

Téma 7 – Struktury dokumentů, metadatová schémata a datové struktury

- Typy struktur
- Modely struktur dokumentů/dat
- SGML

ÚISK – KSA T07

1

1

K čemu je klasifikace a systémová analýza při organizaci informací?

- 1) Pojmové modely a ontologie
= modely z pojmů
- 2) Strukturování dokumentů / dat
neboli
informační architektura
= modely dokumentů
... a další a další

ÚISK – KSA T07

2

2

Metafory organizace znalostí – typy struktur

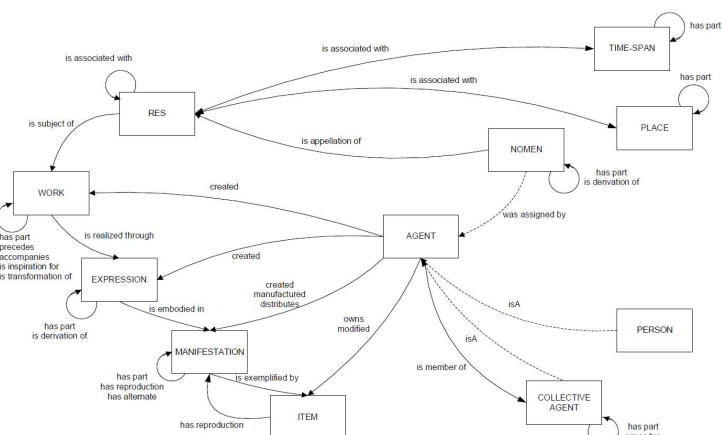
- seznam
- kartotéka
- index, rejstřík
- řetězec
- kruh (cyklus, encyklopedie)
- oddenky (rhizoma)
- strom (hierarchie)
- síť/graf
- tabulka
- vrstvy, úrovně
- stavebnice (fasety, moduly, objektový přístup)

ÚISK – KSA T07

3

3

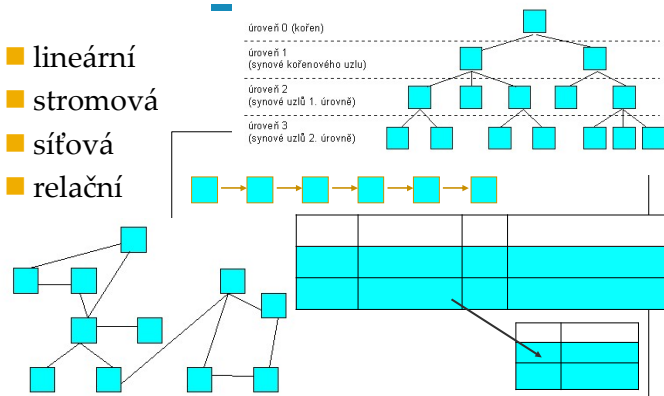
IFLA LRM (Library reference model) – model struktury metadat



4

Základní typy struktur

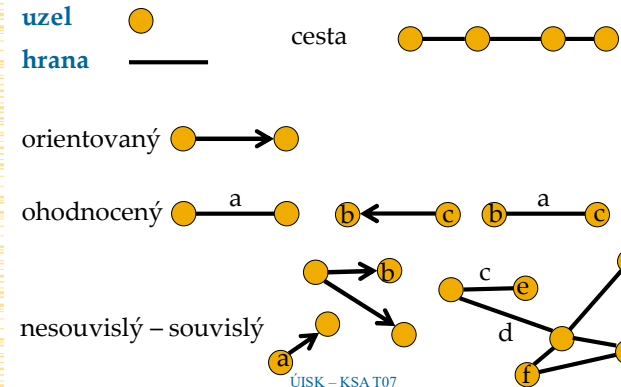
- lineární
- stromová
- síťová
- relační



ÚISK – KSA T07

5

Lineární, stromová a síťová struktura = graf



ÚISK – KSA T07

6

Fakta / obsah

Čtenář4 . na týden „
z knihu World dne
architecture Prahy()
for . the signatura si
půjčil 4 Koutecký
F20295 2025 4 14.
WideWeb Vladimír
Information

Čtenář Vladimír Koutecký z Prahy 4
si dne 14. 4. 2025 na týden půjčil knihu
„Information architecture for the World
Wide Web“ (signatura F202954).

