



---

# **Metadata a identifikátory**

## **(se zaměřením na WWW zdroje)**

Mgr. Filip Vojtášek

*vojtasek@ikaros.cz*

# Podstata metadat

- = data o datech
- = strukturovaná data, která umožňují interpretovat jiná data (primární data)
- = „...pro počítač srozumitelné informace o webových zdrojích a jiných objektech.“

*(T. Berners-Lee)*

- = data o samotném zdroji či nad jeho rámec
  - Přidaná hodnota k primárním datům určená pro počítače (HTTP) nebo pro člověka (digitální kopie)
  - Metadata a primární data: různý charakter a funkce, ale tvoří logickou jednotku



# Typy (a funkce) metadat

- Popisná → indexace a vyhledávání
  - formální a obsahové znaky zdroje
- Administrativní
  - autorská práva, e-business, e-podpis aj.
- Technická
  - hlavičky HTTP, konfigurace snímacího zařízení, atributy grafických souborů, specifikace hardwarové platformy pro emulaci aj.

# Metadata a katalogizace

- Metadata svou podstatou analogií katalogizačních/bibliografických záznamů
- ALE:
  - Zpracování provádějí často autoři/vydavatelé
  - Předmětem výhradně elektronické zdroje
  - Jednotlivé objekty
  - Přístup pomocí vyhledávacích systémů (search engines, IQ agenti)
  - Žádná standardizace (resp. nelze uplatnit katalogizační pravidla – např. prameny popisu)
  - Volně tvořená klíčová slova x řízené slovníky
  - Přímá vazba metadata → zdroj/objekt

# Značkovací systémy

- Procedurální
  - příkazový charakter
  - interní formátovací nástroj (textové procesory, postskriptové jazyky, HTML + CSS1/2)
  - vizuálně odlišná prezentace digitálních objektů (nadpis, odstavec, tabulka aj.)
  - HTML 4.0 + XML → XHTML 1.0
- Deskriptivní
  - vyjádření obsahově významných objektů pomocí specifické kategorie a přidělené hodnoty (<autor>Petr Novák</autor>)
- Ideální stav: oddělení obou systémů (týž obsah, různá forma podle potřeby)
- GML → SGML → XML

# Syntax a sémantika metadat

- Syntax = pravidla správného utváření metadatové struktury (deklarace elementů, atributů atd.) → DTD (Document Type Definition)
  - HTML vs. XML
  - XML 1.0 (1998, rev. 2000)  
<http://www.w3.org/TR/REC-xml>
  
- Sémantika = pravidla zápisu metadat z obsahového hlediska
  - metadatová schémata (objekty → elementy)
  - XML → RDF (Resource Description Framework)
  - RDF syntax (1999)  
<http://www.w3.org/TR/REC-rdf-syntax/>
  - RDF sémantika (2000)  
<http://www.w3.org/TR/rdf-schema/>

# Uložení metadat

- Metadata součástí zdroje
  - HTML 2.0/3.2: tag `<META name content>` [viz](#)
  - podpora vyhledávacími službami (AltaVista x Excite)
  - HTML + Dublin Core Element Set [viz](#)
  - XML (RDF)
- Metadata v relační databázi (SQL)
  - redakční publikační systém [viz](#)
- Objektově orientovaný přístup (složený digitální dokument – metadata a primární data součástí hierarchického a hypertextového systému organizovaného pomocí SGML „mapového“ souboru)
  - DOBM (variabilní metadatové schéma) [viz](#)

<HTML>

<HEAD>

<META HTTP-EQUIV="Content-Type" CONTENT="text/html; charset=windows-1250">

<META NAME="Author" CONTENT="Filip Vojtášek">

<META NAME="Generator" CONTENT="Mozilla/4.04 [en] (WinNT; I) [Netscape]">

<META NAME="Keywords" CONTENT="metadata, identifikátory, elektronické zdroje">

<META NAME="Description" CONTENT="Prezentace se zabývá obecnými aspekty využití metadat při zpracování elektronických (zejména WWW) zdrojů a dále identifikačními systémy (URL, PURL, URN, DOI a SICI).">

<TITLE>Metadata a identifikátory (se zaměřením na WWW zdroje)</TITLE>

...

