Web a sémantický web (co je web)

...

ISKM89 Organizace dat - sémantický web | podzim 2023

Zuzana Nevěřilová | Centrum zpracování přirozeného jazyka

World Wide Web

In 1989, Sir Tim Berners-Lee invented the World Wide Web (see the original proposal). He coined the term "World Wide Web," wrote the first World Wide Web server, "httpd," and the first client program (a browser and editor), "WorldWideWeb," in October 1990.

He wrote the first version of the "HyperText Markup Language" (HTML), the document formatting language with the capability for hypertext links that became the primary publishing format for the Web. His initial specifications for URIs, HTTP, and HTML were refined and discussed in larger circles as Web technology spread.

https://www.w3.org/about/history/

World Wide Web - komponenty

server: httpd

klient: WorldWideWeb

jazyk: HTML

protokol: HTTP

identifikátory: URI

navigace: hypertext

World Wide Web - co znamenají jednotlivé komponenty

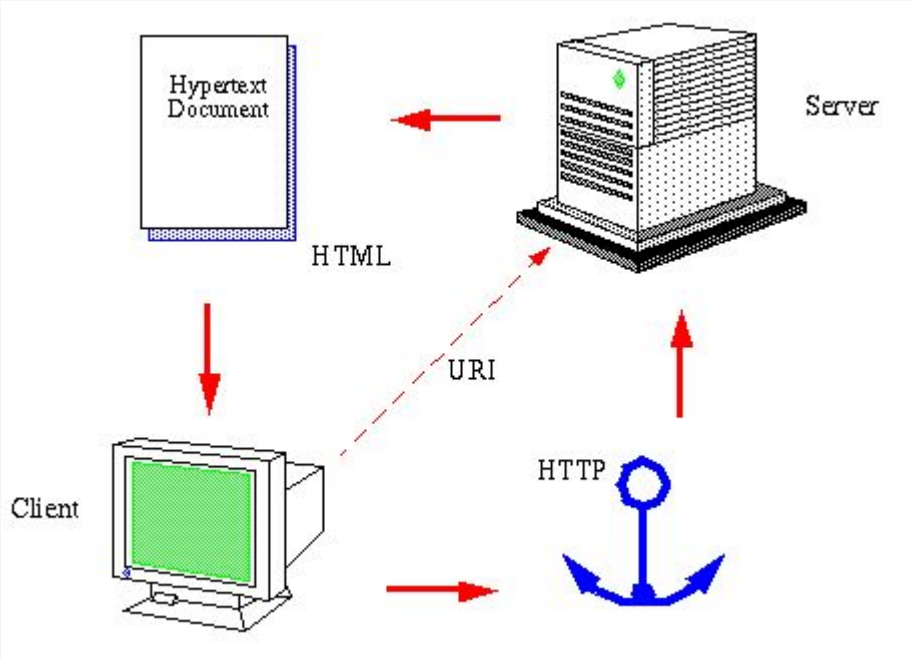
server odpovídá na požadavky klientu

klient zobrazuje objekty transparentně bez ohledu na schéma (access schemes)

HTML obsahuje instrukce, jak má klient stránku zobrazit

HTTP - bezstavový protokol pro výměnu zpráv obsahující schéma požadavku klientu

URI - identifikátor objektu (zadaný explicitně nebo hyperlink)



https://www.w3.org/People/Frystyk/thesis/WWW.html

Příklad zobrazení webové stránky v HTML

URI: https://example.org/index.html
na serveru je soubor `index.html`

```
<!DOCTYPE html>
<html>
  <body>
    <h1>My First Heading</h1>
    <p>My first paragraph.</p>
  </body>
</html>
```

odpověď klientu uživateli

Uniform Resource Identifier

Idea: každý objekt má svůj světově unikátní a univerzální identifikátor

Uniformní - bez ohledu na typ zdroje, syntax URI vypadá stejně

Zdroj (Resource) - cokoliv, co lze odkázat pomocí URI (nejen objekt, ale i kolekce, služba apod)

