Téma 7 – Struktury dokumentů, metadatová schémata a datové struktury

- Typy struktur
- Modely struktur dokumentů/dat
- SGML

1

Metafory organizace znalostí – typy struktur

- seznam
- kartotéka
- řetězec
- kruh (cyklus, encyklopedie)
- strom (hierarchie)
- oddenky (rhizoma)
- vrstvy, úrovně
- síť/graf
- tabulka
- stavebnice (fasety, moduly, objektový přístup)

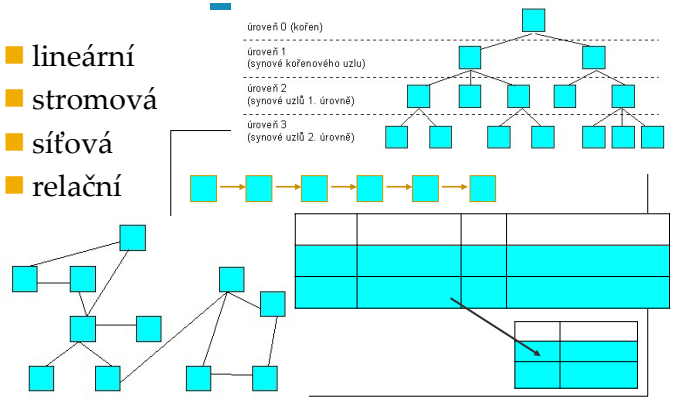
⇒ **Úkol 12**

⇒ **Úkol 13**

2

Základní typy struktur

- lineární
- stromová
- síťová
- relační



3

Lineární, stromová a síťová struktura = graf

uzel

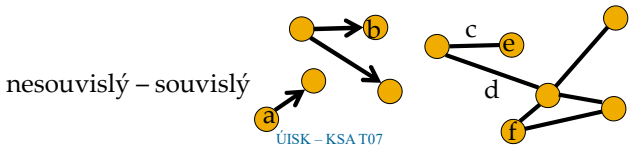
hrana

cesta

orientovaný

ohodnocený

nesouvislý – souvislý



4

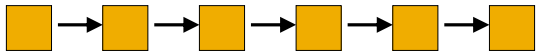
Fakta / obsah

Čtenář4 . na týden „” z knihu World z architecture Prahy() for . the signatura si půjčil 4 Koutecký F202954 20243 14. WideWeb Vladimír Information

Čtenář Vladimír Koutecký z Prahy 4 si dne 14. 4. 2024 na týden půjčil knihu *„Information architecture for the World Wide Web“* (signatura F202954).

5

Lineární (sekvenční) struktura



Vztah: sekvenční (asymetrická) asociace 1:1

6

Lineární (sekvenční) struktura dokumentu

lineární struktura srozumitelná lidem

Čtenář Vladimír Koutecký z Prahy 4 si dne 14. 4. 2024 na týden půjčil knihu *„Information architecture for the World Wide Web“* (signatura F202954).

7

Lineární (sekvenční) struktura dat

VÝPŮJČKY	Jméno	Rodné číslo	Číslo legitimace	Adresa
	Bednář	750512/0235	14	Praha 6

Telefon	Signatura
202832564	A158

Datum	Jméno
12. 7. 2003	Skálová

Adresa	Telefon	Signatura	Autor
Liberec IV	411352765	A526	Tolstoj

Název	Datum	Jméno	Rodné číslo
Vojna a mír	1. 4. 2003	Bednář	750512/0235

Číslo legitimace	Adresa	Telefon	Signatura
14	Praha 6	202832564	A247

Autor	Název	Datum	Jméno
Vian	Pěna dní	3. 8. 2003	Bednář

lineární struktura srozumitelná strojům

8

Lineární (sekvenční) struktura

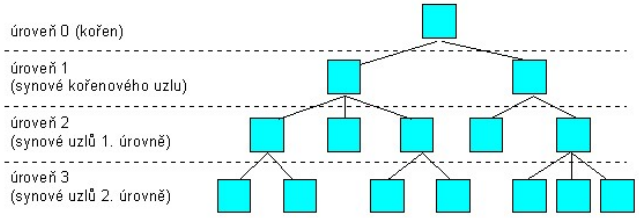
oblasti užití:

- zálohy dat (sekvenční ukládání na magnetickou pásku)
- audio, video
- fulltextové systémy
  (doplněné indexovými soubory)

standardy:

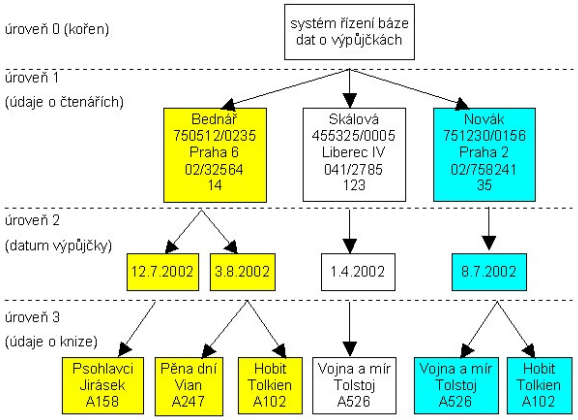
- ISO 2709 – výměnný formát pro bibliografické záznamy MARC

Stromová (hierarchická) struktura

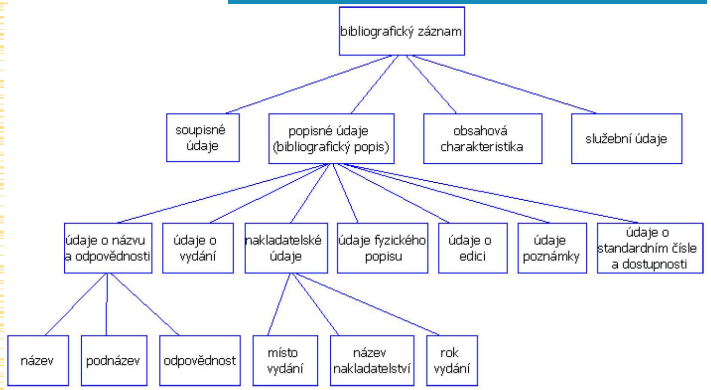


Vztah: monohierarchie 1:N

Stromová (hierarchická) struktura dat



Stromová (hierarchická) struktura bibliografického záznamu



Stromová (hierarchická) struktura

oblasti užití:
- narativní dokumenty
- indexové soubory
- HTML a XML dokumenty
- objektově orientovaná technologie

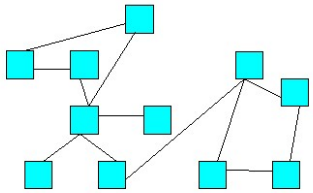
standardy:
- ISO 8879 – SGML
- XML

Stromová (hierarchická) struktura

klady
- možnost vyjádřit hierarchické vztahy (1 : N)
- rychlejší vyhledávání (neprohledává se celý soubor, ale jen příslušné větve)

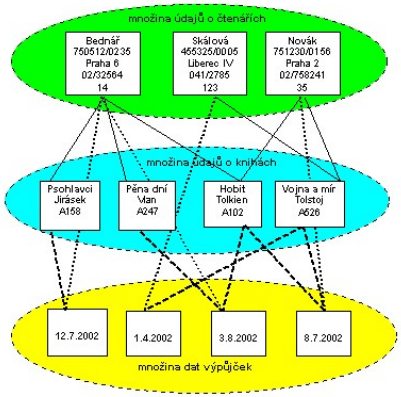
zápory
- nemožnost jednoduchého vyjádření vztahů N : M mezi prvky (bez duplicit)
- strukturu je třeba předem pevně stanovit