</BODY>

</HTML>

<HTML>

<HEAD>

<META HTTP-EQUIV="Content-Type" CONTENT="text/html; charset=windows-1250">

<TITLE>Ikaros c. 8/2000: Virtuální fond rozptýlených vzácných dokumentů: Bachovy autografy zpřístupněny na Internetu</TITLE>

<META NAME="DC.Creator.personalName" CONTENT="Vojtášek, Filip">

<META NAME="DC.Date" SCHEME="ISO8601" CONTENT="2000-09-20">

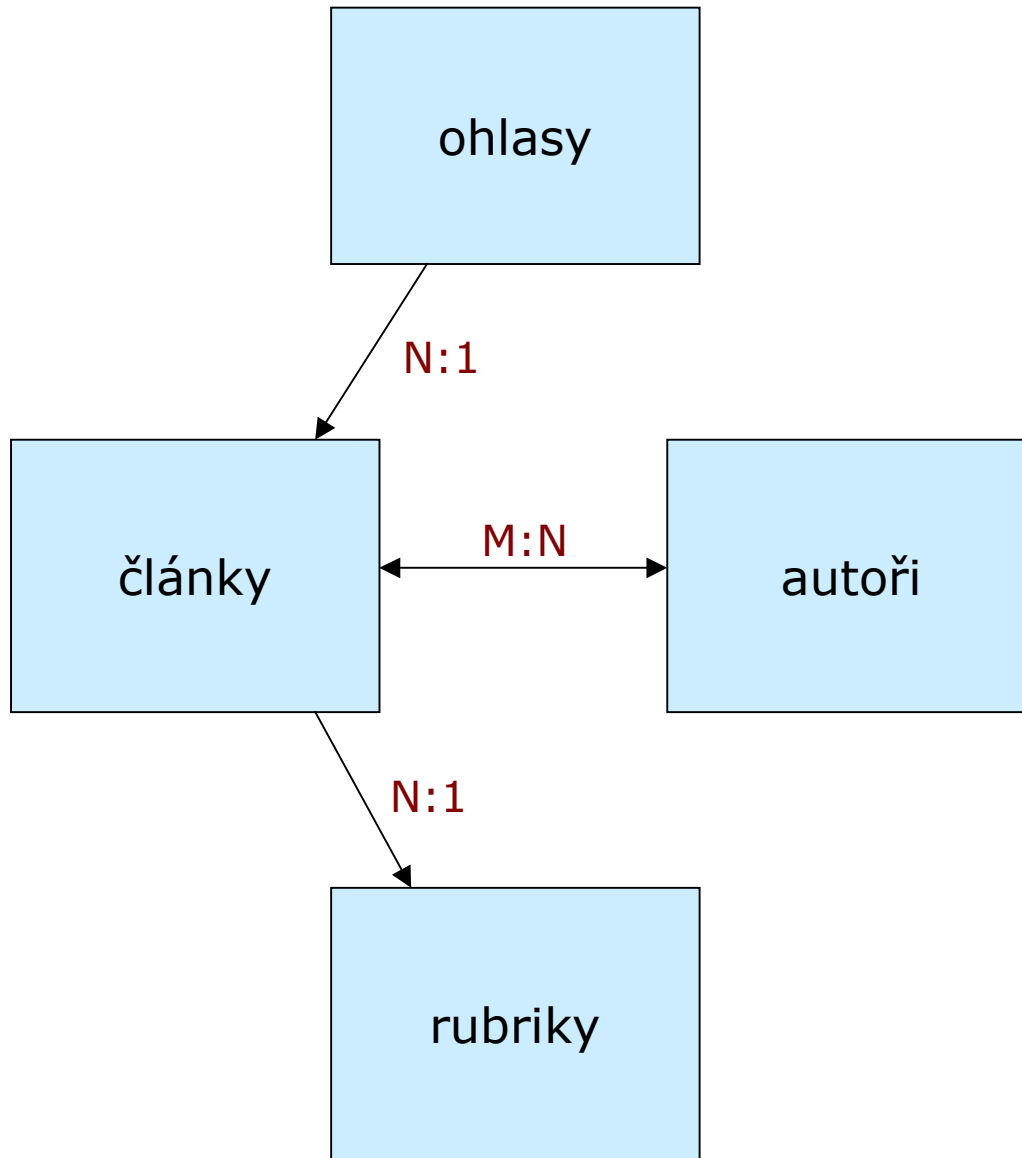
<META NAME="DC.Description" CONTENT="Zpřístupnění digitálních kopií autografů německého skladatele Johanna Sebastiana Bacha (1685-1750)">

