Internet
Sémantický web

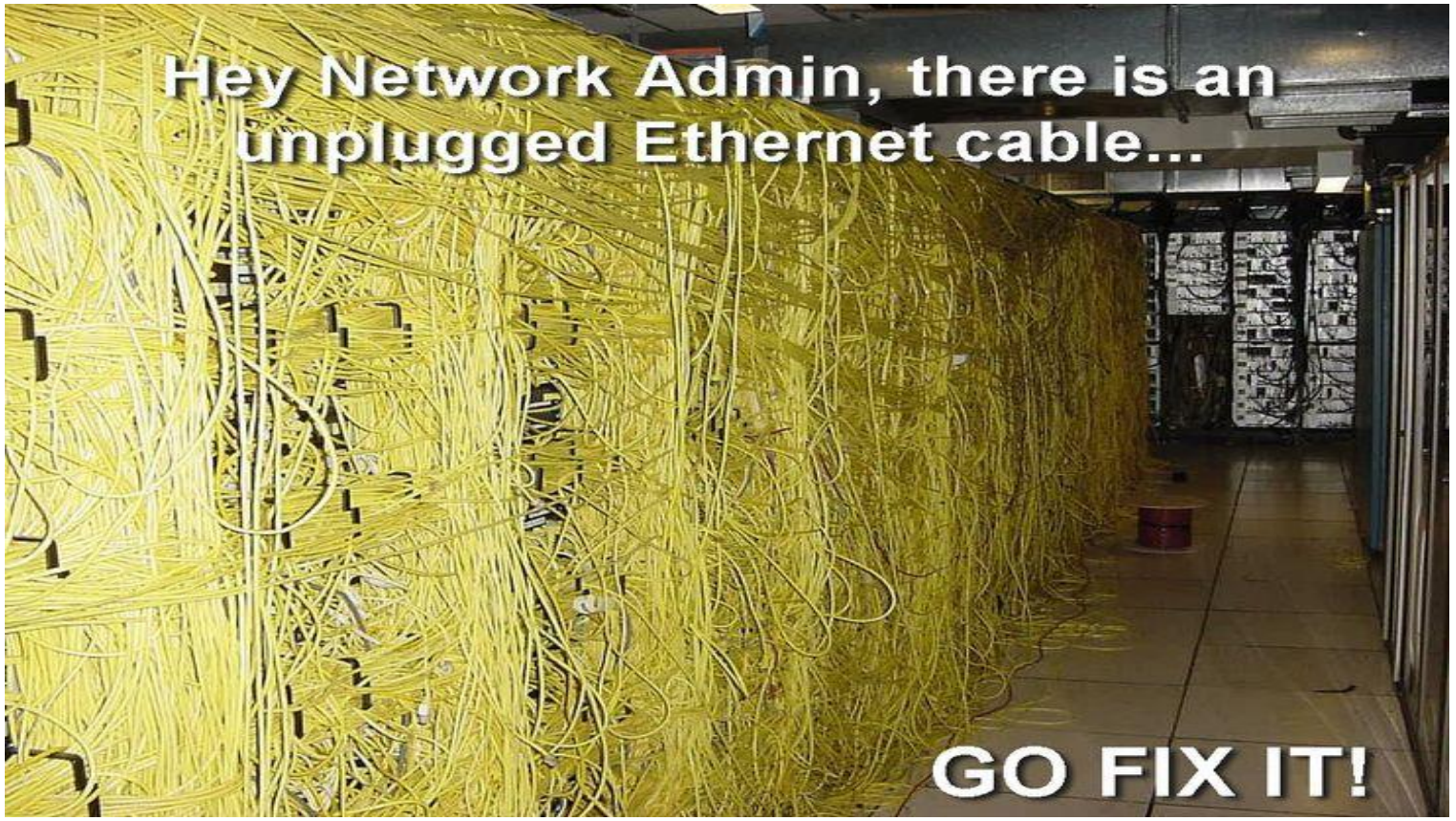
RNDr. Jaromír Plhák, Ph.D.

SIN01 - Sociální informatika

Podzim 2017

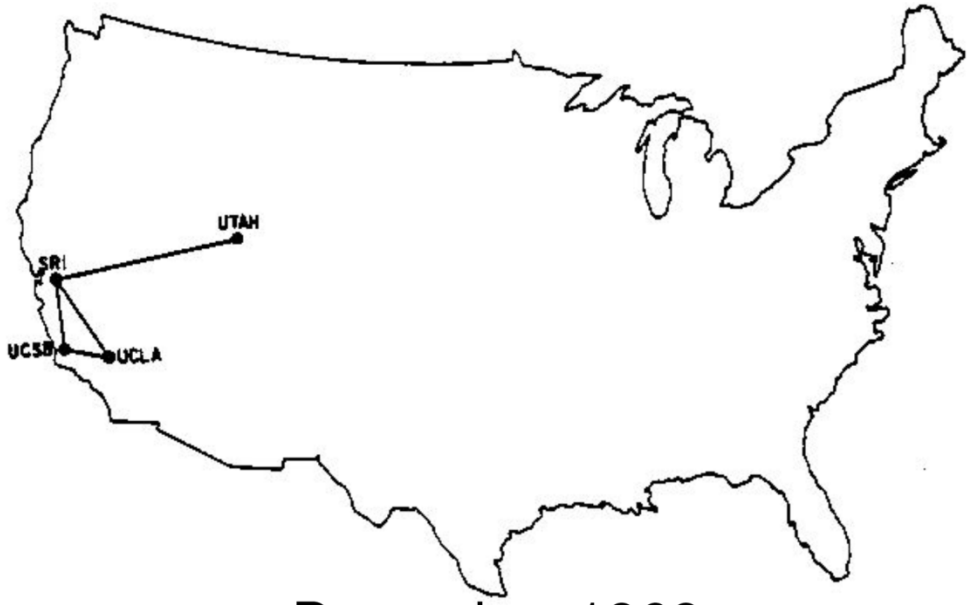
Osnova

- Internet
  - Historie
  - Služby Internetu
- Sémantický web
- Ontologie
- Crowdsourcing

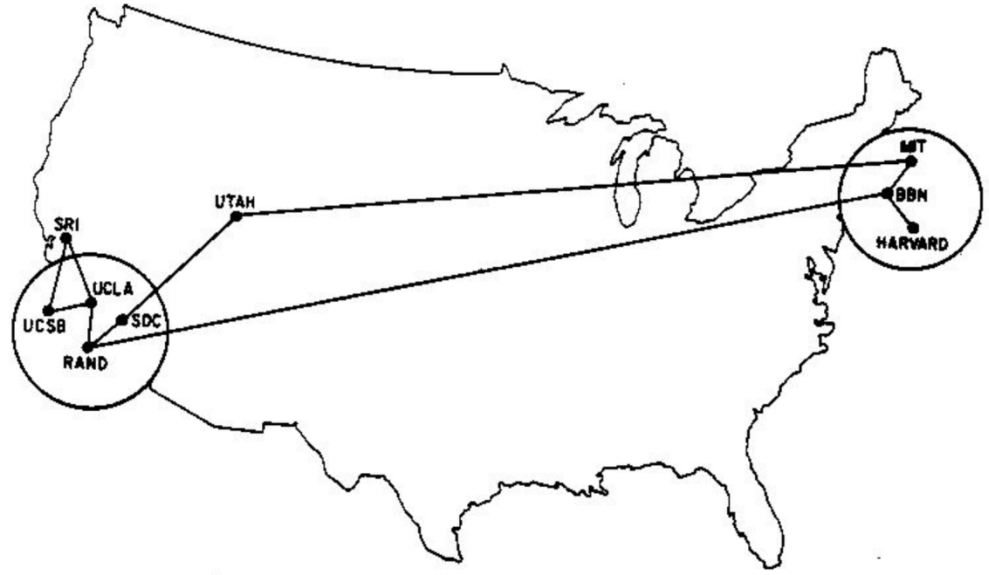


Historie (1)

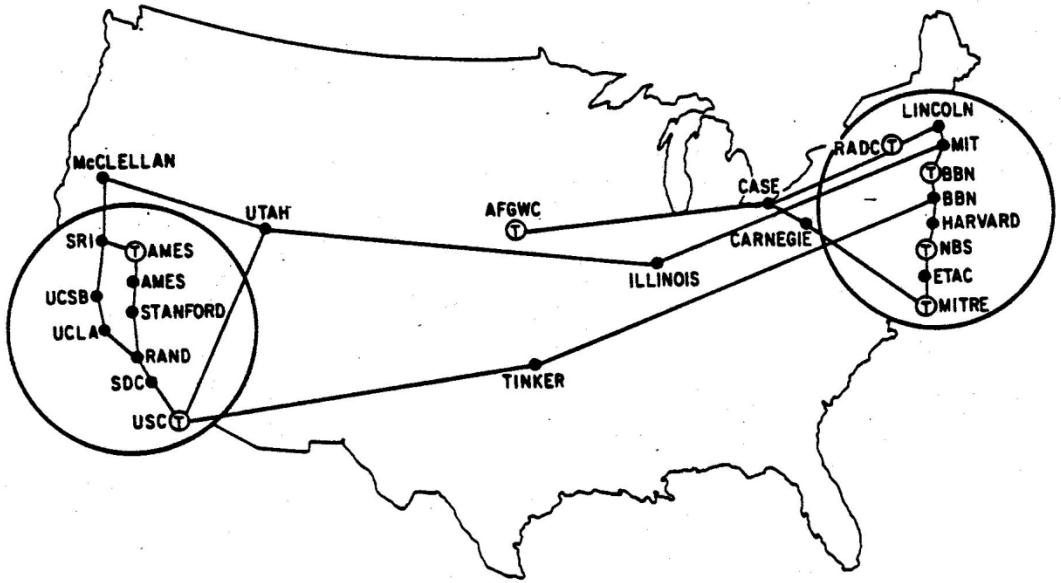
- Síť ARPANET (vývoj začal v roce 1969)
  - Považována za první počítačovou síť vůbec
- Cíle:
  - Zjednodušení komunikace
  - Sdílení SW a HW
  - Sdílení dat a informací
- 1973 - připojena Evropa
  - Norsko