ÚISK – KSA T07

7

Lineární (sekvenční) struktura



Vztah: sekvenční (asymetrická) asociace 1:1

ÚISK – KSA T07

8

Lineární (sekvenční) struktura dokumentu

lineární struktura srozumitelná lidem

Čtenář Vladimír Kouřtecký z Prahy 4 si dne 14. 4. 2025 na týden půjčil knihu „Information architecture for the World Wide Web“ (signatura F202954).

ÚISK – KSA T07

9

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	411352785	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
	Vlan	Pěna dní	3. 8. 2003	Bednář

lineární struktura srozumitelná strojům

ÚISK – KSA T07

10

Lineární (sekvenční) struktura

oblasti užití:

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

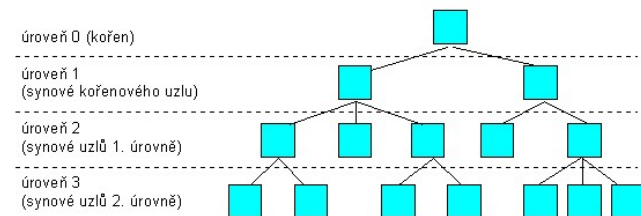
standards:

- ISO 2709 – výměnný formát pro bibliografické záznamy ⇒ MARC

ÚISK – KSA T07

11

Stromová (hierarchická) struktura

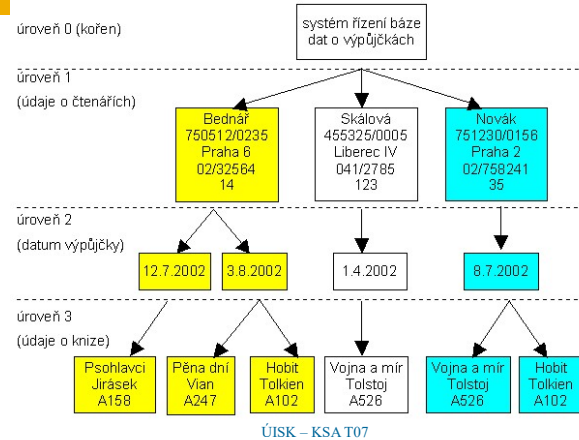


Vztah: monohierarchie 1:N

ÚISK – KSA T07

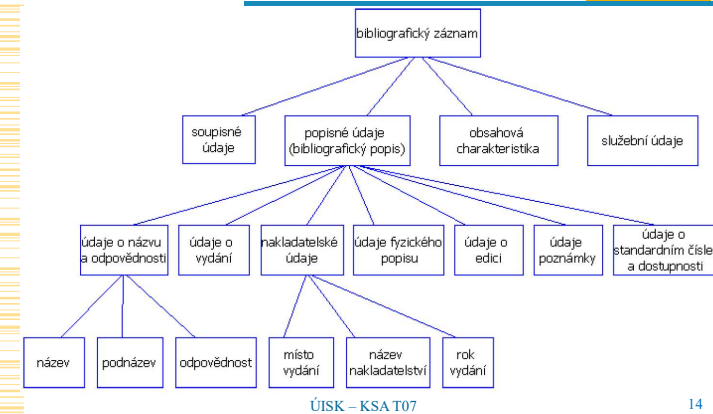
12

Stromová (hierarchická) struktura dat



13

Stromová (hierarchická) struktura bibliografického záznamu (ISBD)



14

Stromová (hierarchická) struktura

klady

- možnost vyjádřit hierarchické vztahy (1 : N)
- rychlé 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

ÚISK – KSA T07

15

15

Stromová (hierarchická) struktura

oblasti užití:

- narativní dokumenty
- indexové soubory
- HTML a XML dokumenty
- objektově orientovaná technologie

standards:

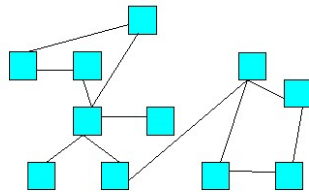
- ISO 8879 – SGML
- XML

ÚISK – KSA T07

16

16

Síťová struktura

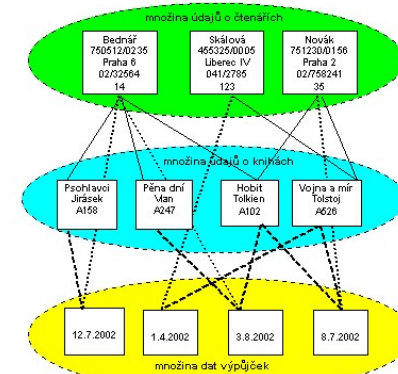


Vztah: asociace, polyhierarchie N:M

ÚISK – KSA T07

17

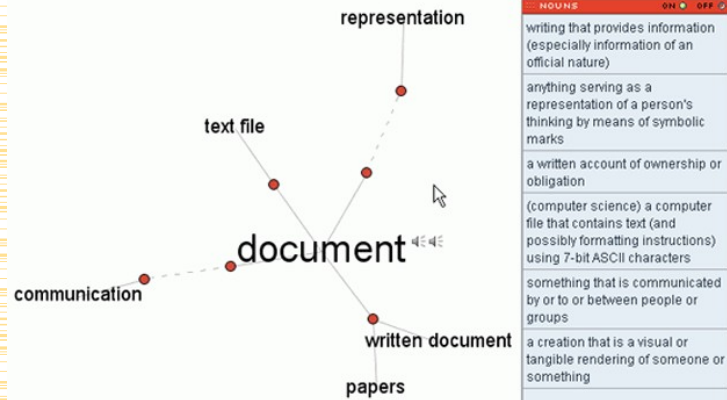
Síťová struktura dat



ÚISK – KSA T07

18

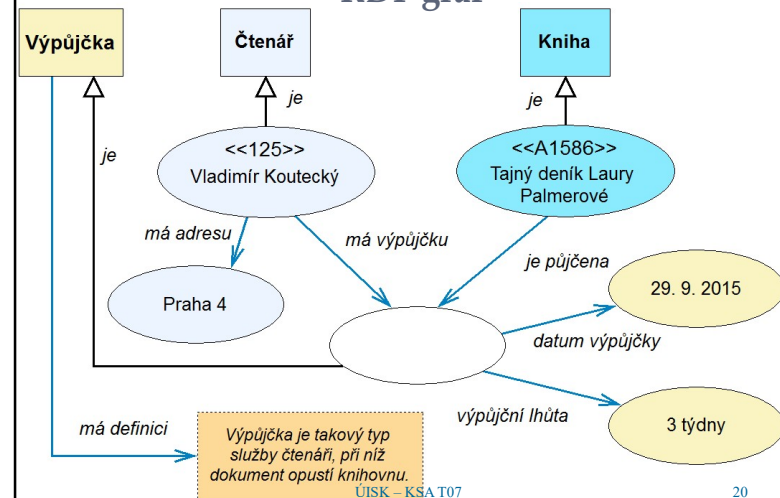
Síťová struktura slovníku WordNet



ÚISK – KSA T07

19

Grafický model obsahu zdroje podle RDF – RDF graf



ÚISK – KSA T07

20

17

18

19

20

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

ÚISK – KSA T07

21

21

Síťová struktura

oblasti užití:

- hypertext, hypermédia
- World Wide Web
- sociální sítě
- znalostní grafy
- propojená data

standards:

- CODASYL
- UML
- RDF

ÚISK – KSA T07

22

22

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	Signatura	Datum	Číslo legitimace
Jirásek	Psohlavci	A158	A526	8.7.2003	35
Tolkien	Hobit	A102	A158	12.7.2003	14
Tolstoj	Vojna a mír	A526	A247	3.8.2003	14
Vian	Pěna dní	A247	A102	3.8.2003	14
			A102	8.7.2003	35
			A526	1.4.2003	123

Vztah: asociace (relace) N:M

ÚISK – KSA T07

23

23

Relační struktura

oblasti užití:

- transakční (relační) databáze

standards:

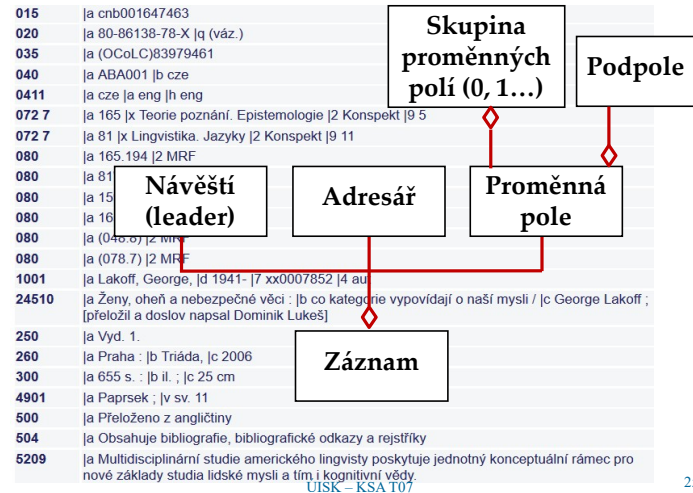
- ISO/IEC 9075 – SQL

ÚISK – KSA T07

24

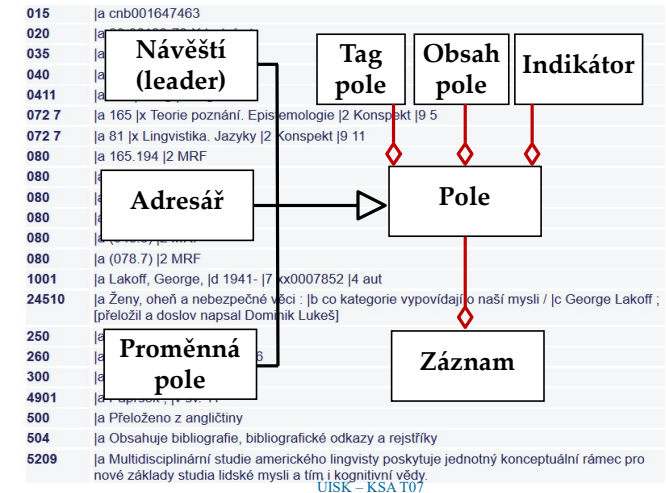
24

MARC21 – jaký typ struktury dat?



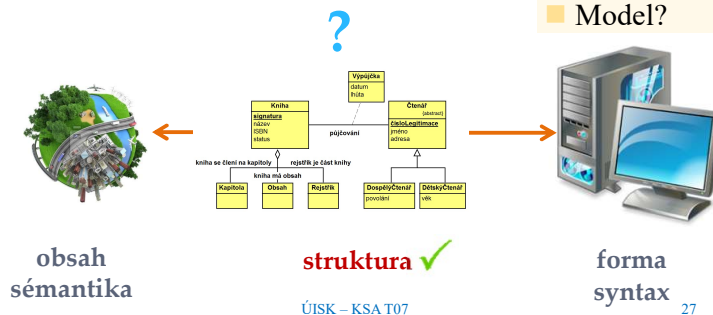
25

MARC21 – jaký typ struktury dat?



26

Sémantická a syntaktická mikroanalýza – modely struktury zdrojů/dokumentů



27

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 aktéry a s funkcemi / případy užití



ÚISK – KSA T07

28

28

29

29

30

30

31

31

32

32

Co definiuje model struktury dokumentu

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

[illegible]

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

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

ÚISK – KSA T07

34

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

ÚISK – KSA T07

35

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		

ÚISK – KSA T07

36

Realita: Standardy pro struktury elektronických dokumentů



ÚISK – KSA T07

37

37

Proč tolik modelů?

Rekapitulace

Co klasifikujeme/identifikujeme/
modelujeme při organizaci znalostí?

Jednotka organizace/identifikace/klasifikace

1) Zaznamenané znalosti

- díla a jejich vyjádření (DOI, VIAF, ISTD, ISAN a V-ISAN, ISWC)
- provedení (manifestace) děl (DOI, ISBN, ISSN, GTIN)
- jednotka (signatura, DOI)
- kolekce a služby (ISCI)
- části / složky / komponenty děl/provedení/jednotek (DOI, SICI)
- metadata – vyhledávací atributy / přístupové body / selekční údaje (OID, URI)

