23. Internet a jeho využití

(1) Účel, vlastnosti, generace

1.1 Co je to Internet

Internet je mezinárodní, **silně decentralizovaná**, robustní a veřejně přístupná **síť přenášející data přepojováním paketů** na **bázi IP** protokolu.

Účelem je poskytování služeb založených na výměně zpráv a sdílení prostředků.

Principem je rozdělení zprávy (např. http požadavku) na samostatně přenášené elementární části (pakety).

Vlastnosti uzlů

- Každé dva uzly v Internetu jsou si navzájem rovnocenné (**peer-to-peer**)
- Každý uzel je **autorita pro vytváření, předávání a přijímání zpráv**

Přepojování paketů (Packet switching)

Cesta paketů k cíli není podstatná, protože **každý paket obsahuje všechny informace potřebné k doručení** a různé pakety patřící k jednomu spojení (jedné *zprávě*), tak mohou být **přeneseny různými cestami**, což umožňuje aby se paket dostal k cíli i v případě, že je některá část sítě nefunkční.

1.2 Historie

- 60. léta – IPTO (odnož organizace ARPA[1]) zkoumá možnosti robustního propojení armádních radarů a zajištění kontinuity státní zprávy po nuklerním útoku pomocí technologie přepojování paketů (RAND Corporation)
- Roku 1969 se po velkém usilí spojily první dva uzly – UCLA (Kalifornská univerita) a SRI (Stanford Research Institute), byl použit přenosový protokol byl NCP
- Ukázková síť posloužila jako důkaz funkčnosti a začaly se připojovat další státy – Anglie, Kanada, Hong Kong a Austrálie
- Tehdejším primárním cílem bylo především sdílení prostředků (superpočítače), ale rychle se začalo prosazovat i použití pro vzdálenou komunikaci
- ARPANET rychle rostl a v roce 1972 měl už 37 uzlů, k 1.1.1983 se protokol ARPANETu změnil z NCP na TCP/IP
- Stavba nejdříve 56 kbps a následně 1,5 Mbps páteřní sítě
- Roku 1988 umožněn vstup soukromého sektoru

1.3 Generace webu

- První stupeň byly statické stránky
- Druhým stupněm je současný dynamický web (tzv. Web 2.0 s technologiemi jako SVG nebo Ajax)
- Společné pro oba stupně je zaměření obsahu webu na lidského čtenáře, místo na stroj

1.4 Koncepce adresace

Každý uzel Internetu (a obecně každý uzel v síti používající protokoly TCP/IP) má dva identifikátory – MAC adresu a IP adresu.

MAC adresa

- Unikátní 48 bitové číslo napevno přiřazené každé vyrobené síťové kartě
- Jednoznačně identifikuje kartu samotnou i jejího výrobce
- Neobsahuje informaci o tom kde (v jaké síti) se karta fyzicky nachází

1) **ARPA**, *Advanced Research Projects Agency*. Americká agentura založená v únoru 1958, určená k získání technologického náskoku před Sovětkým svazem po vypuštění družice Sputnik.

IP adresa

– Spolu s **maskou** vyjadřuje, k jaké síti je uzel (síťová karta) připojen
– Slouží jako lokátor (aby jeden uzel našel druhý), ale není míněna jako unikátní identifikátor, protože je často přidělována dynamicky (DHCP), případně za jednou adresou může být více uzlů (NAT)
– Správu adresového prostoru zajišťuje IANA, která ji deleguje na regionální internetové registrátory (např. RIPE NCC pro Evropu nebo ARIN pro Severní Ameriku)
– Existují dvě verze – IPv4 *(RFC 791)* a IPv6 *(RFC 1883)*