<META NAME="DC.Format" CONTENT="text/html">

<META NAME="DC.Identifier" CONTENT="http://ikaros.ff.cuni.cz/ikaros/2000/c08/bach.htm">

...

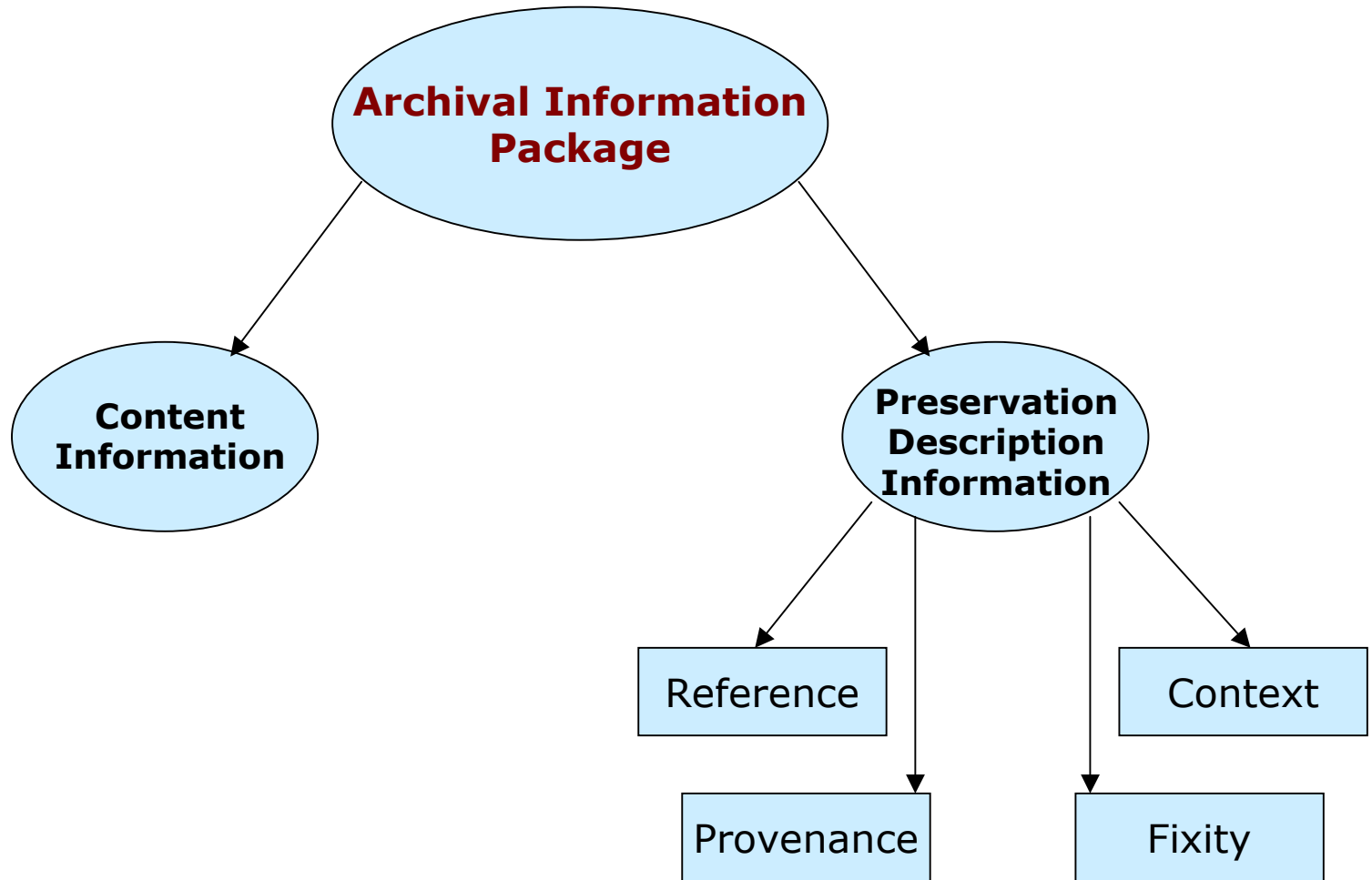
</HEAD>



# Metadata a archivace elektronických zdrojů

- Prostředek k zajištění integrity a autenticity zdroje → dlouhodobé uchování a zpřístupnění (kontext zdroje)
- Podmínka: Co nejmenší svázanost s konkrétním digitálním prostředím (uložení metadat v prostém textu)
- Dosavadní řešení:
  - SGML (→ nutný speciální prohlížeč nebo dynamická konverze do HTML)
  - Obohacení HTML o prvky pro obsahově významné objekty (DOBM)
- Open Archival Information System: funkční model datového toku v digitálním archivu
  - Aplikace: projekty NEDLIB (EU) a Cedars (UK)

# Open Archival Information System



# Identifikátory

- Tradiční publikování: zavedené identifikační systémy ISSN a ISBN (důležité komunikační prostředky na knižním trhu – nakladatelé, knihovny, bibliografické agentury)
- Elektronické publikování: spíše světlo na konci tunelu než řešení na dosah ruky
- Proměnlivost WWW: průměrná životnost těchto zdrojů je 45 dní