Síťová struktura

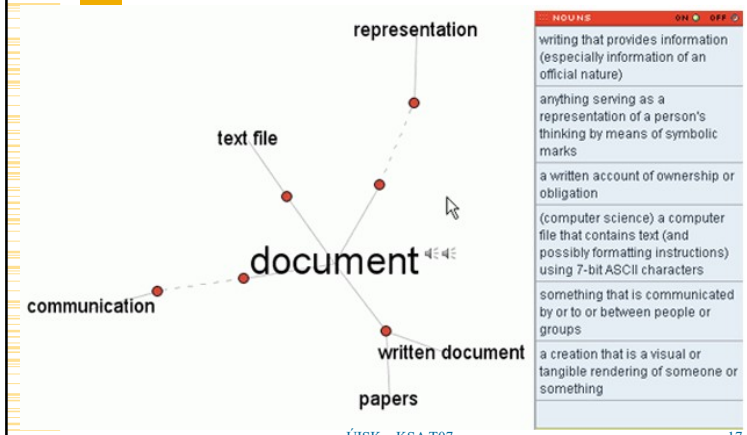


Vztah: asociace, polyhierarchie N:M

Síťová struktura dat

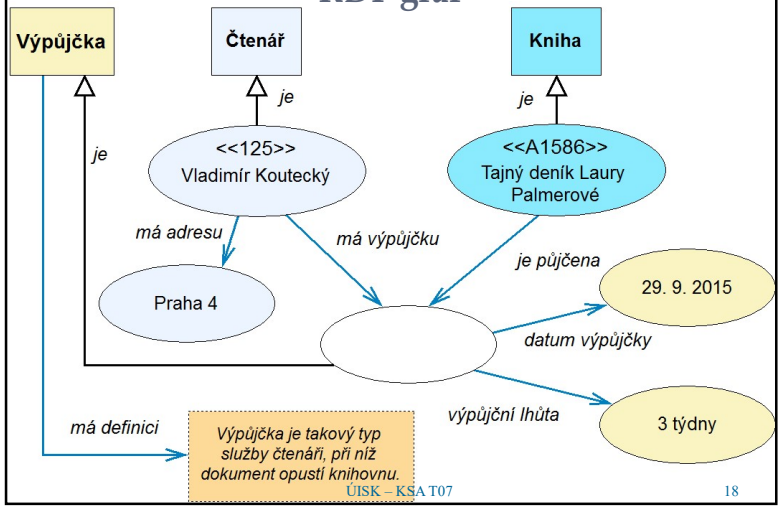


Síťová struktura – WordNet



17

Grafický model obsahu zdroje podle RDF – RDF graf



18

Síťová struktura

klady

- flexibilita při vytváření vztahů 1 : N, N : 1 a N : M
- rychlé vyhledávání (neprohledáváme celý soubor, ale sledujeme přímou cestu k danému prvku stanovenou definovanou vazbou – přímé skoky)
- odstranění redundance

zápory

- vzájemné vztahy mezi prvky se mohou stát velmi komplikovanými a obtížně se modelují
- umožňuje rychle zodpovídat pouze předem připravené dotazy, komplikované dotazy vyžadují provést více kroků

19

Síťová struktura

oblasti užití:

- hypertext
- World Wide Web
- propojená data

standardy:

- CODASYL
- UML
- RDF

20

Relační struktura dat

Jméno	Rodné číslo	Adresa	Telefon	Číslo legitimace
Bednář	751230/0235	Praha 6	202832564	14
Skálová	455325/0005	Liberec IV	411352785	123
Novák	610723/0156	Praha 2	275824741	35

Autor	Název	Signatura
Jirásek	Psohlavci	A158
Tolkien	Hobit	A102
Tolstoj	Vojna a mír	A526
Vian	Pěna dní	A247

Signatura	Datum	Číslo legitimace
A526	8.7.2003	35
A158	12.7.2003	14
A247	3.8.2003	14
A102	3.8.2003	14
A102	8.7.2003	35
A526	1.4.2003	123

Vztah: asociace (relace) N:M

21

Relační struktura

oblasti užití:

- transakční (relační) databáze

standardy:

- ISO/IEC 9075 – SQL

22

MARC21 – jaký typ struktury dat?

015	|a cnb001647463
020	|a 80-86138-78-X |q (váz.)
035	|a (OCoLC)83979461
040	|a ABA001 |b cze
0411	|a cze |a eng |h eng
072 7	|a 165 |x Teorie poznání. Epistemologie |2 Konspekt |9 5
072 7	|a 81 |x Lingvistika. Jazyky |2 Konspekt |9 11
080	|a 165.194 |2 MRF
080	|a 81
080	|a 15
080	|a 16
080	|a (0[illegible]) |2 MRF
080	|a (078.7) |2 MRF
1001	|a Lakoff, George, |d 1941- |7 xx0007852 |4 aut
24510	|a Ženy, oheň a nebezpečné věci : |b co kategorie vypovídají o naší mysli / |c George Lakoff ; [přeložil a doslov napsal Dominik Lukeš]
250	|a Vyd. 1.
260	|a Praha : |b Triáda, |c 2006
300	|a 655 s. : |b il. ; |c 25 cm
4901	|a Paprsek ; |v sv. 11
500	|a Přeloženo z angličtiny
504	|a Obsahuje bibliografie, bibliografické odkazy a rejstříky
5209	|a Multidisciplinární studie amerického lingvisty poskytuje jednotný konceptuální rámec pro nové základy studia lidské mysli a tím i kognitivní vědy.