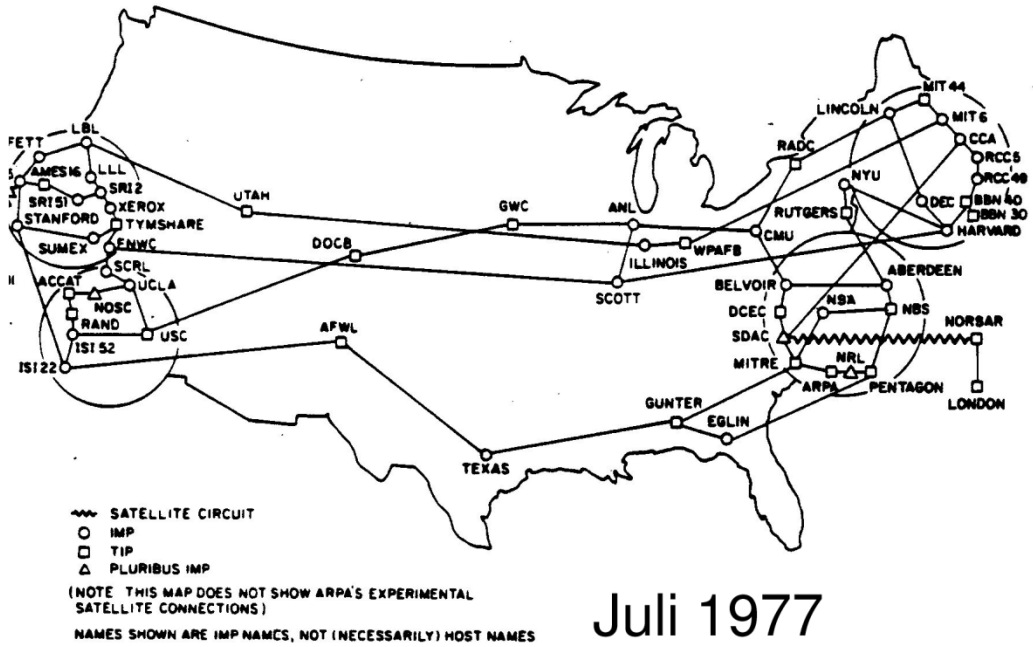
Dezember 1969



Juni 1970



März 1972



Juli 1977

Historie (2)

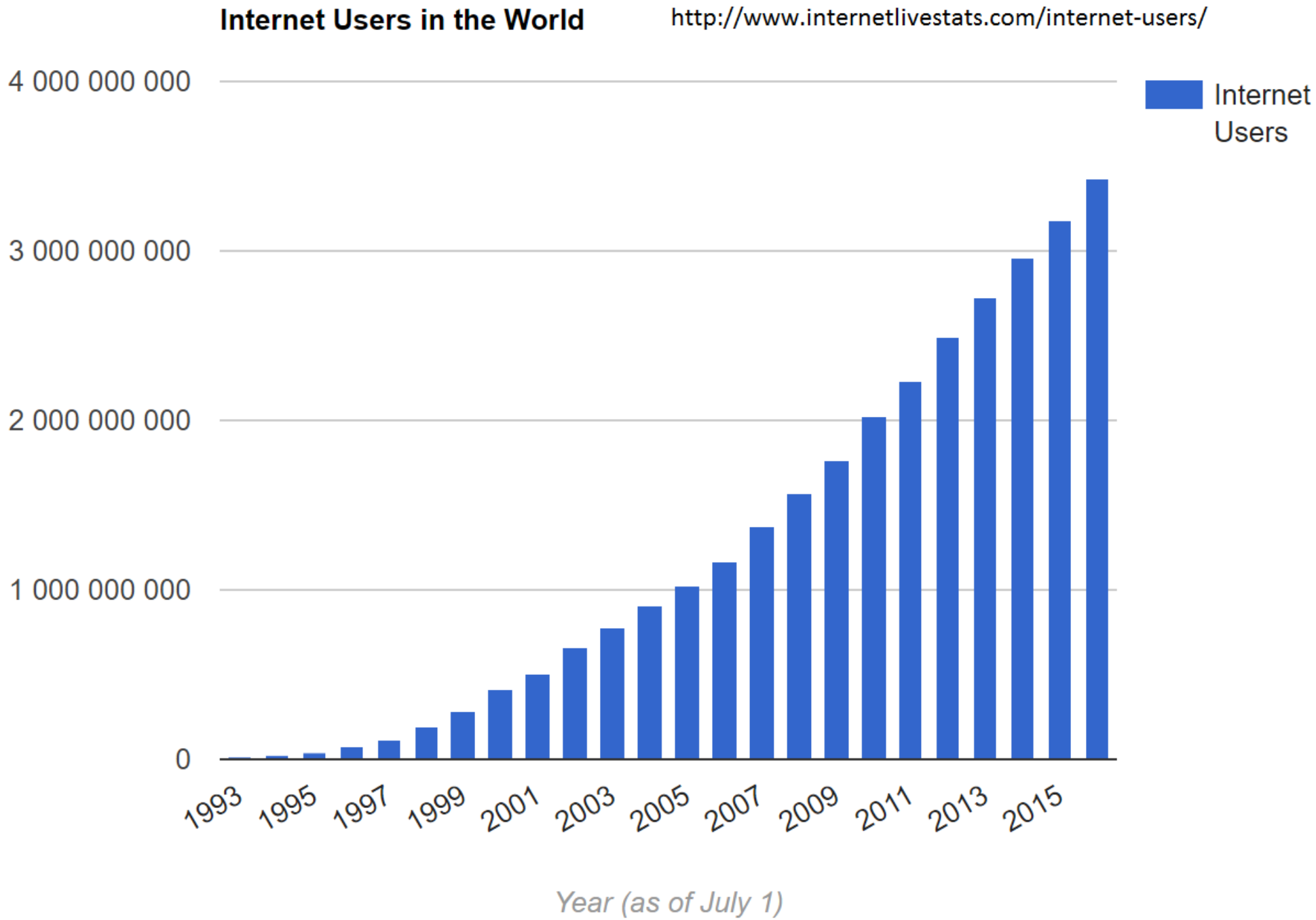
- 1992 – připojeno Československo
  - ČVUT
- 1995 – Amazon, e-Bay
- 1996 – 55 milionů uživatelů, Seznam
- 1998 – Google, PayPal
- 2000 – 250 milionů uživatelů
- 2003 – 600 milionů uživatelů
- 2005 – 900 milionů uživatelů
- 2006 – více než miliarda uživatelů

Oblasti působnosti Internetu

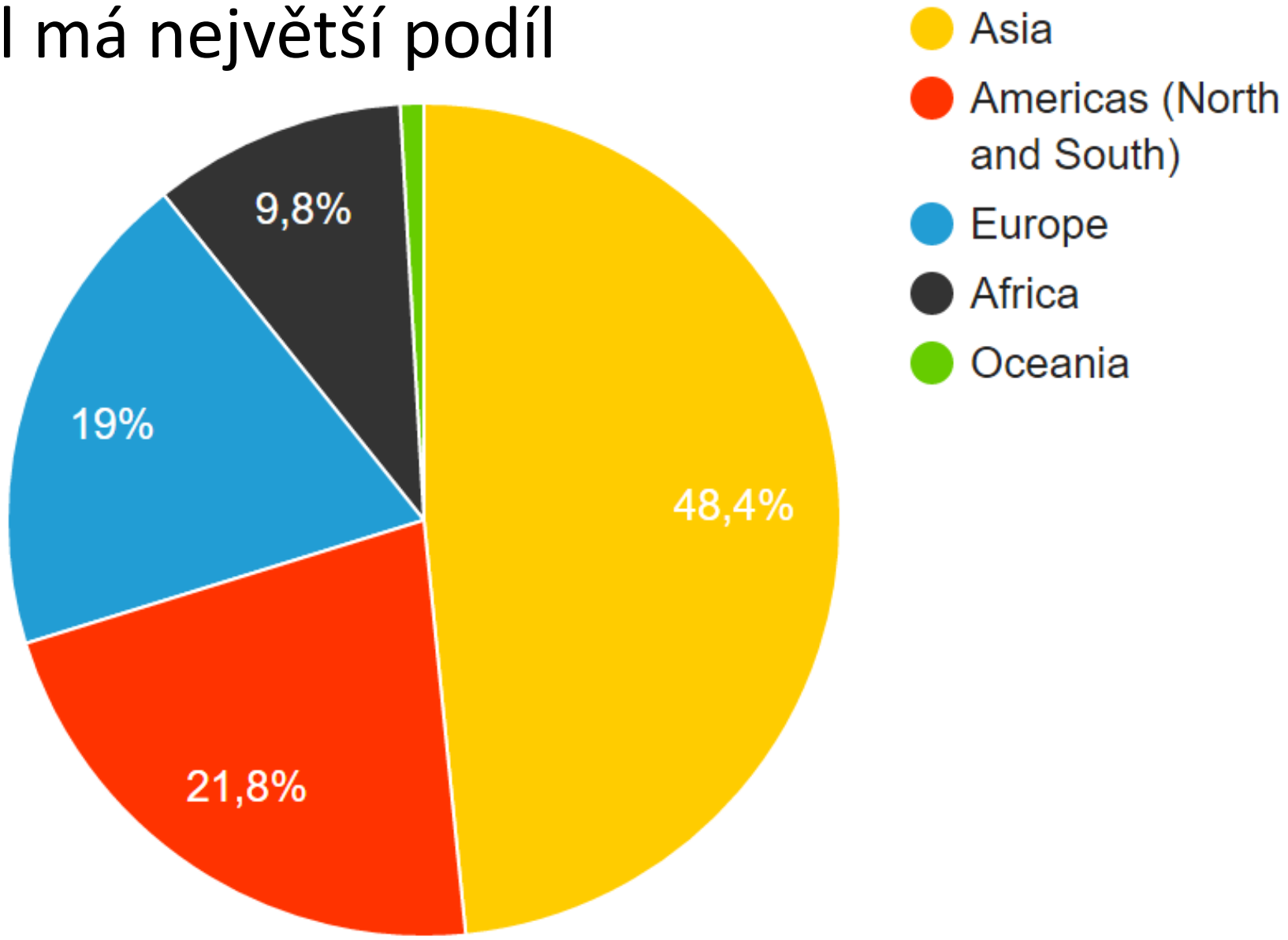
- 70. léta 20. století
  - Zejména vojenský sektor, zajištění obrany USA
- 80. léta 20. století
  - Přechod z vojenského sektoru do akademického, zajištění komunikace v nekomerční sféře
- 90. léta 20. století
  - Přechod ze sektoru akademického do komerčního a postupně do všech oblastí lidských činností
- Prudký rozvoj ICT, požadavky na kvalitu služeb a zabezpečení přenášených informací