- Unikátní a v čase stabilní sekvence znaků (ve standardizované a mezinárodně podporované podobě), nesvázaný s konkrétním aplikačním SW, vztahující se k danému zdroji (či jeho instanci)
- Přidělení identifikátoru: úkon technické, nikoliv administrativní povahy!
- Primární popisný údaj WWW zdrojů (?)
- Další funkce: hypertextové odkazy, citace na dílčí digitální objekty

# Not Found

The requested URL /xsl/N9068.html was not found on this server.

---

*Apache/1.3.14 Server at www.dpawson.co.uk Port 80*

# URL (Uniform Resource Locator)

- Nejrozšířenější „identifikátor“ WWW zdrojů
- Jednoduchý, flexibilní a srozumitelný (často odrážející obsah zdroje)

*http://server/adr1/adr2/soubor*

- Marketingový nástroj (doména II. řádu)
- Zachycuje aktuální místo uložení zdroje (resp. instrukci pro přístup k němu), nikoliv trvale platné označení!
- Příčinou nejsou technologické nedostatky, ale „lidský faktor“:
  - Vydavatel (fyzická/právnícká osoba) ukončí svou činnost
  - Vydavatel zdroj přesune
  - Vydavatel zdroj zcela odstraní
  - Nové za staré
  - Změna struktury serveru (jiná doména)