Skupina proměnných polí (0, 1…)

Podpole

Návěští (leader)

Adresář

Proměnná pole

Záznam

23

MARC21 – jaký typ struktury dat?

015	|a cnb001647463
020	|a
035	|a
040	|a
0411	|a
072 7	|a 165 |x Teorie poznání. Epistemologie |2 Konspekt |9 5
072 7	|a 81 |x Lingvistika. Jazyky |2 Konspekt |9 11
080	|a 165.194 |2 MRF
080	|a
080	|a
080	|a
080	|a
080	|a (078.7) |2 MRF
1001	|a Lakoff, George, |d 1941- |7 xx0007852 |4 aut
24510	|a Ženy, oheň a nebezpečné věci : |b co kategorie vypovídají o naší mysli / |c George Lakoff ; [přeložil a doslov napsal Dominik Lukeš]
250	|a
260	|a
300	|a
4901	|a
500	|a Přeloženo z angličtiny
504	|a Obsahuje bibliografie, bibliografické odkazy a rejstříky
5209	|a Multidisciplinární studie amerického lingvisty poskytuje jednotný konceptuální rámec pro nové základy studia lidské mysli a tím i kognitivní vědy.

Návěští (leader)

Tag pole

Obsah pole

Indikátor

Adresář

Pole

Proměnná pole

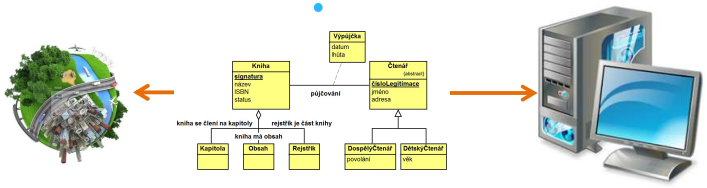
Záznam

24

Sémantická a syntaktická mikroanalýza – modely struktury zdrojů/dokumentů

- Která struktura?
- Obsah?
- Forma?
- Model?

?



obsah sémantika — **struktura** ✓ — forma syntax

25

Informační zdroje jsou…

- Dokumenty
- Objekty
- Data

- Aplikace
- Služby
- Zprávy

- Metadata

Zajímá nás
- obsah
- forma
- počet
- způsob vzniku / vytvoření / získání
- frekvence aktualizace / počet přírůstků
- propojení s funkcemi a s uživateli

26

Úrovně granularity prvků klasifikace zaznamenaných znalostí

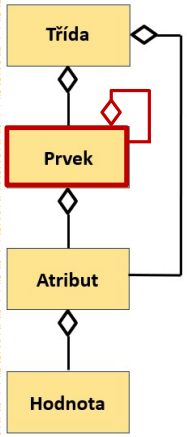
mikro-

- Údaj
- Výrok
- Část (prvek, element) dokumentu
- Dokument
- Agregát, integrační zdroj
- Kolekce, datová sada
- Repozitář (úložiště)

makro-

27

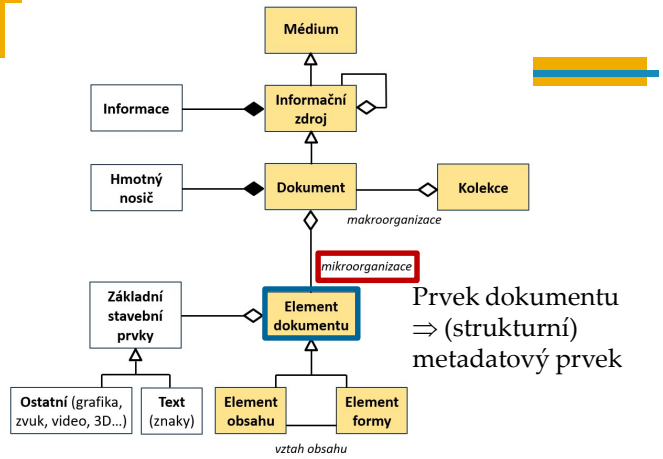
Rekapitulace pravidel systémové analýzy a jejich aplikace v analýze struktury dokumentů



1. **Prvek <je instancí> třídy.**
   Třída/kategorie je množina prvků se stejnými vlastnostmi.
2. **Prvek/třída/kategorie <má> atribut.**
   Atribut je specifikací prvku nebo třídy.
3. **Atribut <má> hodnotu.**
   Třídy ani prvky nemají hodnoty.
4. **Vztah <je> atribut** patřící více prvkům nebo třídám.
5. **Metadata**
   = atributy/prvky (dokumentů) **+ hodnoty.**
6. **Atribut nebo třída/kategorie ⇒ faseta**