- ne všechno, co má URI, musí být dostupné na Internetu (lidé, knihy atd.)
- také abstraktní koncepty mohou mít URI (např. operace sčítání, relace “rodič”)

Identifikátor - rozdílné věci mají rozdílné identifikátory

https://datatracker.ietf.org/doc/html/rfc3986

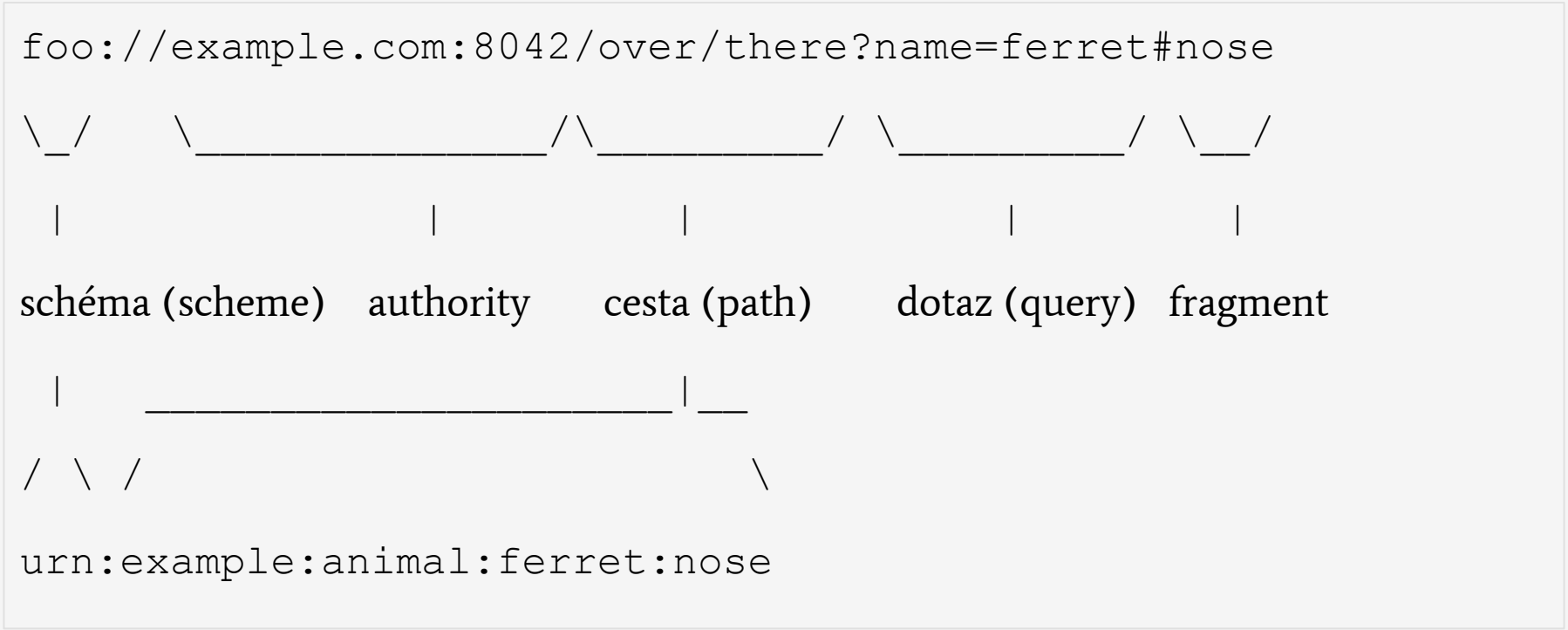
Uniform Resource Identifier - syntax

```
URI = scheme ":" hier-part [ "?" query ] [ "#" fragment ]

      hier-part = "//" authority path-abempty

                / path-absolute

                / path-rootless

                / path-empty
```

":" hierarchie ke specifičtějšímu (zleva-doprava)