IPv4

– 32 bitové číslo, obvykle reprezentována v decimálním tvaru s maskou za lomítkem (např. 192.168.10.110/24 je adresa v síti 192.168.10.0, číslo masky vyjadřuje kolik bitů z adresy je adresa uzlu a kolik sítě)
– původně byl prostor dělen na třídy A, B, C podle velikosti (16M, 64K a 256), což vedlo k jeho nehospodárnému využití a proto bylo v roce 1993 zavedeno beztřídní směrování (CIDR)
– maska určuje do jaké sítě daná IP adresa patří,čímž je umožněno „jemnější“ dělení adresového prostoru
– speciální IP adresy
  - 127.0.0.1 (loopback)
  - 224.0.0.0/4 (broadcast)
  - 10.0.0.0/8, 172.16.0.0/16 a 192.168.0.0/16 (pro lokální sítě)

IPv6

  - 128 bitové číslo, adresa reprezentována v hexadecimálním tvaru
  - např. `2001:0DB8:85A3:08D3:1319:8A2E:0370:7334`

1.5 Koncepce DNS

– Hierarchický systém distribuovaných, decentralizovaných databází doménových jmen, sloužící lidem k snažší orientaci v Internetu
– Provoz zabezpečuje množství navzájem zrcadlených, kořenových serverů obsahujících autoritativní informace o doménách nejvyšší úrovně (TLD)

Domény

– Prostor doménových jmen je tvořen stromem, administrativně děleným na zóny, umožňující snadnou delegaci práv
– Každá větev stromu obsahuje informace o sobě podřízené části doménového jména (menších větví), jehož správu nepředala na nižší úroveň
– Kořenem stromu je tzv. kořenová doména, která se zapisuje tečkou

Příklad DNS relace

1. Klient má požadavek na přeložení doménového jména linux.slashdot.org
2. Resolver pošle dotaz lokálnímu jmennému serveru (LNS) a očekává jednoznačnou (autoritativní) odpověď
3. Lokální nameserver pošle některému root serveru (jehož adresu má na disku předem zapsánu v souboru root.hint) dotaz na seznam nameserverů domény .org
4. LNS odešle dotaz na jméno *slashdot* některému nameserveru domény .org, odpověď obsahuje seznam nameserverů pro doménu slashdot.org
5. LNS odešle dotaz na jméno *linux* některému nameserveru v doméně slashdot, odpověď je autoritativní a obsahuje IP adresu serveru poskytujícího obsah v doméně linux.slashdot.org

Server spravující doménu .org tedy obsahuje jen informace o doménovém jméně slashdot ve svém prostoru, ale o jméně linux ve jmenném prostoru (doméně) slashdot (dohromady linux.slashdot.org) už informace nemá, protože tuto správu delegoval a odkáže klienta na nameserver uvedený v DNS záznamu domény slashdot.org.

(2) Identifikace v Internetu (URI)

2.1 Vlastnosti

- Dovoluje určit rozdílné typy objektů, ke kterým se přistupuje odlišným způsobem
- Neexistuje žádné omezení pro typ identifikovaného objektu – může to být elektronický dokument, obraz, zvuková nahrávka, informační zdroj nebo třeba kniha v knihovně
- Objekt pouze identifikuje, neříká nic o jeho dostupnosti
- **URI reference** se skládá ze **schématu** a na něm závislé **hierarchické části** (autorita, cesta), **dotazu** a **fragmentu**
  - URI reference je **relativní**, pokud obsahuje jen některé části (např. jen cestu a dotaz), je navíc nutná báze
  - URI je vždy plnohodnotný řetězec ukazující na zdroj, takže pokud je URI reference relativní, URI se získá až kombinací s bází

2.2 Způsob zápisu a skladba

http://login:pass@slashdot.org:80/content/article.psp?page=3&rubric=31#comments

Scheme (schéma)

- Reprezentuje určitou síťovou službu
- Určuje syntaxi a sémantiku identifikátoru URI
- Nedefinuje přístupový protokol
- Přiděluje IANA

Authority (autorita)

- Reprezentuje jmenný prostor, kde se zdroj nachází
- Jmenný prostor může být vyjádřen registrovaným jménem nebo IP adresou (v různých tvarech)
- Pole začíná `//` a končí `/` nebo `?` nebo `#`
- Tři části:
  - userinfo – musí být ukončeno znakem `@` *(nepovinné)*
  - host – IP adresa nebo jméno
  - port – odděluje se dvojtečkou *(nepovinné)*

Path (cesta)