2) Kontext

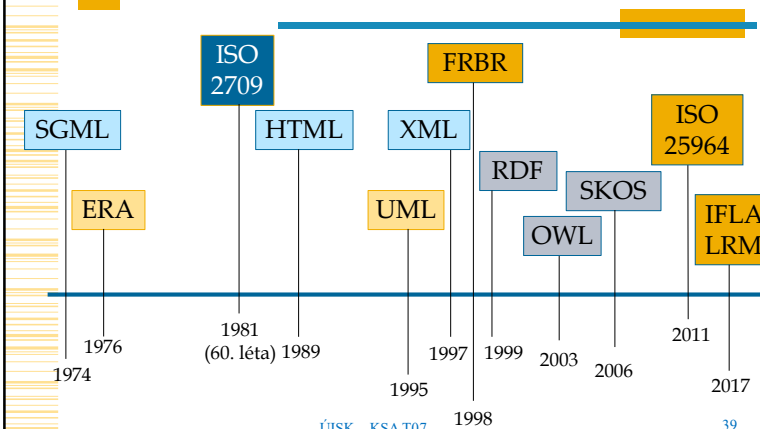
- lidé, např. autoři (IČO, VIAF, ISNI, ORCID)
- organizace (IČ, DIČ, GLN, VIAF, ISNI, ISIL, sigla)

ÚISK – KSA T07

38

38

Modely dokumentů v čase



ÚISK – KSA T07

39

39

Jazyky pro modely datových struktur dokumentů

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

ÚISK – KSA T07

40

40

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

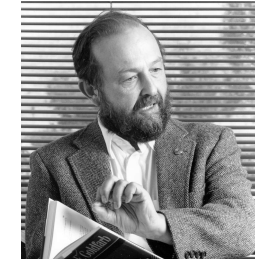
ÚISK – KSA T07

41

41

SGML – Standard Generalized Markup Language

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



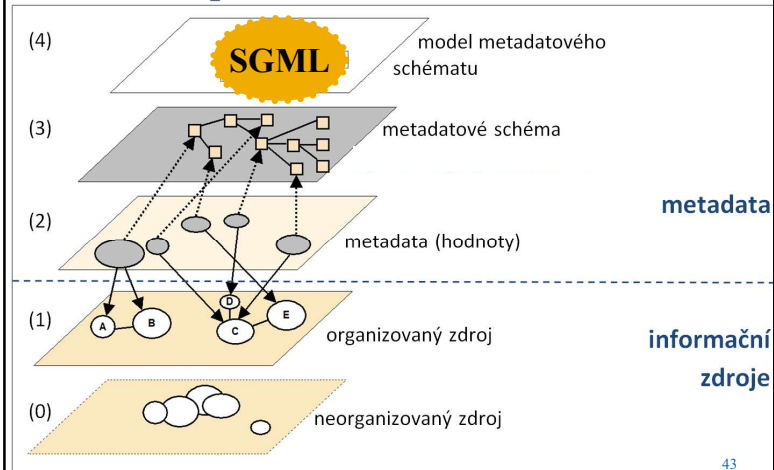
Charles F. Goldfarb
(26. 11. 1939 – 21. 4. 2023)

ÚISK – KSA T07

42

42

SGML jako model metadatového schématu pro strukturní metadata



43

43

Proč nás zajímá SGML?

1) HTML – Hypertext Markup Language

2) XML – eXtensible Markup Language

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



ÚISK – KSA T07

44

44

Typ dokumentu (document type)

⇒ Úkol 12

- **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*

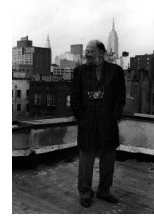
ÚISK – KSA T07

45

45

Typ dokumentu?

Samoobsluha v Kalifornii
Allen Ginsberg
1926 – 1997



Jak jsem na tebe myslel dnes večer, Walte Whitmane, když
jsem kráčet 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

46

46

Co definuje DTD

- **elementy** – název, typ obsahu (datový typ), atributy
- **vztahy mezi elementy**
 - hierarchie celek – část
 - pořadí elementů
- **výskyt elementu v dokumentu**
 - povinnost výskytu
 - možnost vícenásobného (opakovaného) výskytu