28

Granularita informačních zdrojů



Prvek dokumentu
⇒ (strukturní) metadatový prvek

29

Obsah a forma informace

1) **Informace = data, která mají smysl**

   ⇒ obsah
   zpráva, sdělení, message, content

2) **Informace = znalosti, které jsou sdělitelné**

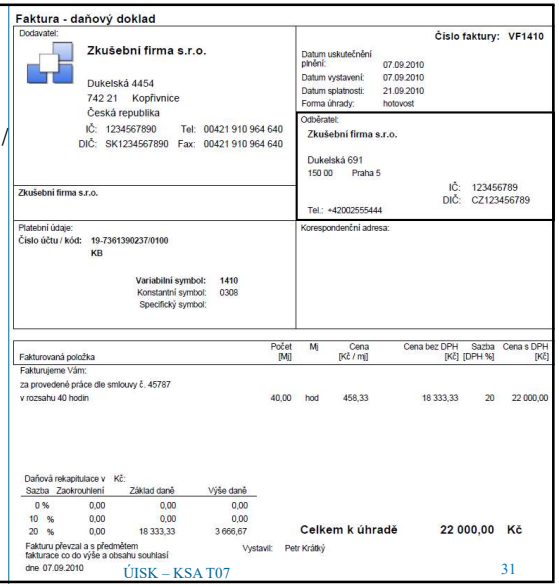
   ⇒ komunikace – formát, kód
   kanál, médium, container, carrier, code

MARC21
336 Typ obsahu
337 Typ média
338 Typ nosiče

30

Co definuje model struktury dokumentu

- **prvky** (elementy) / entity
- **atributy**
- **vztahy**
  - **ekvivalence**
  - **hierarchie** (generická, partitivní)
  - **asociace** (obecná, sekvenční)
- **výskyt prvku v dokumentu**
  - **povinnost**
  - **násobnost**



Faktura - daňový doklad

Dodavatel: Zkušební firma s.r.o.
Dukelská 4454
742 21 Kopřivnice
Česká republika
IČ: 1234567890 Tel: 00421 910 964 640
DIČ: SK1234567890 Fax: 00421 910 964 640

Číslo faktury: VF1410
Datum uskutečnění plnění: 07.09.2010
Datum vystavení: 07.09.2010
Datum splatnosti: 21.09.2010
Forma úhrady: hotovost

Odběratel: Zkušební firma s.r.o.
Dukelská 691
150 00 Praha 5
IČ: 123456789
DIČ: CZ123456789
Tel.: +42002555444

Zkušební firma s.r.o.

Platební údaje:
Číslo účtu / kód: 19-7361390237/0100
KB
Variabilní symbol: 1410
Konstantní symbol: 0308
Specifický symbol:

Korespondenční adresa:

Fakturovaná položka	Počet [MJ]	MJ	Cena [Kč / mj]	Cena bez DPH [Kč]	Sazba [DPH %]	Cena s DPH [Kč]
Fakturujeme Vám: za provedené práce dle smlouvy č. 45787 v rozsahu 40 hodin	40,00	hod	458,33	18 333,33	20	22 000,00

Daňová rekapitulace v Kč:

Sazba	Zaokrouhlení	Základ daně	Výše daně
0 %	0,00	0,00	0,00
10 %	0,00	0,00	0,00
20 %	0,00	18 333,33	3 666,67

Celkem k úhradě 22 000,00 Kč

Fakturu převzal a s předmětem fakturace co do výše a obsahu souhlasí dne 07.09.2010

Vystavil: Petr Krátký

31

Model / schéma / typ / prototyp / šablona / profil / architektura dokumentu

1 : N

- **model:**
  zjednodušená a zobecněná (abstraktní) reprezentace struktury dokumentu
  forma
  generická (obecná) struktura
- **instance:**
  konkrétní obsah dokumentu s danou strukturou
  obsah
  specifická struktura

32

Modely struktur dokumentů / dat

Cíl:

Najít **obecnou / otevřenou / generickou** strukturu (formát, architekturu)

- použitelnou pro co největší počet typů zdrojů (heterogenita)
- nezávislou na platformě, tj. na použitém hardwaru a softwaru
- umožňující distribuovat informační zdroje
- srozumitelnou lidem i počítačovým programům