Současnost Internetu

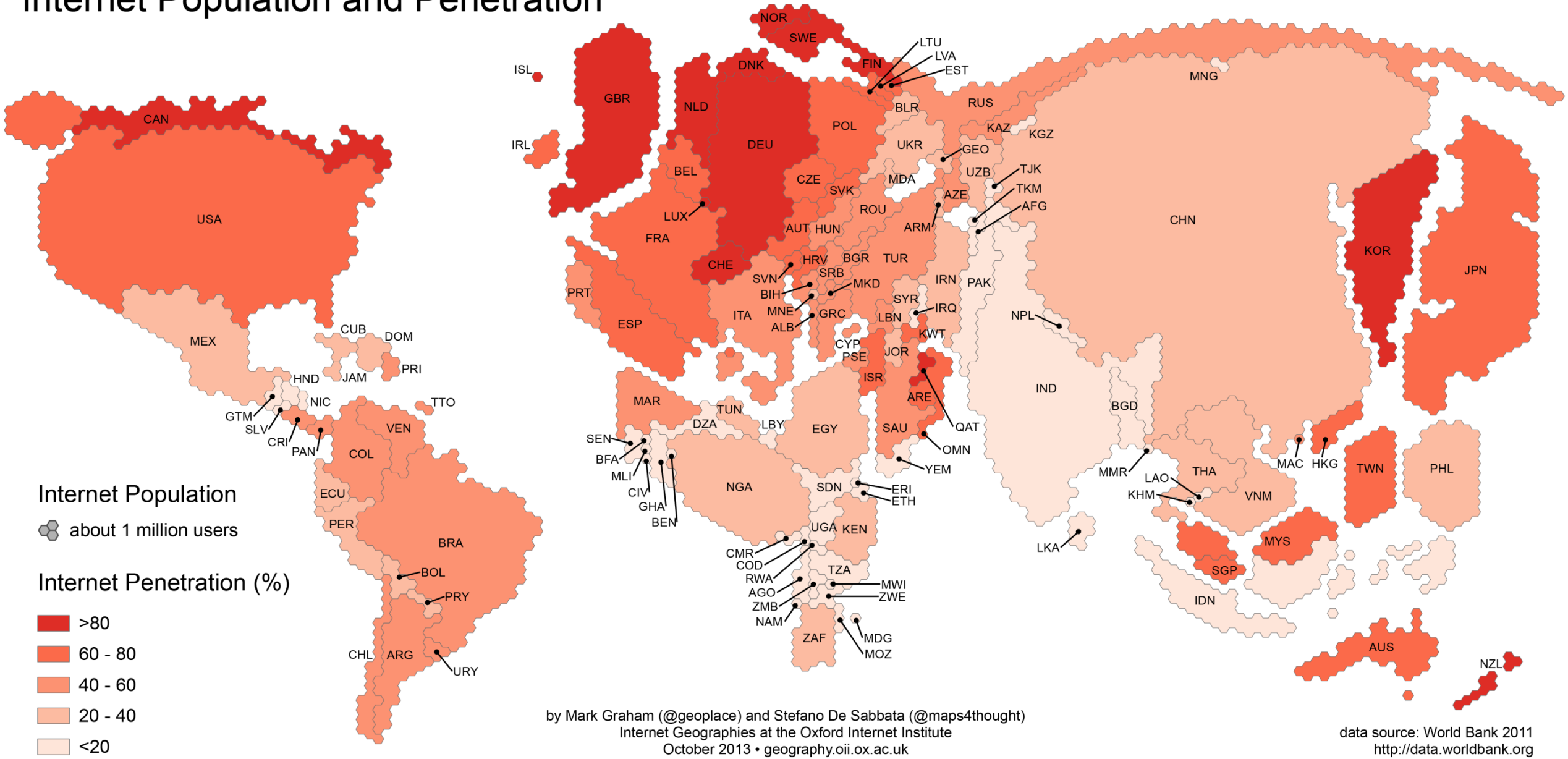
- Veřejná síť typu WAN (Wide Area Network)
  - Nemá vlastníka
- Infrastruktura je provozována na komerční bázi
  - ISP (Internet Service Provider)
- Negativní jevy
  - Riziko omezení soukromí
  - Zázemí pro kriminální a nelegální aktivity pod rouškou zdánlivé anonymity
  - Prohlubování rozdílů mezi technologicky vyspělým a technologicky zaostalým světem


Internet Users in the World
http://www.internetlivestats.com/internet-users/
4 000 000 000
3 000 000 000
2 000 000 000
1 000 000 000
0
Internet Users
1993
1995
1997
1999
2001
2003
2005
2007
2009
2011
2013
2015
Year (as of July 1)

Který světadíl má největší podíl uživatelů?



Internet Population and Penetration



Internet Population

about 1 million users

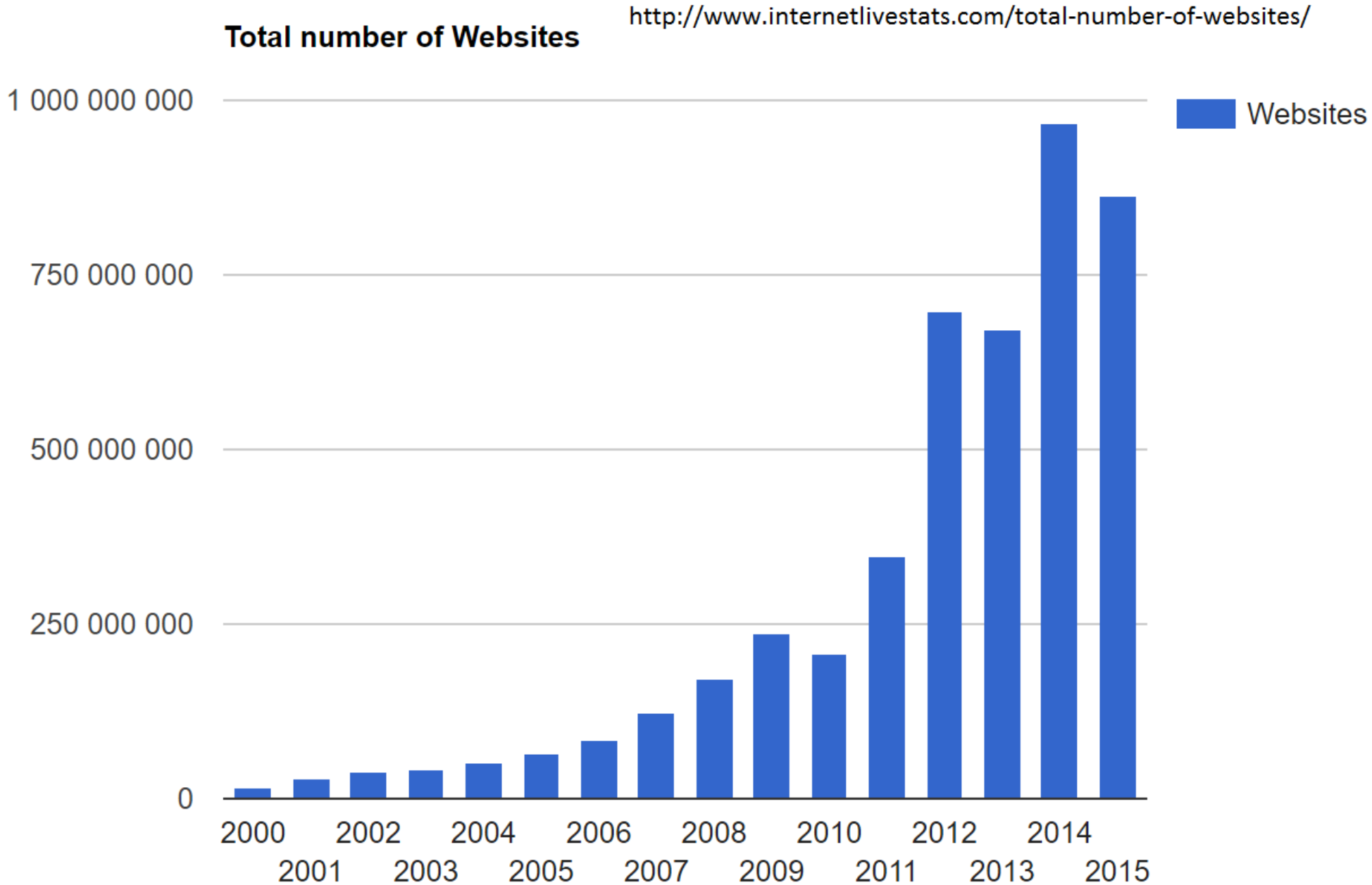
Internet Penetration (%)