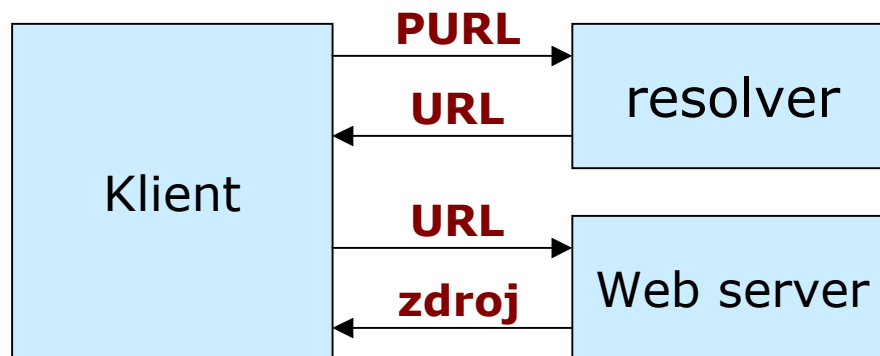
**Zdroj může „fyzicky“ existovat, ale de facto není k dispozici!**

# PURL (Persistent URL)

- \* 1995 (OCLC)
- Dočasné řešení (než se prosadí identifikace pomocí URN)
- PURL se přiřazuje danému URL na základě uživatelem vybraného jména

*http://resolver/adr1/adr2/soubor*

- Nepřímá komunikace klient-server (přesměrování): K zadanému PURL vyhledá server, který spravuje databázi PURL (resolver), právě odpovídající URL, které vrátí zpět klientovi, aby dokončil transakci.



## PURL (Persistent URL)

- Nutná aktualizace databáze při změně URL
- Podpora prohlížeči (stejný mechanismus protokolu HTTP jako u URL)
- Centrální resolver ([purl.oclc.org](http://purl.oclc.org)): obsahuje 565 000 PURL (z toho 4 v doméně .cz)
- Národní a institucionální resolvers
  - National Library of Australia [purl.nla.gov.au](http://purl.nla.gov.au)
  - Dansk BiblioteksCenter [www.purl.dk](http://www.purl.dk)
  - US Government Printing Office [purl.access.gpo.gov](http://purl.access.gpo.gov)

# URN (Uniform Resource Name)

- Vývoj: pracovní skupina IETF
- Perspektivní náhrada URL: jednoznačná identifikace zdroje nezávislá na jeho uložení
- Princip resolvingu
- Aplikace bibliografických identifikátorů jako tzv. jmenných prostorů (ISSN, ISBN, SICI, NBN)

*urn: <NID> ":" <NSS>*

- Syntax: RFC 2141 (1997)
- Plug-in URN:ISSN (0.3beta) [viz urn.issn.org](#)



# DOI (Digital Object Identifier)

- \* 1997 (Association of American Publishers a Corporation for National Research Initiatives)
- Od 1998 spravuje International DOI Federation
- Cíl: efektivnější ochrana majetkových autorských práv
- Přesměrování na server vlastníka, který rozhoduje co a za jakých podmínek zpřístupněno (bibliografický záznam, abstrakt, plný text)

*<http://dx.doi.org/10.nakIID/sufix>*

- Využití: komerční poskytování informačních služeb - elektronické verze odborných časopisů (Academic Press, Blackwell Science, Elsevier Science, Institute for Scientific Information, John Wiley & Sons, Springer Verlag aj.) a elektronické knihy
- Agregátor CrossRef [www.crossref.org](http://www.crossref.org) (71 nakladatelů, 3800 titulů, 3 milióny článků)

# SICI (Serial Item and Contribution Identifier)

- Norma ANSI/NISO Z39.56 (1991, revize 1996)
- Určen pro tištěné a elektronické seriály (úroveň: titul – číslo – článek – část článku)  
*Viz Wiley InterScience*
- Integrace s existujícími standardy pro účely analytického zpracování:
  - Extenze ISSN
  - Sufix DOI
  - Jmenný prostor URN
- Další využití: citace, automatizovaná akvizice (transakční systém EDI)
- SICI generátor  
<http://www.ep.cs.nott.ac.uk/~sgp/sicisend.html>