- Pole začíná znakem `/` a končí znaky `?` nebo `#`
- Hierarchická sekvence segmentů oddělená lomítky
- relativní nebo absolutní
- Relativní cesta může obsahovat znaky `.` a `..`
- Při relativní cestě musí být definována báze

Query (dotaz)

- Nemá hierarchickou strukturu
- Pole začíná znakem `?` a končí `#`
- Struktura KLÍČ = HODNOTA
- Obvykle se realizuje dotazem do databáze

Fragment

- Nepřímo identifikuje sekundární zdroj
- Začíná znakem `#` a končí koncem URI
- Může představovat část primárního zdroje nebo jiný způsob prezentace primárního zdroje
- Typický příklad použití je u http, kde označuje konkrétní místo v dokumentu

2.3 Rozdíl mezi URN a URL

URN a URL jsou podmnožiny URI. Základní rozdíl spočívá v tom, že URN určuje pouze identitu (název), kdežto URL určuje, kde se daný zdroj nachází.

Typickým příkladem URN je ISBN systém – unikátně identifikuje každou knížku, ale neříká nic o jejím umístění. Naproti tomu http://www.slashdot.org/ identifikuje zdroj a naznačuje, že ho lze získat protokolem http ze serveru na adrese www.slashdot.org.

V technických publikacích a zvláště standardech od organizací IETF a W3C se termín URL nepoužívá, protože v praxi není třeba rozlišovat mezi URI a URL.

2.4 IRI (Internationalized Resource Identifier)

Internationalized Resource Identifier (IRI) je nadmnožinou URI (Uniform Resource Identifier), umožňující odkazovat na zdroj pomocí znaků v kódování UTF-8.

(3) Sémantický Web

3.1 Úvod

Účelem je učinit data dostupná na webu srozumitelná strojům přidáním metadat specifikujících sémantický obsah objektu.

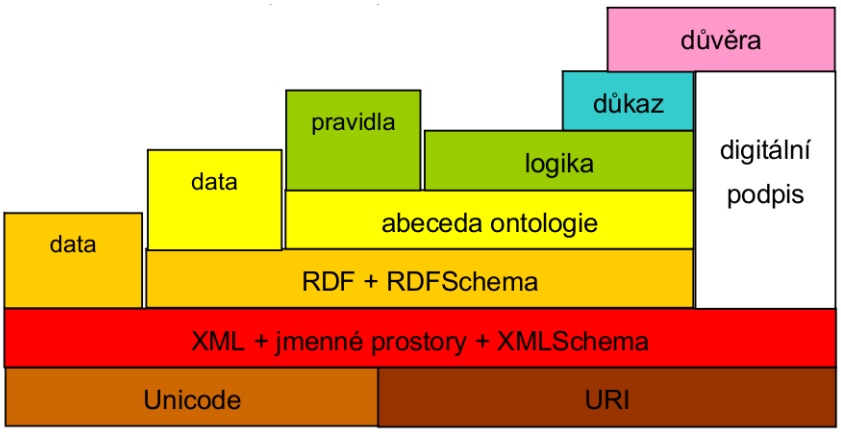
Principem je vývoj standardů, které umožní automatizovaným systémům najít spojitosti mezi distribuovanými daty čímž se vytvoří síť *znalostí*, namísto současné sítě *informací*. Pro fungování sémantického webu je nutné, aby počítače měli přístup ke strukturovanému souboru informací a odvozovacích pravidel, které mohou použít k dedukci.

Sémantický web je soubor technologií, designových pravidel a filozofie. Některé součásti jsou v současnosti pouze výhledy do budoucna, čekající na svoji praktickou realizaci (ontologie a softwaroví agenti), jiné jsou již dnes široce rozšířené (XML).

První veřejná zmínka o sémantickém webu je v časopise Scientific American z května 2001 v článku *The Semantic Web*. Cílem je univerzální médium pro výměnu znalostí, dat a informací.