- >80
- 60 - 80
- 40 - 60
- 20 - 40
- <20

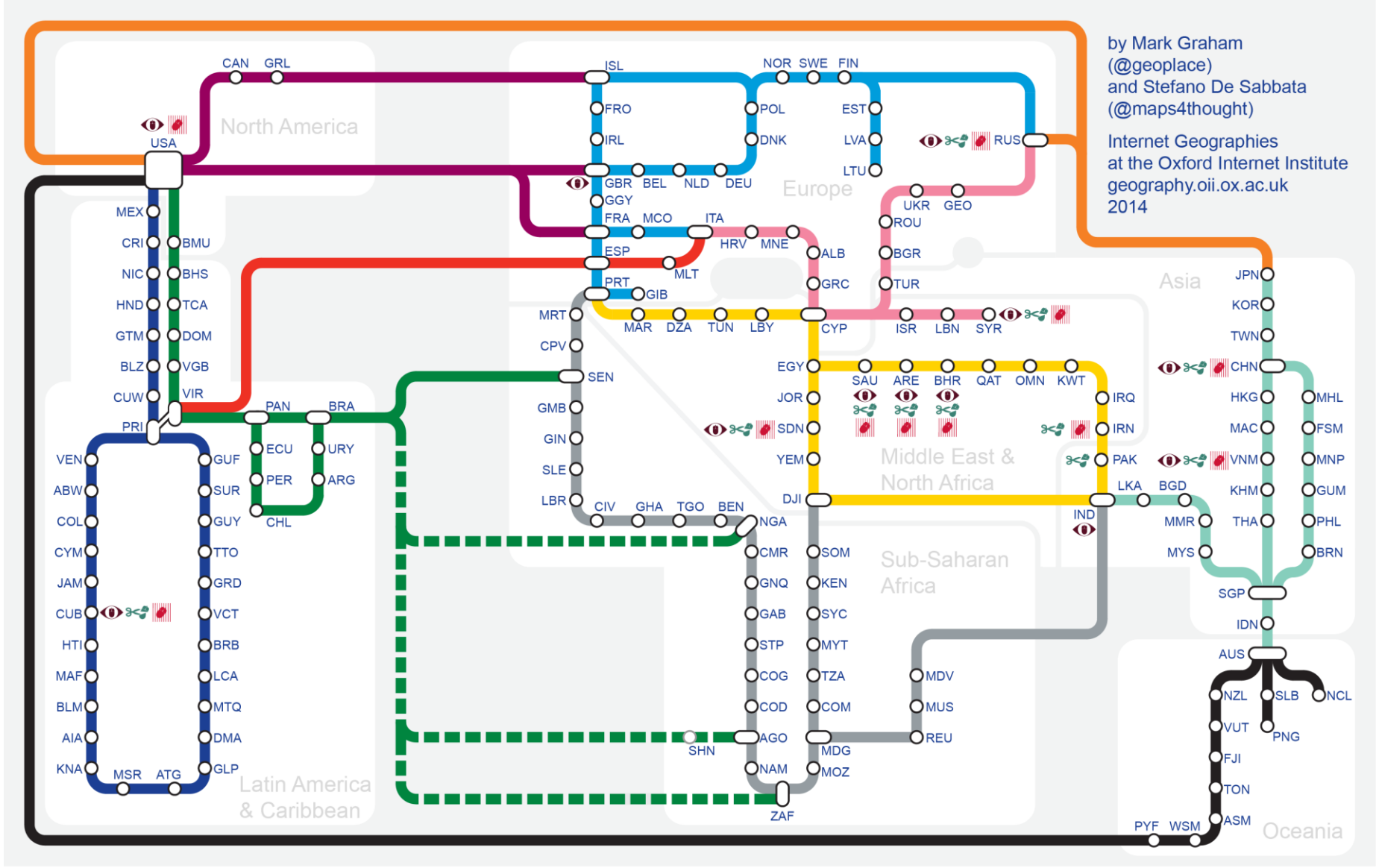
by Mark Graham (@geoplace) and Stefano De Sabbata (@maps4thought)
Internet Geographies at the Oxford Internet Institute
October 2013 • geography.oii.ox.ac.uk

data source: World Bank 2011
http://data.worldbank.org

http://www.internetlivestats.com/total-number-of-websites/




AGE
of
INTERNET
EMPIRES
Top Sites
Google - 62 countries
Facebook - 50 countries
Baidu - 2 countries
Yahoo! - 2 countries
AlWatan Voice - 1 country
Mail.ru - 1 country
VK - 1 country
Yandex - 1 country
data incognita
NORTH AMERICA
EUROPE
ASIA
AFRICA
SOUTH AMERICA
AUSTRALIA
GREENLAND
ARCTIC OCEAN
NORTH ATLANTIC OCEAN
SOUTH ATLANTIC OCEAN
INDIAN OCEAN
NORTH PACIFIC OCEAN
SOUTH PACIFIC OCEAN
by Mark Graham (@geoplace) and Stefano De Sabbata (@maps4thought)
Internet Geographies at the Oxford Internet Institute
August 2013 • geography.oii.ox.ac.uk
data source: Alexa, www.alexa.com • made with Natural Earth
0 1,000 2,000 3,000 4,000
Miles



Internet Tube

An abstraction of the global
submarine fibre-optic cable network

Milionář Lugner má rakovinu a rozvádí se s manželkou mladší o 57 let

VIDEO: Smažák vám klidně připraví i Pohlreich. A jako dietní pokrm

Mladá žena se pustila do proměny bytu, nezastavila ji ani nosná zeď

Výjimečná smůla: loď kotvou přerušila internetové kabely na mořském dně

29. listopadu 2016 13:51

Nejméně tři podmořské internetové optické kabely mezi Velkou Británií a Normandskými ostrovy přerušila lodní kotva. Konektivitu tak provizorně zajišťuje kabel z Francie, oprava potrvá několik dní.

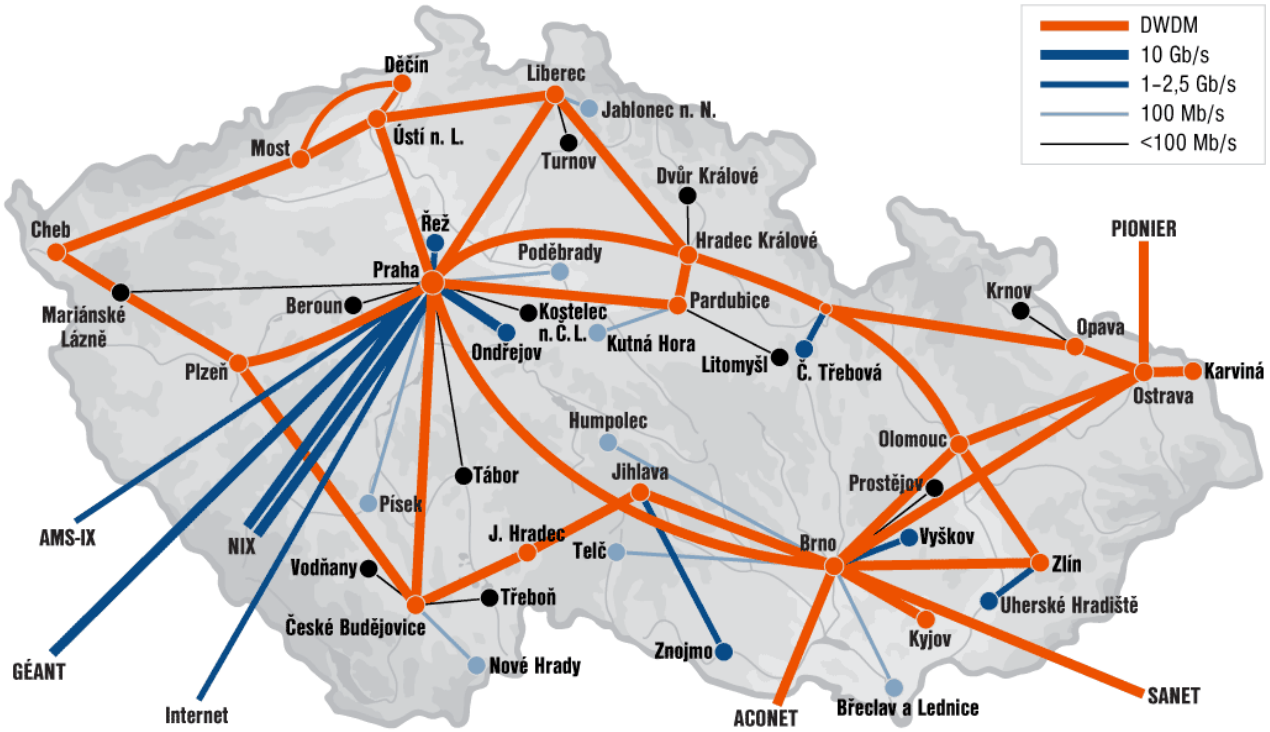
Lodě podmořské kabely pokládají, někdy je však jiné kotvou poškodí. | foto: Urs Hölzl, Google