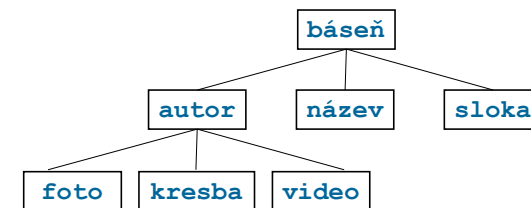
= **strukturní metadata**

ÚISK – KSA T07

47

47

Příklad grafického znázornění struktury DTD



ÚISK – KSA T07

48

48

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

ÚISK – KSA T07

49

49

Příklad: DTD pro záznam výpůjčky

⇒ Cvičení 21

Instance dokumentu / zdroje – záznam výpůjčky

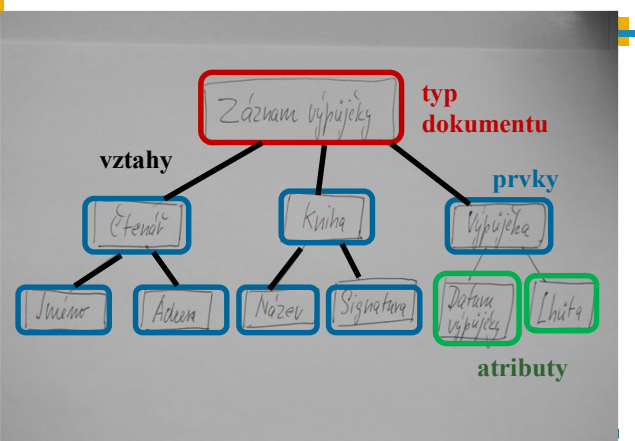
Čtenář Vladimír Koutecký z Prahy 4
si dne 14.4.2025 na týden půjčil
knihu *Information architecture for
the World Wide Web* (signatura
F202954).

ÚISK – KSA T07

50

50

Analýza struktury dokumentu



51

DTD (definice typu dokumentu) = strukturní metadata

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



Převod
hierarchické
struktury do
lineárního zápisu

ÚISK – KSA T07

52

52

DTD (definice typu dokumentu) = strukturní metadata

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

Odlišení typu
dokumentu,
prvků a atributů

ÚISK – KSA T07

53

53

DTD (definice typu dokumentu) = strukturní metadata

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

Určení typu obsahu
/ datových typů
pro prvky a atributy

ÚISK – KSA T07

54

54

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

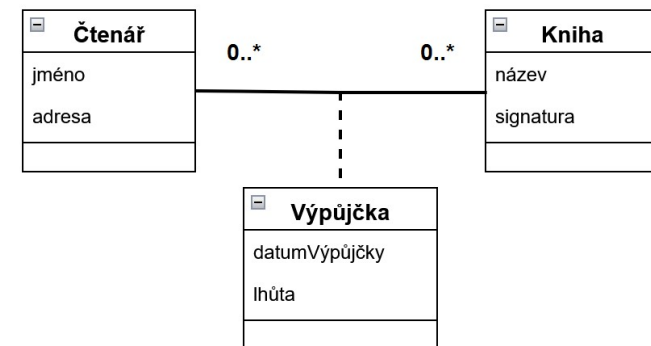
```
<ZáznamVýpůjčky>
<čtenář><jméno>Vladimír Koutecký</jméno>
<adresa>z Prahy 4</adresa></čtenář>
<výpůjčka datumVýpůjčky="si dne 14.4.2025";
lhůta="na týden">půjčil</výpůjčka>
<kniha>knihu <název>Information architecture for
the World Wide Web</název>
<signatura>(F202954)</signatura></kniha>
</ZáznamVýpůjčky>
```

ÚISK – KSA T07

55

55

Model v UML (diagram tříd) grafická notace



ÚISK – KSA T07

56

56

Model v UML textový zápis v programovacím jazyce

```

/* 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

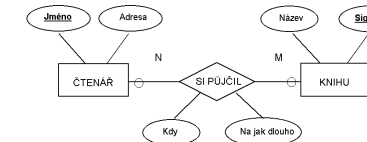
ÚISK – KSA T07

57

57

ERA diagram – varianty notace

Peter Chen



James Martin:
Information
Engineering (IE)



Merise



ÚISK – KSA T07

58

58

Model v ERA textový zápis

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

R: VÝPŮJČKA (datumVýpůjčky, Lhůta)

Převod síťové
struktury do
lineárního zápisu

ÚISK – KSA T07

59

59

Instance ERA / UML modelu – záznamy v databázi

Čtenář	
jméno	adresa
Vladimír Koutecký	Praha 4

Výpůjčka	
datumVýpůjčky	lhůta
14. 4. 2025	1 týden

Kniha	
název	signatura
Information architecture for the World Wide Web	F202954

ÚISK – KSA T07

60

60