"/" stromová struktura uspořádání dokumentů

Uniform Resource Identifier - syntax



Uniform Resource Identifier - další podmínky

`foo://example.com:8042/over/there?name=ferret#nose`

schéma (scheme), authority, cesta (path), dotaz (query), fragment

- používají alfanumerické znaky z US-ASCII sady (písmena, čísla)
- další nevyhrazené znaky jsou `- _ . ~`
- vyhrazené znaky jsou `! * ' ( ) ; : @ & = + $ , / ? % # [ ]`
- jiné znaky se musejí kódovat (escaping)

URI a IRI

- alfanumerické znaky z US-ASCII sady (písmena, čísla)
- další nevyhrazené znaky jsou `- _ . ~`
- vyhrazené znaky jsou `! * ' ( ) ; : @ & = + $ , / ? % # [ ]`
- jiné znaky se musejí kódovat (escaping)

Nověji (2005) definovaný standard je IRI (Internationalized Resource Identifier).

IRI a URI jsou kompatibilní, pokud je text v Unicode kódovaný jako UTF-8 pomocí znaku `%`

https://en.wiktionary.org/wiki/Ῥόδος je totéž co
https://en.wiktionary.org/wiki/%E1%BF%AC%CF%8C%CE%B4%CE%BF%CF%82

Písmeno Ῥ (ró) je v Unicode kódované jako `1FEC`
https://www.compart.com/en/unicode/U+1FEC

URI, URL, URN

Klasické pojetí (do pol. 90. let 20. stol.)

Identifikátor (URI) obsahuje buď umístění (Locator), nebo název (Name).
URI je buď URL, nebo URN.

Moderní pojetí

URI obsahuje schéma (např. `http` nebo `urn`) a může obsahovat jmenný prostor (namespace)

`urn:isbn:n-nn-nnnnnn-n` je jmenný prostor ISBN ve schématu URN

URL zůstalo neformálním pojmenováním URI obsahujícím lokaci (síťovou)

Schémata URI

Registrovaná schémata (IANA)

např. http, https, ftp, ldap, mailto - jsou stálá (Permanent)

ssh, spotify, ms-appinstaller - jsou prozatímní (Provisional)

http://www.iana.org/assignments/uri-schemes/uri-schemes.xhtml

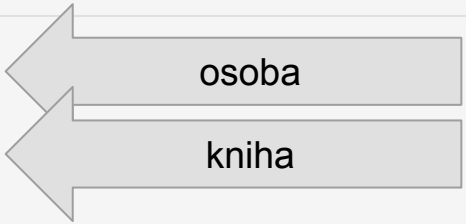
Neregistrovaná a privátní schémata

Užívaná specifickou aplikací, např. streamovací službou.

Nevíme o nich.

Jmenné prostory (namespaces)

```
<title>Dr.</title> <name>Jones</name>

<title>Harry Potter</title>
```



Jmenný prostor určuje kontext objektů a jejich vlastností.

```
xmlns:pn = http://person.com
xmlns:bk = http://book.com

<pn:title>Dr.</pn:title> <pn:name>Jones</pn:name>

<bk:title>Harry Potter</bk:title>
```

World Wide Web - standardy

V roce 1994 vzniklo konsorcium W3C (MIT + CERN, podpora DARPA a EK), následovaly další instituce, které se střídaly v hostování W3C.

W3C vydalo doporučení (recommendations) pro komponenty webu (HTTP, URI, HTML, ...)

Některá doporučení se transformovala na standardy. Standardy jsou popsány pomocí RFC (Request for Comment), např. RFC2616 popisuje HTTP/1.1 (https://datatracker.ietf.org/doc/html/rfc2616)

Standardy, které se vyvíjejí, spravuje WHATWG (Web Hypertext Application Technology Working Group, 2004), např. URL: https://url.spec.whatwg.org/

W3C a WHATWG spolupracují: https://www.w3.org/2019/04/WHATWG-W3C-MOU.html