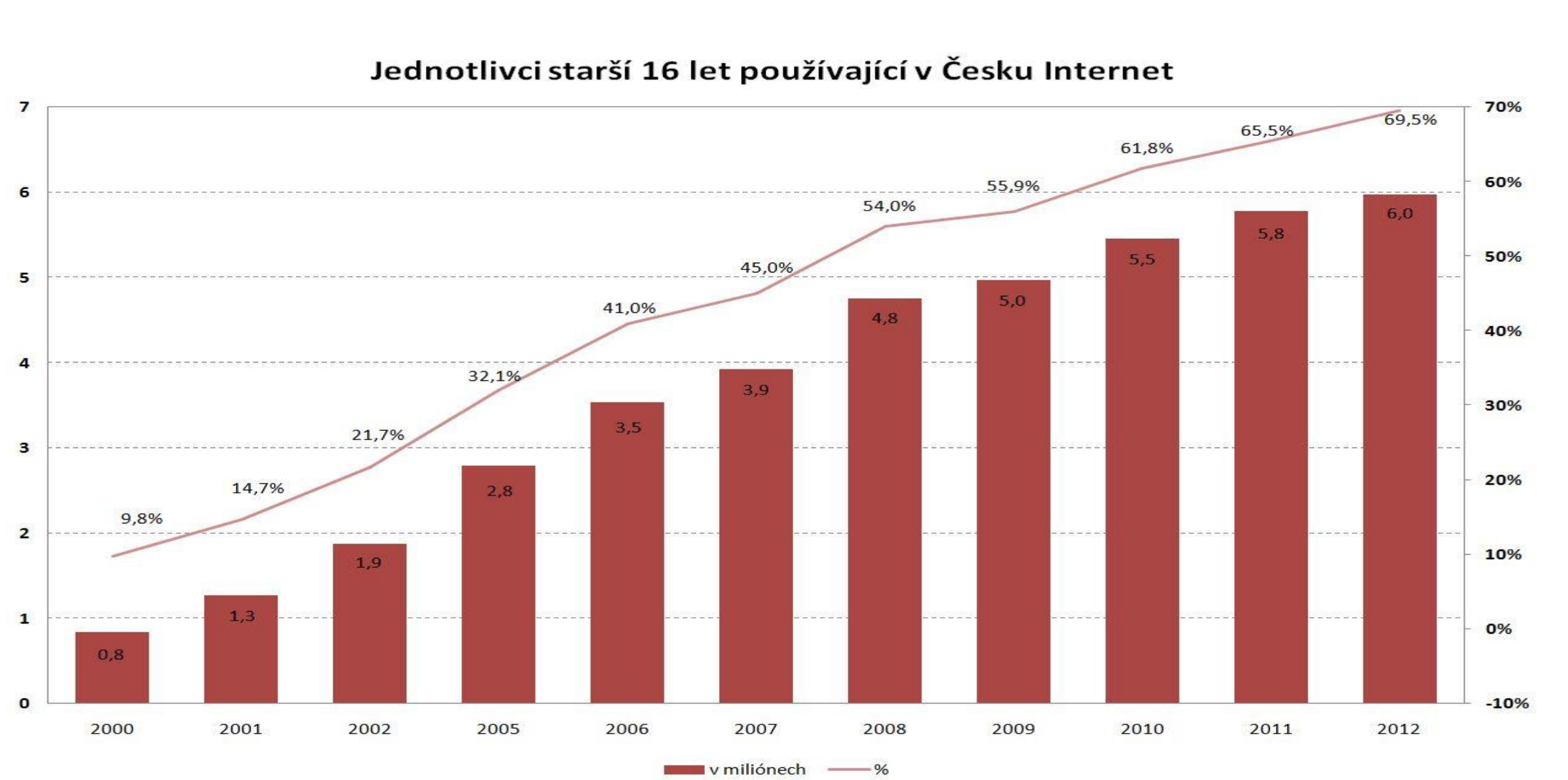
„Je to nejzávažnější komplikace, kterou jsme u podmořských kabelů kdy viděli," prohlásil Daragh McDermott, ředitel vnějších vztahů telekomunikačního operátora JT, jemuž poškozené kabely patří. „Poškození tří podmořských kabelů v jednom dni, to je nesmírně smolná a zcela nová situace", doplnil McDermott.

Všechny tři podmořské kabely zajišťovaly konektivitu Normandských ostrovů s Velkou Británií a podle všech indicií je poškodila lodní kotva tažená po mořském dně. Která loď škodu způsobila se zatím vyšetřuje, spekuluje se také o poškození dalších kabelů patřících jiným společnostem. Konektivitu provizorně zajišťuje kabel z Francie, uživatelé vysokorychlostního internetu si musí dočasně zvyknout na nižší propustnost přetíženého spoje.

Internet v ČR

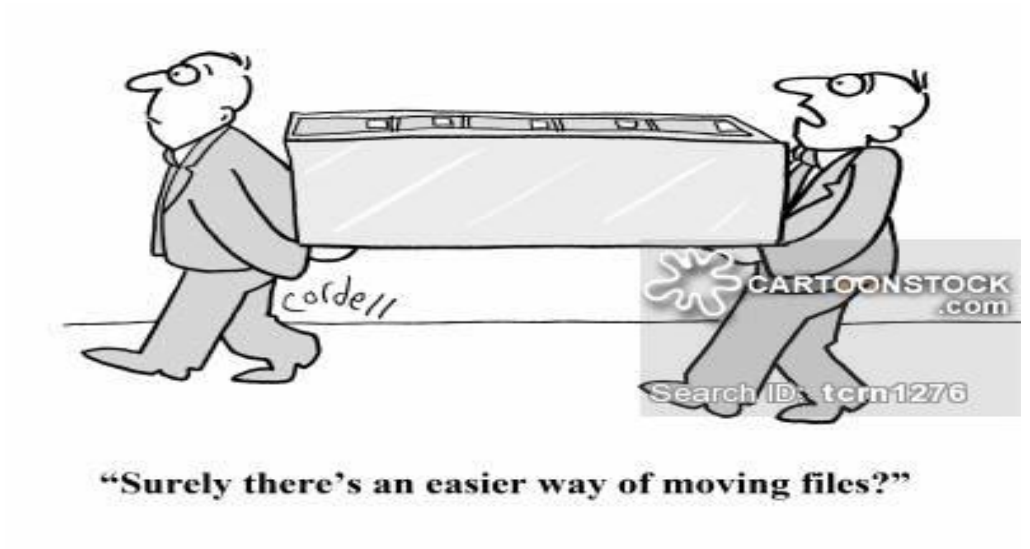
- První organizací, zajišťující správu na celém území republiky, bylo sdružení CESNET
  - Czech Education and Scientific NETwork
  - Česká akademická síť
  - CESNET 2 od roku 2001




Jednotlivci starší 16 let používající v Česku Internet
7
6
5
4
3
2
1
0
70%
60%
50%
40%
30%
20%
10%
0%
-10%
9,8%
14,7%
21,7%
32,1%
41,0%
45,0%
54,0%
55,9%
61,8%
65,5%
69,5%
0,8
1,3
1,9
2,8
3,5
3,9
4,8
5,0
5,5
5,8
6,0
2000
2001
2002
2005
2006
2007
2008
2009
2010
2011
2012
v miliónech
%

Služby Internetu

- Výměna dat a informací
  - Přenos souborů
  - Vzdálený přístup
  - WWW
- Komunikace
  - Elektronická pošta
  - Diskusní systémy
  - IRC, ICQ, Jabber
  - IP telefonie
  - Videokonference



WWW – World Wide Web

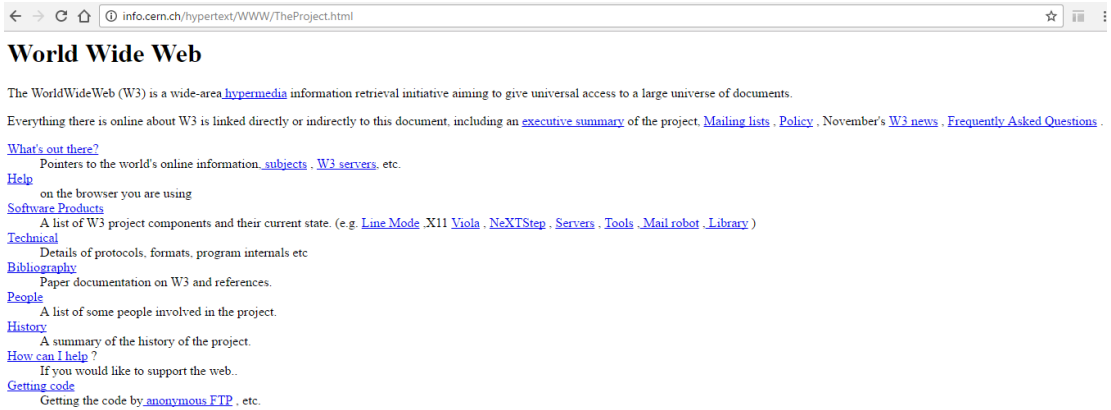
- Systém serverů, které uchovávají hypertextové dokumenty a další zdroje a umožňují k nim přístup
- Distribuovaný systém podporovaný protokolem **HTTP** (HyperText Transfer Protocol)
- **Hypertext** = způsob organizace informační jednotky
  - Nelineární dokument, obsahující odkazy na další související informační jednotky
- Komunikace na principu klient - server

Internet vs. WWW (1)

- Nejsou synonyma, přestože se tak používají
- **Internet** – celosvětová síťová struktura počítačů uzpůsobených k vzájemnému přenosu dat a tudíž k výměně informací
- **WWW** – podsystém Internetu, soustava dokumentů navzájem propojených odkazy a spolupracujících pomocí protokolu HTTP

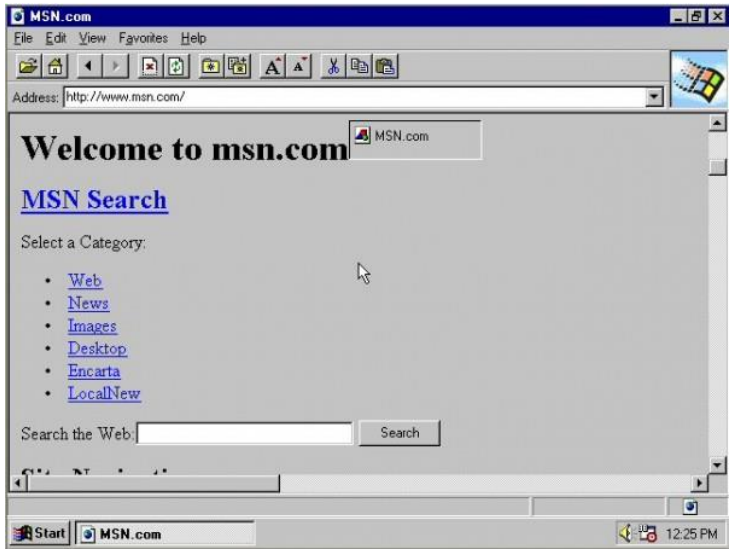
Internet vs. WWW (2)