Problémem současného webu je orientace na lidského čtenáře, počítač sice může snadno určit co je nadpis, co odkaz na další stránku, ale nemá sémantické vnímání k určení, že toto je web pražské školy se čtyřmi, technicky zaměřenými obory a adresou V Úžlabině 320, Praha 10. Takové informace lze získat jen robotem naprogramovaným speciálně pro tento web a v okamžiku, kdy se webu jen trochu změní struktura, tak je nutné robota upravit.

Příklad použití sémantického webu: Chceme zjistit ceny všech lednic s objemem menším než 100 litrů, za méně než 15 000Kč v obchodech v okruhu 20km, které mají ve středu otevřeno do 18:00. Dnes jsou pro takový úkol nutní vyhledávací roboti specificky naprogramovaní pro každý prohledávaný obchod (zdroj). Sémantický web umožňuje skrze své standardy (RDF, N3, SPARQL a další) přidat k informacím relevantní metadata umožňující snažší strojní zpracování a větší důvěru ve výsledek.



Obr. 1: Schéma struktury sémantického webu

3.2 Technologie

RDF (Resource Description Framework, Systém popisu zdrojů)

– Navržen W3C jako formát pro reprezentaci metadat (metadata model)
– Nyní je to abstraktní model pro modelování (reprezentaci) informací skrze různé metody serializace[1]
– Považován W3C konsorciem za další evoluční stupeň[2] webu
– Jednoduchost předurčuje použití i v jiných oblastech než sémantický web, například expertní systémy
– Založen na výrocích o zdrojích ve tvaru **subject-predicate-object** *(předmět-výrok-objekt)*
  - *Předmět* je zdroj
  - *Predikát (výrok, sloveso)* je rys nebo vlastnost, vyjadřuje vztah mezi *předmětem* a *objektem*
  - *Objekt* specifikuje pro nějaký předmět hodnotu vlastnosti
  - Předmět, výrok i objekt jsou definovány pomocí URI, nový výrok vzniká uvedením svého URI kdekoliv na webu
  - Dohromady tvoří tyto tři prvky statement (výrok)
– **Příklad:** Výrok „Auto má žlutou barvu“ zapíšeme jako trojici řetězců. Předmět je „auto“, výrok „má barvu“ a objekt je „žlutá“. V N3 by výrok mohl vypadat takto: „`<#auto> <#barva> <#žlutá> .`“.

XML (eXtensible Markup Language, Rozšiřitelný značkovací jazyk)

– Vyvinut a standardizován W3C v rámci RDF
– Často používán jako synonymum pro RDF
– Nejčastější forma serializace RDF (=> XML RDF)
– Umožňuje sice autorovi definovat vlastní tagy, které ale nic neříkají o sémantice – ta vzniká až spojením s ontologií
– Příklad v notaci XML RDF

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
 xmlns:dc="http://purl.org/dc/elements/1.1/"
 xmlns:foaf="http://xmlns.com/0.1/foaf/" >
 <rdf:Description rdf:about="">
 <dc:creator rdf:parseType="Resource">
 <foaf:name>Petr Praus</foaf:name>
 </dc:creator>
 <dc:title>Semantic web in a nutshell</dc:title>
 </rdf:Description>
</rdf:RDF>
```

Tento kus kódu říká, že název popisovaného objektu (např. článku) je „Semantic web in a nutshell“ a jméno autora je „Petr Praus“.

První řádek definuje DTD (Definice typu dokumentu), druhý a třetí jmenné prostory. Čtvrtý otevírá „popisnou část“, pátý jmenný prostor s názvem *dc* a upravuje jeho atribut *name*, šestý definuje jméno autora pomocí atributu *name ve* vnořeném jmenném prostoru foaf. Osmý řádek definuje název popisovaného objektu pomocí atributu *title*.

N3 (Notation3)

– Uveden W3C konsorciem jako další metoda serializace RDF
– Snáze čitelný než XML RDF zápis
– Lze strojově převádět z/do XML RDF
– Příklad v notaci N3