Realita: Formáty elektronických dokumentů

text	obraz (image)	zvuk (audio)	video
▪ASCII, TXT	▪JPEG	▪MP3	▪MPEG
▪PostScript	▪GIF	▪FLAC	▪MOV
▪PDF	▪TIFF	▪RAM	▪WMV
▪DOC, DOCX, RTF	▪PNG	▪WMA	▪AVI
▪SGML, HTML, XML	▪BMP	▪AAC	▪SWF
▪ODT	▪SVG		

Realita: Standardy pro struktury elektronických dokumentů

Rekapitulace

Jazyky pro modelování dokumentů / dat

1) SGML/XML DTD
2) XML Schema

elektronické dokumenty (texty)

3) ERA – entity relationship attribute
4) ✓UML – unified modeling language (ISO 19501)

softwarové aplikace

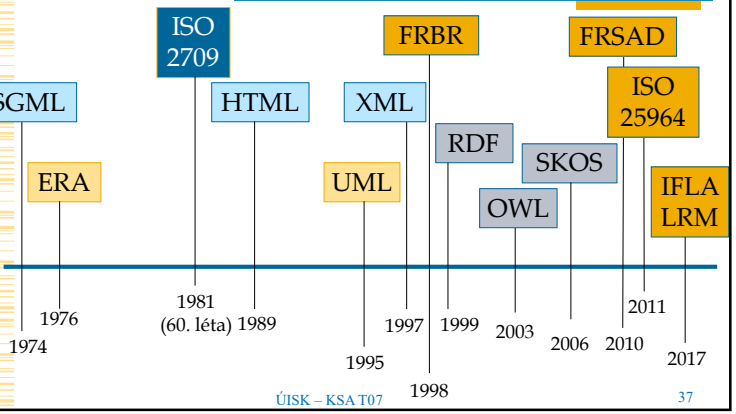
5) ✓RDF – Resource description framework
formát metadat
6) RDFS – RDF schema
7) OWL – Web ontology language
obsah metadat
8) ✓SKOS

(sémantický) web, propojená otevřená data (LOD)

9) Topic maps – mapy námětů (ISO 13250)

umělá inteligence

Modely dokumentů v čase



37

Jazyky pro modely datových struktur dokumentů

- **lineární** – MARC
- **stromové** (hierarchické) – DTD (SGML, HTML), XML Schema
- **síťové** – ERA, UML, RDF, RDFS

38

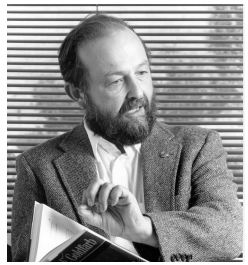
Jazyky pro operace s dokumenty / daty

- **programovací jazyk** – popis operací s dokumenty / daty
  procedurální
  = **Pascal, Java, C++, CSS, XSL** …
- **značkovací (vyznačovací) jazyk** ***(markup language)*** – popis dokumentů / dat
  deklarativní
  = **MARC, SGML, XML, HTML** …
- **modelovací** – vyjádření modelů dokumentů / dat
  = **UML**

39

SGML – Standard Generalized Markup Language

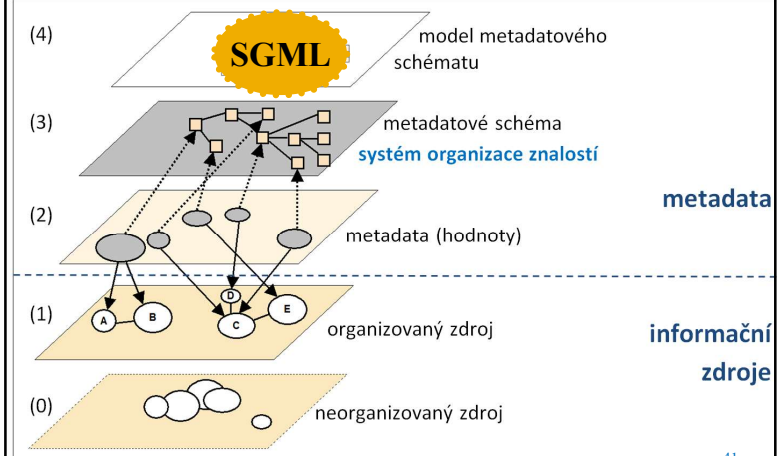
- obecný jazyk určující **syntaxi** značkovacích jazyků (metajazyk)
- základem je stromová struktura prvků dokumentu
- ISO 8879:1986



Charles F. Goldfarb
(26. 11. 1939 – 21. 4. 2023)

40

SGML jako model metadatového schématu pro strukturní metadata



41

41

Proč nás zajímá SGML?