- Internet existuje nezávisle na WWW, ale WWW by bez Internetu existovat nemohlo
- Vznik WWW v roce **1989** jako prostředek komunikace mezi odlehlými pracovišti
  - Tim Berners-Lee, CERN
- WWW systém je tvořen webovými stránkami
- Populární díky jednoduchosti jazyka
  - (X)HTML (5)

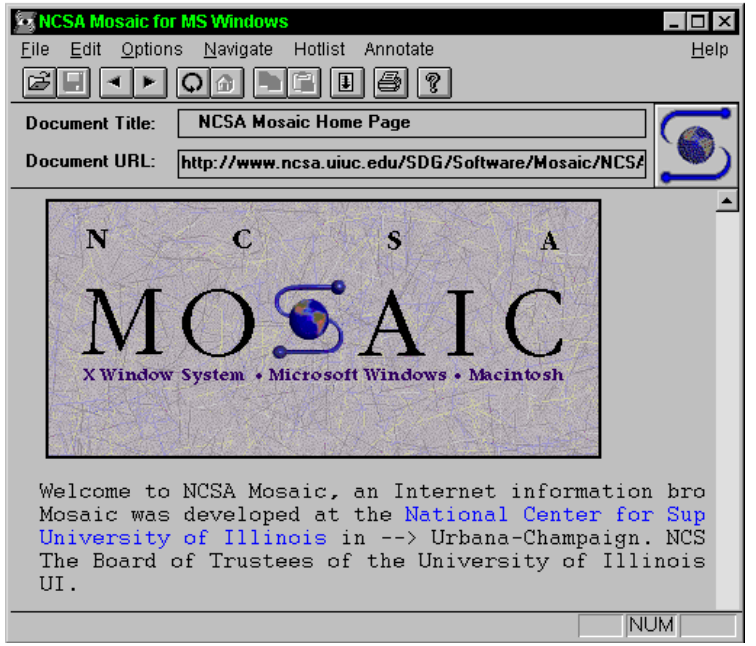


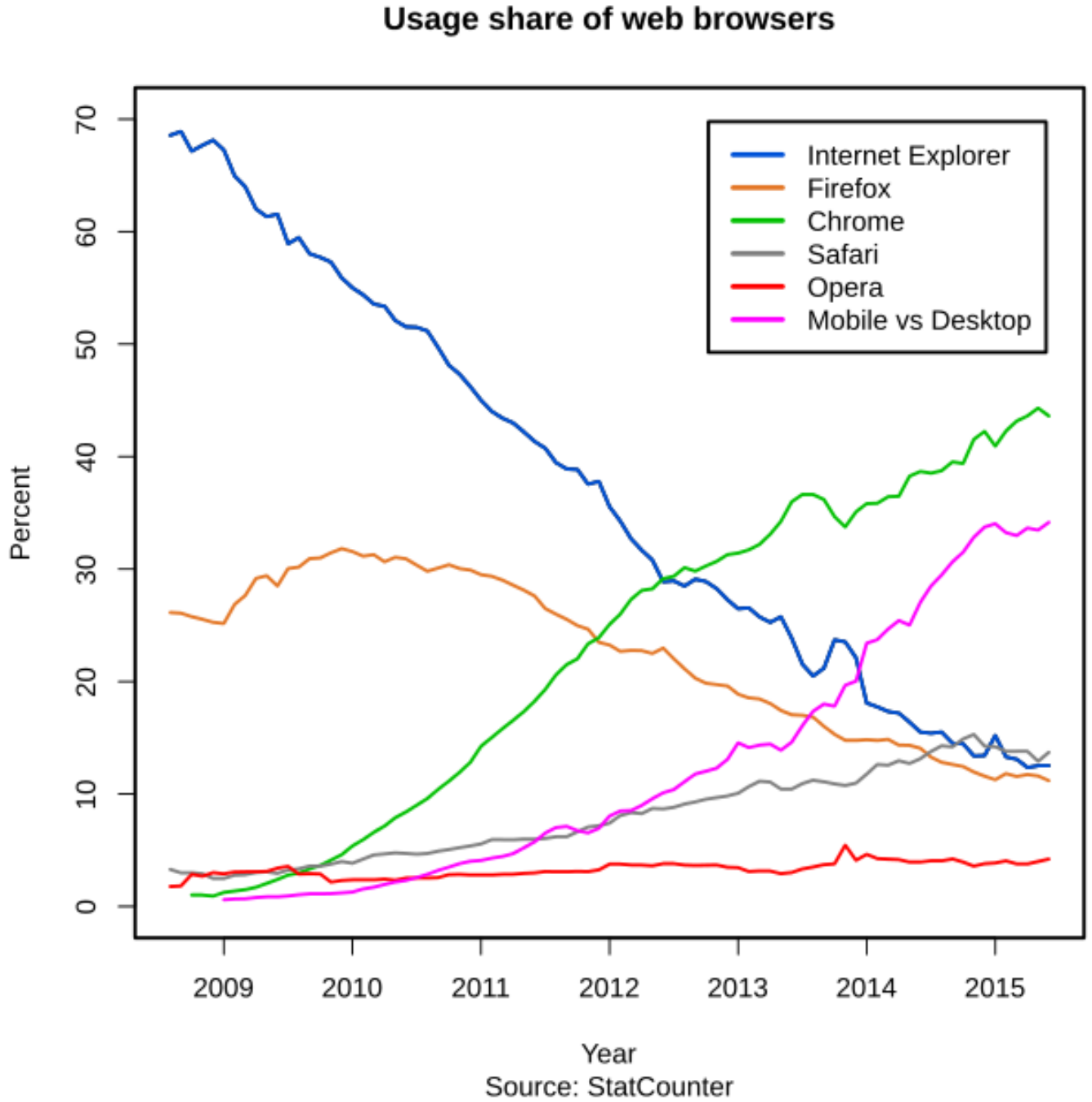
Webový klient

- V roce 1993 vznikl první grafický prohlížeč Mosaic
- V současnosti pestrá nabídka webových prohlížečů
- Nejpopulárnější **webové prohlížeče**
  - **Mozilla Firefox**
  - **Chrome**
  - **Opera**
  - **Safari**
  - **Internet Explorer**



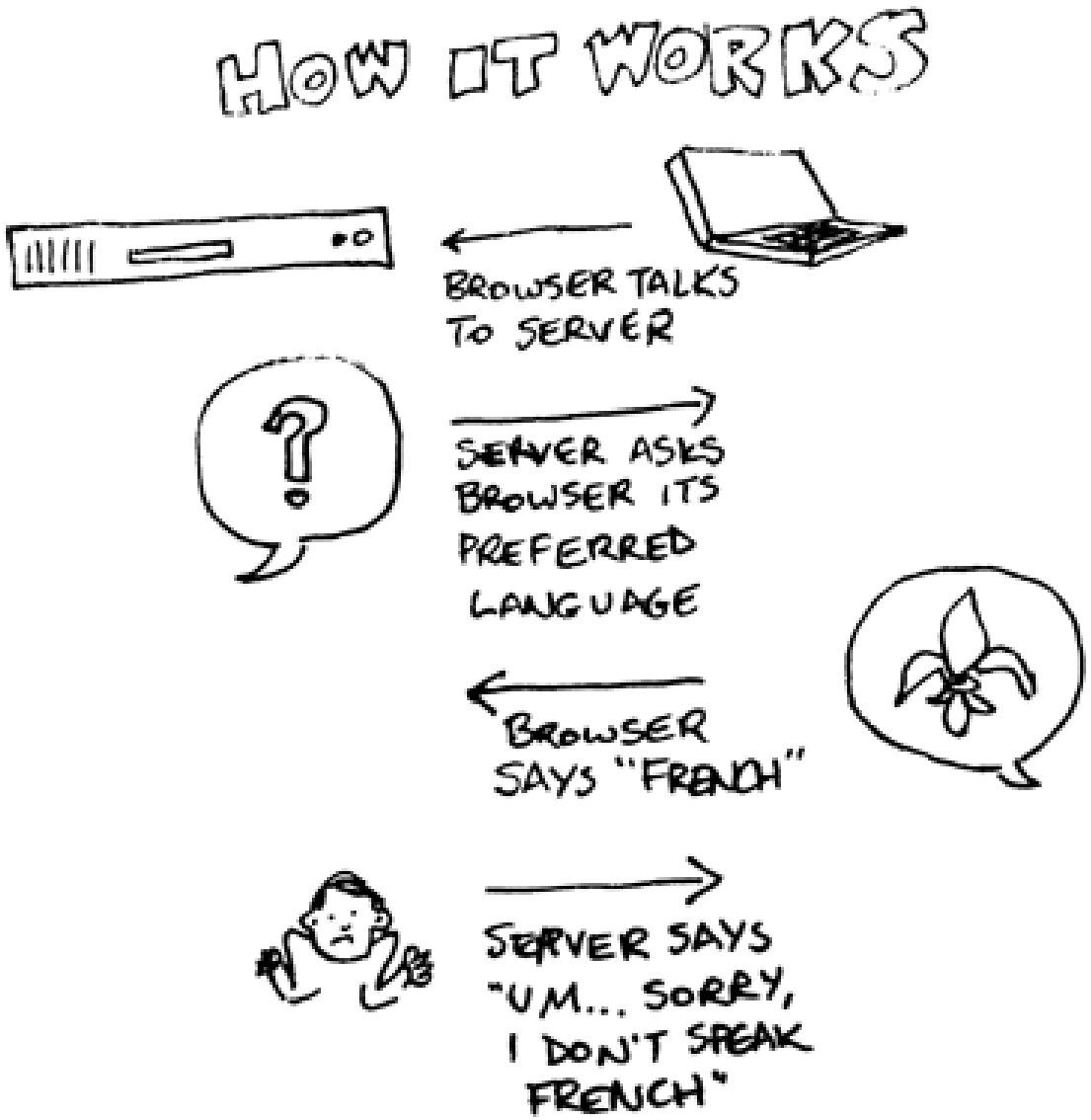
From Computer Desktop Encyclopedia
Reproduced with permission.
© 2004 National Center for Supercomputing Applications