```
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix foaf: <http://xmlns.com/0.1/foaf/> .
<> dc:title
 "Semantic web in a nutshell".
<> dc:foaf:name
```

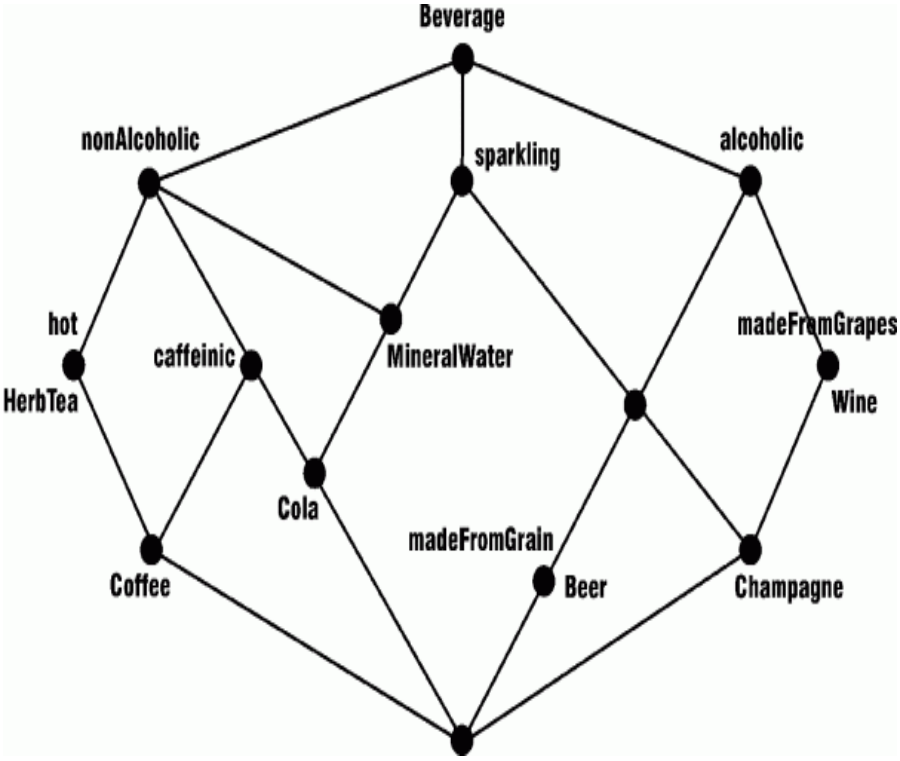
1) Serializace je proces uložení objektu na úložné médium. Serializací je například uložení objektu (třídy) do binárního souboru pomocí pickle v Pythonu, zápis abstraktního RDF modelu v XML formátu nebo uložení dokumentu ve Wordu do formátu doc.
2) Prvním stupněm byly statické stránky, druhým je současný dynamický web (tzv. Web 2.0 s technologiemi jako SVG nebo Ajax). Společné pro oba stupně je zaměření obsahu webu na lidského čtenáře.