1) HTML – Hypertext Markup Language

2) XML – eXtensible Markup Language

- rozšiřitelný vyznačovací metajazyk využitelný v prostředí Internetu
- podmnožina SGML – zjednodušení některých příliš složitých a vývojem překonaných pravidel
- objektově orientovaný

ÚISK – KSA T07 42

42

Typ dokumentu *(document type)*

⇒ Úkol 12

⇒ Úkol 13

- **Třída dokumentů**, které mají podobné charakteristiky

  dokument `<má vlastnost>`, dokument `<má část>`

  časopis, článek, dopis, faktura…

- **DTD – document type definition**

  slovník, kterým jsou popisovány dokumenty daného typu

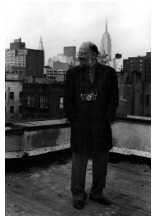
  = *metadatové schéma*

ÚISK – KSA T07 43

43

Typ dokumentu?

Samoobsluha v Kalifornii
Allen Ginsberg
1926 – 1997



Jak jsem na tebe myslel dnes večer, Walte Whitmane, když jsem kráčel postranními uličkami pod stromy, bolela mě hlava a plaše jsem hleděl na měsíc v úplňku.
Utahaný a hladový, chtěl jsem nakoupit obrazy, a tak jsem vešel do neónové samoobsluhy s ovocem a snil o tvých enumeracích!
Jaké broskve a jaké odstíny! Celé rodiny nakupující v noci! Uličky plné manželů! Ženy u avocados, děti v rajčatech! – A ty, Garcío Lorco, co tys tam hledal mezi melouny?

Přeložil Jan Zábrana

ÚISK – KSA T07 44

44

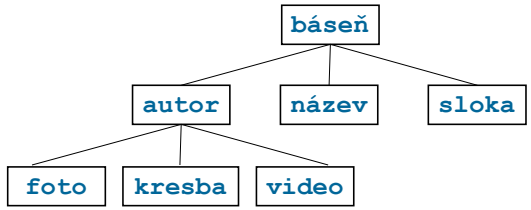
Co definuje DTD

- **elementy** – název, typ obsahu (datový typ), atributy
- **vztahy mezi elementy**
  - hierarchie celek – část
  - pořadí elementů

= strukturní metadata

- **výskyt elementu v dokumentu**
  - povinnost výskytu
  - možnost vícenásobného (opakovaného) výskytu

Příklad grafického znázornění struktury DTD



Příklady notace DTD

- `báseň (název, autor, sloka)`
  báseň tvoří název, jméno autora a sloky (přesně v tomto pořadí)
- `báseň (název & autor & sloka)`
  báseň tvoří název, jméno autora a sloky (v libovolném pořadí)
- `autor (foto | kresba | video)`
  u autora se uvádí buď fotografie, nebo kresba nebo video
- `název?`
  dokument může mít jen 1 název, nemusí mít žádný
- `název (#PCDATA)`
  obsah elementu název je tvořen textem

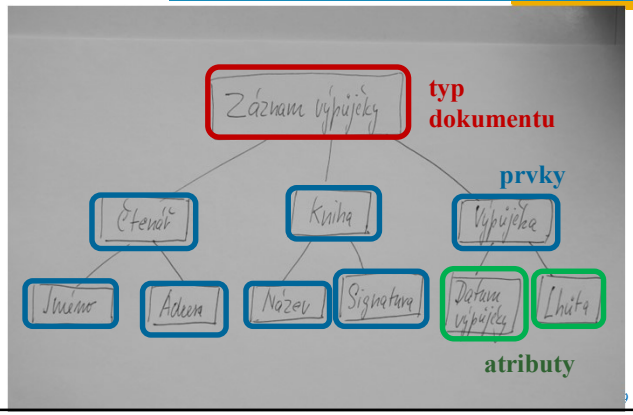
Příklad: DTD pro záznam výpůjčky

⇒ Cvičení 22

Instance dokumentu / zdroje – záznam výpůjčky

Čtenář Vladimír Koutecký z Prahy 4 si dne 14.4.2024 na týden půjčil knihu *Information architecture for the World Wide Web* (signatura F202954).