Usage share of web browsers
Internet Explorer
Firefox
Chrome
Safari
Opera
Mobile vs Desktop
Percent
0
10
20
30
40
50
60
70
2009
2010
2011
2012
2013
2014
2015
Year
Source: StatCounter

Webový server

- Nejznámější **webové servery**
  - **Apache**
    - Pro operační systémy třídy Unix
  - **Internet Information Services**
    - Pro Windows
  - **Lotus Domino**
    - Komerční, pro platformy IBM
  - **Nestcape Enterprise, Fast Track**
    - Komerční
  - **Netware**
    - Komerční systém firmy Novell

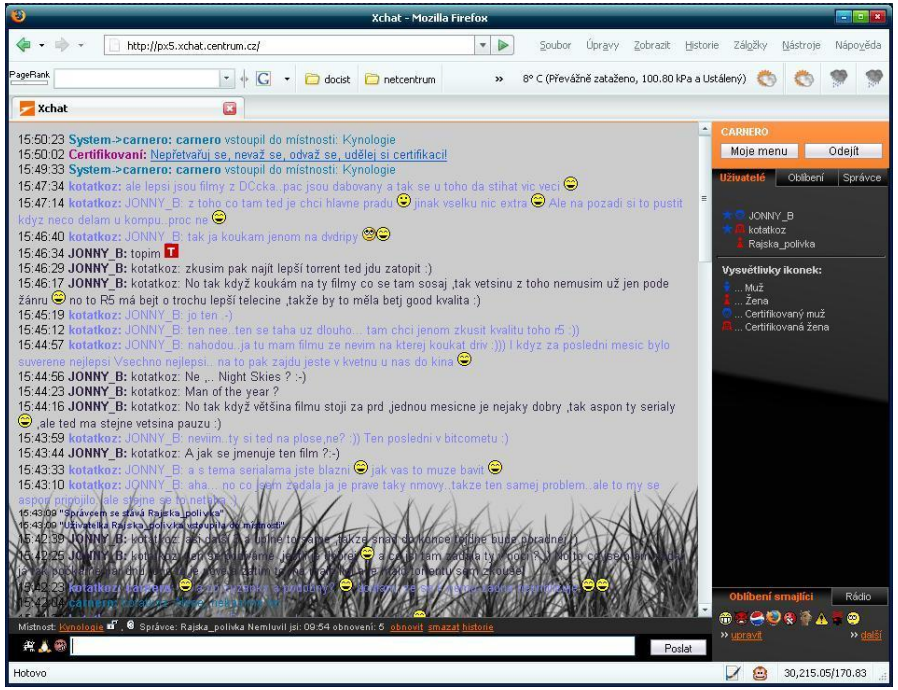


E-mail

- Způsob odesílání, doručování a přijímání zpráv přes elektronické komunikační systémy
- SMTP
  - Simple Mail Transfer Protocol
- Tělo zprávy
- Přílohy
  - Každá příloha je určena typem obsahu (Content-Type)
  - Příklady: text/plain, text/html, image/jpeg, image/png, application/msword

Instant messaging

- Zpráva je doručena ve velmi krátké době od odeslání
- Status
  - Online / Offline / Cofee break / …
- Internet Relay Chat
  - Skupinová komunikace v místnostech – kanálech
- ICQ, Jabber
- Viber, WhatsApp, Facebook Messenger, …



IP telefonie

- **VoIP** (Voice over Internet Protocol)
- Přenos hlasového (telefonního) signálu prostřednictvím počítačové sítě
  - V současnosti i videopřenos
- Příklady klientských aplikací
  - Skype
  - Microsoft NetMeeting
  - …

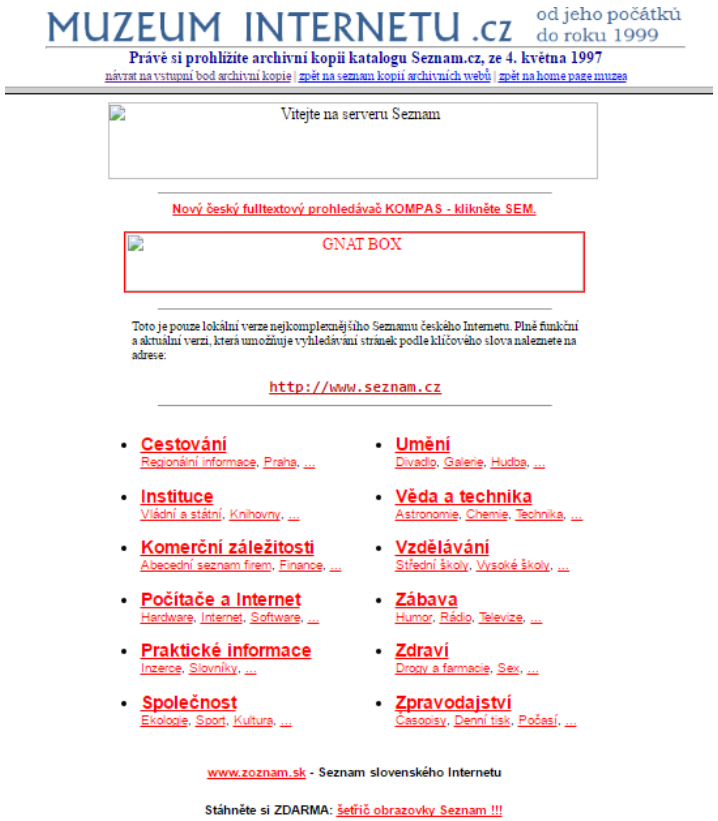
Videokonference

- Nejpokročilejší forma **dvoustranné** nebo skupinové komunikace na Internetu
- Založeny na přenosech audia a videa
- Technicky náročný proces, vytvářející komplexní prostředí pro distribuovanou týmovou spolupráci
- Méně náročná je **jednosměrná** relace, kdy vysílající uzel je pouze jeden a všichni účastníci multimediální data pouze přijímají
  - Využití zejména pro distanční výuku (E-learning)

Unified communications

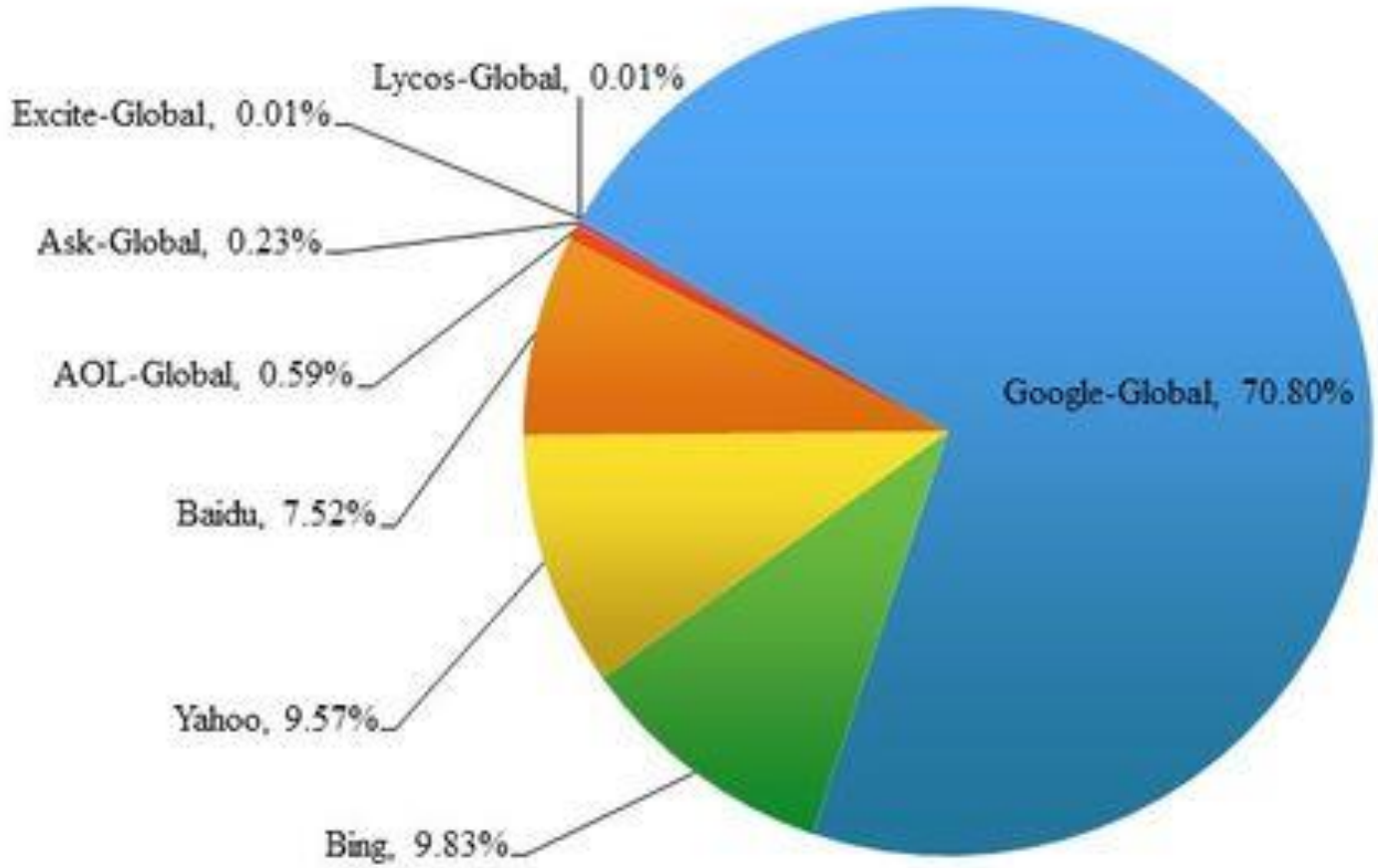
Vyhledávací služby na Internetu

- Google – http://www.google.com
- Bing - https://www.bing.com/
- Seznam – http://www.seznam.cz
- Centrum – http://www.centrum.cz
- Baidu - www.baidu.com/
- …