```
"Petr Praus".
```

Příklad vyjadřuje to samé, co příklad výše, ale podstatně čitelnější a jednodušší formou.

3.3 Ontologie

Předmětem ontologie je zkoumat kategorie věcí (objektů), které existují ve zkoumaném oboru. Produktem takového zkoumání je ontologie – katalog objektů, které existují ve zkoumané oblasti a vztahů mezi nimi. Existují dva druhy vztahů – hyponymické[1] a meronymické[2]. Nejznámější ontologií je WordNet[3]. Seskupuje *literály* (slova nebo skupiny slov) do *synsetů* (skupin synonym), které jsou provázány výše zmíněnými vazbami. Dalším příkladem je lékařská ontologie UMLS[4] nebo evropský projekt EuroWordNet (ukončen v červnu 1999). Pro skutečný sémantický web však bude pravděpodobně zapotřebí podstatně obsáhlejší a obecnější ontologie.



Obr. 2: Ontologická mřížka (graf)

3.4 Praktické použití, problémy a kritika

Většina toho, co sémantický web nabízí zůstává prozatím nevyužita, ale pomalu se objevují první vlaštovky, např. iniciativa Semantic Technologies Center firmy Oracle. Některé části Web 2.0 již splňují definici sémantického webu. Dobrým příkladem sémantické Web 2.0 služby je RSS – je to forma automatizované konsolidace informací z více zdrojů, což je jeden z cílů sémantického webu. Verze 1.0 dokonce používala stejné DTD jako RDF (v rámci zjednodušení však byly prvky RDF později vypuštěny).

Sémantický web jako celek rozhodně ještě nemá vyhráno jako další vývojový stupeň Webu (ačkoliv má velmi dobře našlápnuto), může dojít jen k částečné adopci jeho prvků a zbytek bude nahrazen velmi pokročilou umělou inteligencí. Současná vize sémantického webu jako celku je stále spíše akademickým projektem, než prakticky použitelnou technologií a hlavně z tohoto důvodu stále nemá podporu hlavních hráčů na trhu vyhledávání (Google a Yahoo). Douglas Merill (člověk, který stál u zrodu vyhledávání v Google) se o sémantickém webu na své nedávné přednášce na půdě ČVUT vyjadřoval velmi chladně – lze tedy předpokládat, že podpory ze strany Google se v brzké době nedočkáme (přestože je členem konsorcia W3C).

Existuje však i stinná stránka – pro některé subjekty může být nežádoucí zjednodušovat strojové zpracování (jinak také data-mining) svých stránek. Nabízí se příklad TV programu – hlavním zdrojem příjmů pro jeho poskytovatele je reklama a v případě strojového zpracování poskytovatel ztrácí zisk, protože je strojovým zpracováním odfiltrována.

1) Hyponym/Hypernym: *šarlatová, rumělková, rudá, karmínová* jsou hyponymy *červené* (jejich hypernymu) která je zase hyponymem slova *barva*. Tento typ vztahu je podstatně častější než meronym.
2) Meronym je naopak část většího celku, např. zápěstí je část ruky.
3) Ontologický slovník (pod BSD-style licencí) obsahující ~150 000 anglických slov ve ~115 000 synsetech. Je pod správou Princetonské univerzity, http://wordnet.princeton.edu/ (lze prohlížet on-line, doporučuji vyzkoušet).
4) http://www.nlm.nih.gov/research/umls/about_umls.html

Zajímavým projektem je FoaF (Friend of a Friend, www.foaf-project.org) popisující mezilidské vztahy pomocí RDF.

Základní požadavky pro úspěšné fungování sémantického webu lze definovat takto:

1) Inteligentní informační služby
   - informace na stránkách uložené ve zpracovatelné formě, např. pomocí XML RDF
2) Univerzální vyjadřování
   - jazyková nezávislost – součástí ontologií musí být vztahy mezi stejnými slovy v různých jazycích
   - nová metoda – Fuzzy logika
3) Syntaktická interoperabilita softwarových komponent (= jednotná syntaxe)
   - nikoli uspořádání pomocí syntaktických pravidel nebo uspořádání do katalogu
4) Sémantická interoperabilita zdrojů
   - definice vztahů mezi pojmy pomocí ontologií
   - sémantická organizace webu
   - automatizované sbírání znalostí (softwaroví agenti, dnes weboví roboti)

3.5 Zdroje

Účel, vlastnosti, generace

http://en.wikipedia.org/wiki/Internet

http://en.wikipedia.org/wiki/Classless_Inter-Domain_Routing

http://www.fi.muni.cz/~kas/p090/referaty/2006-podzim/ct/dns.html

Identifikace v Internetu

http://en.wikipedia.org/wiki/URI

http://www.w3.org/Addressing/

http://www.w3.org/Addressing/9710-uri-vs-url.html

http://www.ietf.org/rfc/rfc3987.txt

http://gbiv.com/protocols/uri/rfc/rfc3986.html

http://annevankesteren.nl/2005/02/iri

Sémantický web

http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21&catID=2

http://www.w3.org/2001/sw/

http://www.jfsowa.com/ontology/

http://www-ksl.stanford.edu/kst/what-is-an-ontology.html

http://en.wikipedia.org/wiki/Ontology_(computer_science)

http://www.oracle.com/technology/tech/semantic_technologies/index.html

http://infomesh.net/2001/swintro/#whatIsSw