:
 - Karel Hynek Mácha, K. H. Mácha, Karel H. Mácha

36

Aktivní skupiny na poli standardizace

- **Organizace ISO** a její skupina Metadata Working Group
 - Cíl skupiny: specifikace elementů metadat, klasifikační a kódovací schémata, správa metadat s možnostmi jejich výměny (sdílení)
 - Výsledek: ISO 11179: Specification and Standardization of Data Elements a ISO 15836: Metadata Dublin Core
- **Organizace ANSI** (American National Standards Institute)
 - Cíl: vývoj metamodelu pro reprezentaci dat (pojmenování, identifikace, definice, klasifikace a registrace metadat, ...)
- **Organizace W3C** (World Wide Web Consortium):
 - RDF : specifikace jednoduchých ontologií
 - PICS (Platform for Internet Content Selection: pro kódování a přenos metadat odvozená ze specifikací Dublin Core)

37

Praktické problémy a překážky

- Existence různých projektů pro definici metadat
- Koordinace spíše na vyšší úrovni Dublin Core
- Různé skupiny mohou definovat vlastní způsob tvorby metadat => kombinace již existujících způsobů specifikace metadat
- Problém s kompatibilitou obsahů různých metadat
- Příklad: v Dublin Core specifikaci je element `ResourceType` (typ zdroje) využíváný spíše knihovnami než v oblasti softwaru

38

Praktické problémy konkrétněji

- **Výběr elementů, sub-elementů a schématu**
- Co přitom uvažovat?
 - Specifické potřeby skupiny, která bude metadata v praxi využívat při vyhledávání zdrojů a jejich správě
 - Účelnost zavedení metadatových popisů
 - „Více škody než užitku?“
 - Jednoduchý popis je zřejmě lepší než žádný, protože napomáhá ve vyhledávání zdrojů

39

Shrnutí funkcí metadat

- Shrnutí (popisná funkce): sumarizace obsahu
- Vyhledávání: prohledávání metadat s cílem vyšší přesnosti výsledků
- Doporučení: umožňuje uživateli určit, která data potřebuje
- Vybírání: pomoc při rozhodování, který zdroj informací načíst
- Přístup: zajištění přístupu k datům (např. uvedením přesné lokace)
- Interpretace: instrukce, jak se má s daty zacházet (např. formát, kódování, jazyk, šifrování, ...)
- Specifikace: informace ovlivňující užití dat (např. právní podmínky, velikost, stáří, přístup. práva, ...)
- Historie: popis historie nebo původu dat
- Správa dat: specifikace pro správu objektu (např. datum poslední modifikace, datum vytvoření, ...)
- Propojování a vztahy mezi daty: specifikace vztahů mezi objekty (např. mezi článkem a časopisem, mezi překladem a originálem, ...)

40

Budoucnost metadat


- Tvorba registrů metadatových schémat
 - Pro dostupnost platných verzí rozšířených metadatových schémat
- Některé standardy se stanou vůdčími
- Důležitost vícejazyčnosti při tvorbě metadat (již zakomponováno v RDF)
- Nutné počítat se širokým spektrem zdrojů (obrázky, videa, dynamicky generované objekty, ...)
- Způsoby hodnocení efektivnosti metadat (metriky)
- Teoretické základy pro tvorbu metadat (formální modely, metodologie, úvaha nad právními otázkami)

41

Použitá literatura

- Přednáška byla připravena mj. s použitím následujících zdrojů:
 - Sklenák, V. a kol. Data, informace, znalosti a internet. C. H. Beck pro praxi, 2001, vydání 1. ISBN 80-7179-409-0
 - Kashyap, V. a kol. The Semantic Web: Semantics for Data and Services on the Web. Springer, 2010. ISBN 978-3-642-09530-6
 - Pollock, J. T. Semantic Web for Dummies. Wiley, 2009. ISBN 978-0-470-39679-7

42



Příští přednáška dne 27. 2. 2014 na téma

Reprezentace metadat pomocí
námětových map

43