Analýza struktury dokumentu



49

DTD (definice typu dokumentu) = strukturní metadata

```
ZáznamVýpůjčky
(čtenář, kniha, výpůjčka)
čtenář (jméno, adresa)
kniha (signatura, název)
výpůjčka
datumVýpůjčky
lhůta
```

Převod hierarchické struktury do lineárního zápisu

50

DTD (definice typu dokumentu) = strukturní metadata

```
<! DOCTYPE    ZáznamVýpůjčky [
<! ELEMENT    ZáznamVýpůjčky
(čtenář, kniha, výpůjčka)>
<! ELEMENT    čtenář (jméno, adresa)>
<! ELEMENT    kniha (signatura, název)>
<! ELEMENT    výpůjčka>
<! ATTLIST    výpůjčka
              datumVýpůjčky
              lhůta>
] >
```

Odlišení typu dokumentu, prvků a atributů

51

DTD (definice typu dokumentu) = strukturní metadata

```
<! DOCTYPE    ZáznamVýpůjčky [
<! ELEMENT    ZáznamVýpůjčky
(čtenář, kniha, výpůjčka)>
<! ELEMENT    čtenář (jméno, adresa)>
<! ELEMENT    kniha (signatura, název)>
<! ELEMENT    jméno (#PCDATA)>
<! ELEMENT    adresa (#PCDATA)>
<! ELEMENT    signatura (#PCDATA)>
<! ELEMENT    název (#PCDATA)>
<! ELEMENT    výpůjčka (#PCDATA)>
<! ATTLIST    výpůjčka
              datumVýpůjčky (#CDATA)
              lhůta (#CDATA)>
] >
```

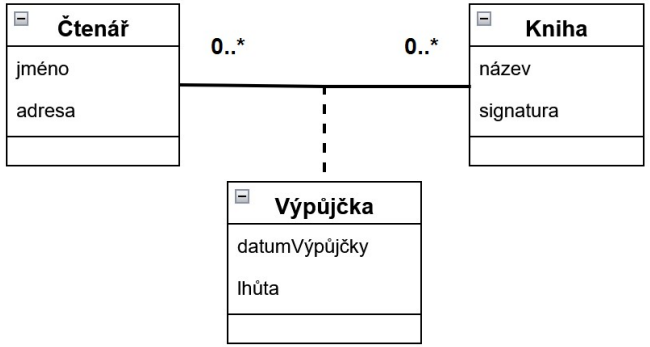
Určení typu obsahu / datových typů pro prvky a atributy

52

Instance SGML dokumentu – text s vyznačením *(markup)*

```
<ZáznamVýpůjčky>
<čtenář><jméno>Vladimír Koutecký</jméno>
<adresa>z Prahy 4</adresa></čtenář>
<výpůjčka datumVýpůjčky="si dne 14.4.2024";
lhůta="na týden">půjčil</výpůjčka>
<kniha>knihu <název>Information architecture for
the World Wide Web</název>
<signatura>(F202954)</signatura></kniha>
</ZáznamVýpůjčky>
```


Model v UML (diagram tříd) grafická notace



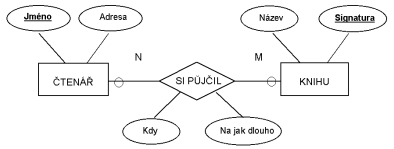
Model v UML textový zápis

```
/* Module:  Kniha.cs
 * Purpose: Definition of the Class Kniha/
public class Kniha{
public string get_název(){return název;}
public void getnázev(string newNázev)
{this.název = newNázev;}
public string get_signatura(){return signatura;}
public void getsignatura(string newSignatura)
{this.signatura = newSignatura;}
   private string název;
   private string signatura;
```

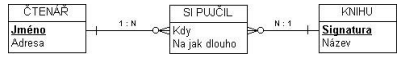
Převod síťové struktury do lineárního zápisu

ERA diagram – varianty notace

Peter Chen



James Martin: Information Engineering (IE)



Merise

Model v ERA
textový zápis

```
E:   ČTENÁŘ (Jméno, Adresa)
     KNIHA (Signatura, Název)

R:   VÝPŮJČKA (datumVýpůjčky, Lhůta)
```

Převod síťové struktury do lineárního zápisu

57

Instance ERA / UML modelu – záznamy v databázi

Čtenář

jméno	adresa
Vladimír Koutecký	Praha 4

Výpůjčka

datumVýpůjčky	lhůta
14. 4. 2024	1 týden

Kniha

název	signatura
Information architecture for the World Wide Web	F202954

58