CIW

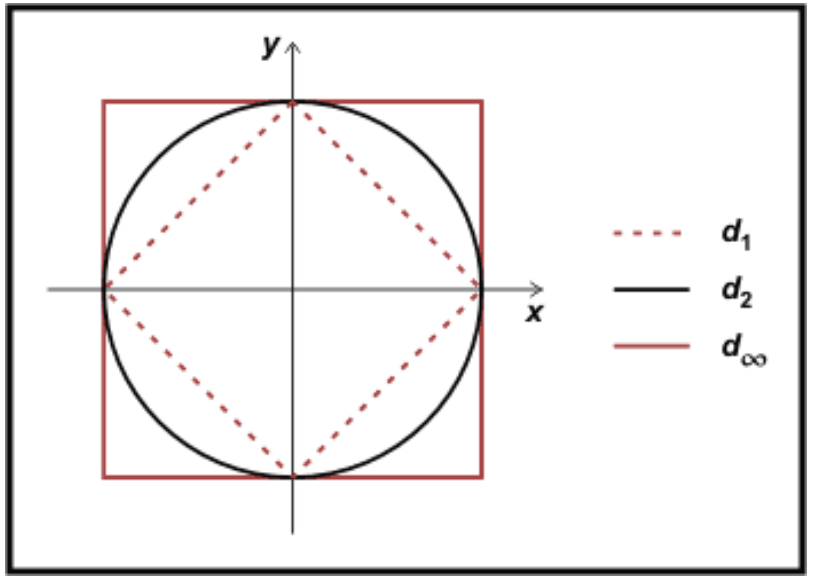
Search Engine Market Share in the World in June 2015



China Internet Watch

Source: Net Applications, July 2015

Podobnostní vyhledávání



- Složité datové struktury, obrázky
- Vyhledávání založené na podobnosti (metrice)
- Metrické prostory
- Škálovatelné algoritmy
  - Polynomiální složitost není postačující
- Podpora grafických algoritmů a AI

Sémantický web

- Metody a techniky pro přiřazení významu (sémantiky) informacím na webu
- Web rozšířený o metadata
- **Metadata** = data o datech
- Postaven na formátu **Resource Description Framework** (RDF) a **Ontology Web Language** (OWL)

Cíle sémantického webu

- **Integrovat data** z různých zdrojů
- Umožnit **výměnu dat** mezi aplikacemi napříč celým webem
- Umožnit **kvalitnější strojové vyhledávání** informací na webu
- Umožnit **popsat vztahy** mezi daty a objekty v reálném světě
- **Přiřadit** informacím na webu přesný **význam**

Google restaurace u pavouka

Internet Obrázky Mapy Nákupy Videa Více Vyhledávací nástroje

Přibližný počet výsledků: 76 300 (0,50 s)

Restaurace U Pavouka
www.upavouka.cz/
Vítáme Vás na nových stránkách **restaurace U pavouka**. I takovéto dobroty u nás máme. Pojďte ochutnat, těšíme se na Vaši návštěvu! Více informací zde.

Vranovská 52, 614 00 Brno
545 577 157
Jídelní lístek - Kontakty - Zahrádka - Nabídky

Jídelní lístek - Restaurace U Pavouka
www.upavouka.cz/jidelni-listek/
Prohlédněte si naši nabídku jídla a pití. Denní nabídka je pravidelně aktualizovaná. Soubory jsou ve formátu PDF, odkazy se po kliknutí zobrazí do nového okna.

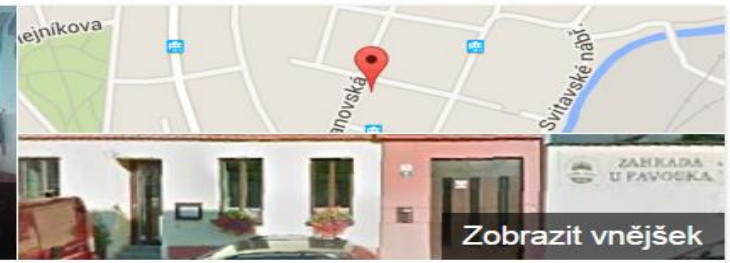
Ano, šéfe! nepomohlo. U Pavouka krachlo manželství!
www.blesk.cz/.../ano-sefe-nepomohlo-u-**pavouka**-krachlo-manzelstvi.ht...
Penzion **U Pavouka** v Hořesedlech se po návštěvě drsného kuchaře Zdeňka Pohlreicha z pořadu Ano, šéfe! otřásl v základech.

Krčma U Pavouka - Medieval Tavern
www.krcmau**pavouka**.cz/
Středověká krčma **U Pavouka**, Praha | The Medieval Tavern **U Pavouka**, Prague, CS.

Aukce Restaurace u Pavouka z insolvenčního řízení | gladys ...
https://www.gladys.cz/.../aukce-**restaurace-u-pavouka**-z-insolvencniho-ri...
Lokalita, Hořesedly. Ulice, Hořesedly 146. ID zakázky, 001694-1905201501. Typ, Prodejní. Způsob, Anglická s postupným snížením nejnižšího podání.

Restaurace U Pavouka (Brno, Husovice) • Firmy.cz
www.firmy.cz/detail/324008-**restaurace-u-pavouka**-brno-husovice.html
Aktuálně ověřené informace, kontakty a hodnocení **Restaurace U Pavouka**, Brno, Husovice Provozujeme restauraci s nabídkou denního menu, minutek, ...

U Pavouka – Necyklopedie – Wikia - Brno
necyklopedie.wikia.com/wiki/U_**Pavouka**
„Jestli si pánové Brňáci myslí, že mezi jinými provozovnami státního podniku Restaurací a jídelen (Ra.J) byla **restaurace u Pavouka** něčím výjimečná, tak já je



Restaurace U pavouka ★

Web. stránky Trasa

3,8 ★★★★☆ 11 recenzí Google
Restaurace

Adresa: Vranovská 52, 614 00 Brno
Telefon: 545 577 157

Recenze
Napište recenzi
11 recenzí Google

Lidé také hledají
Zobrazit další (více než 15)


Restaurace U Štiky

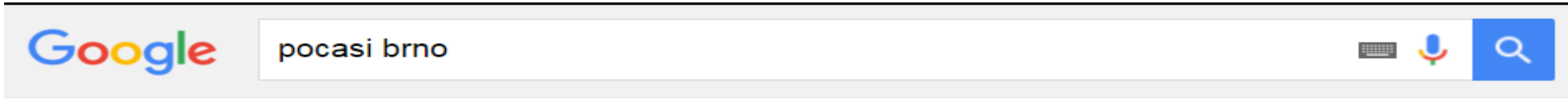

U měděné pánve


Restaurace U MLSNÉ KOZY


Stopkova Plzeňská Pivnice


Restaurace na Vyhlídce aneb u Černé báby

Jste vlastníkem firmy? Zpětná vazba



Internet Mapy Obrázky Videa Zprávy Více Vyhledávací nástroje

Přibližný počet výsledků: 282 000 (0,43 s)

weather.com Zpětná vazba

Předpověď počasí - Brno - Meteocentrum.cz

www.meteocentrum.cz/predpoved-**pocasi**/cz/740/**brno**

Metadata v HTML

- Pomocí **<meta>** tagů:
  ```
  <meta name="keywords" content="HTML, CSS, XML" />
  ```
- Cílem je umožnit kvalitnější vyhledávání, než obyčejný full-text search
  - Zneužíváno ve velké míře spammery
- Neumožňuje definovat vztahy a hierarchie objektů
- Dnes vyhledávače dávají přednost jiným metodám, než prohledávání <meta> tagů

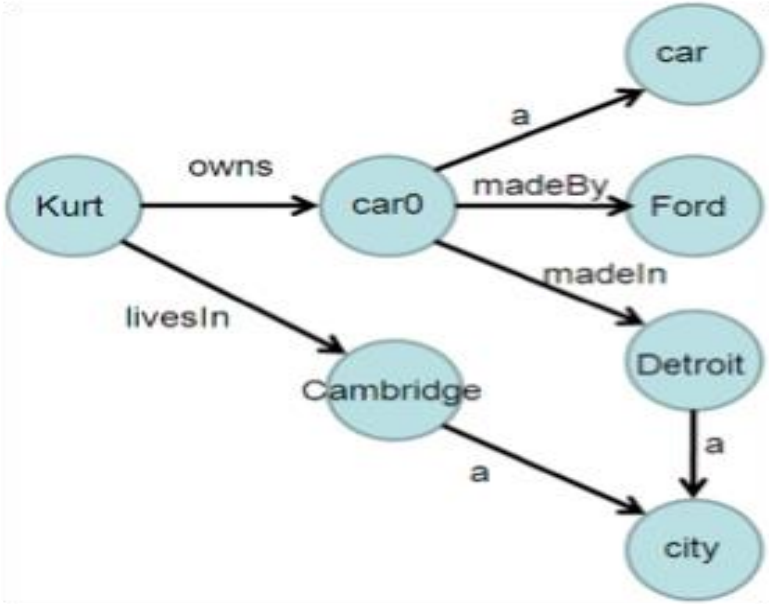
HTML 5

- Některé elementy v HTML5
  - <article>, <aside>
  - <nav>, <section>
  - <footer>, <time>
  - <video>, <audio>
  - <canvas>
  - …

RDF

- **RDF** = Resource Description Framework
- Framework pro popis zdrojů na webu
- Navržen tak, aby byl strojově čitelný a pochopitelný
- Doporučení W3C
- Různé způsoby serializace
  - Uložení do souboru
  - Př. **RDF/XML**

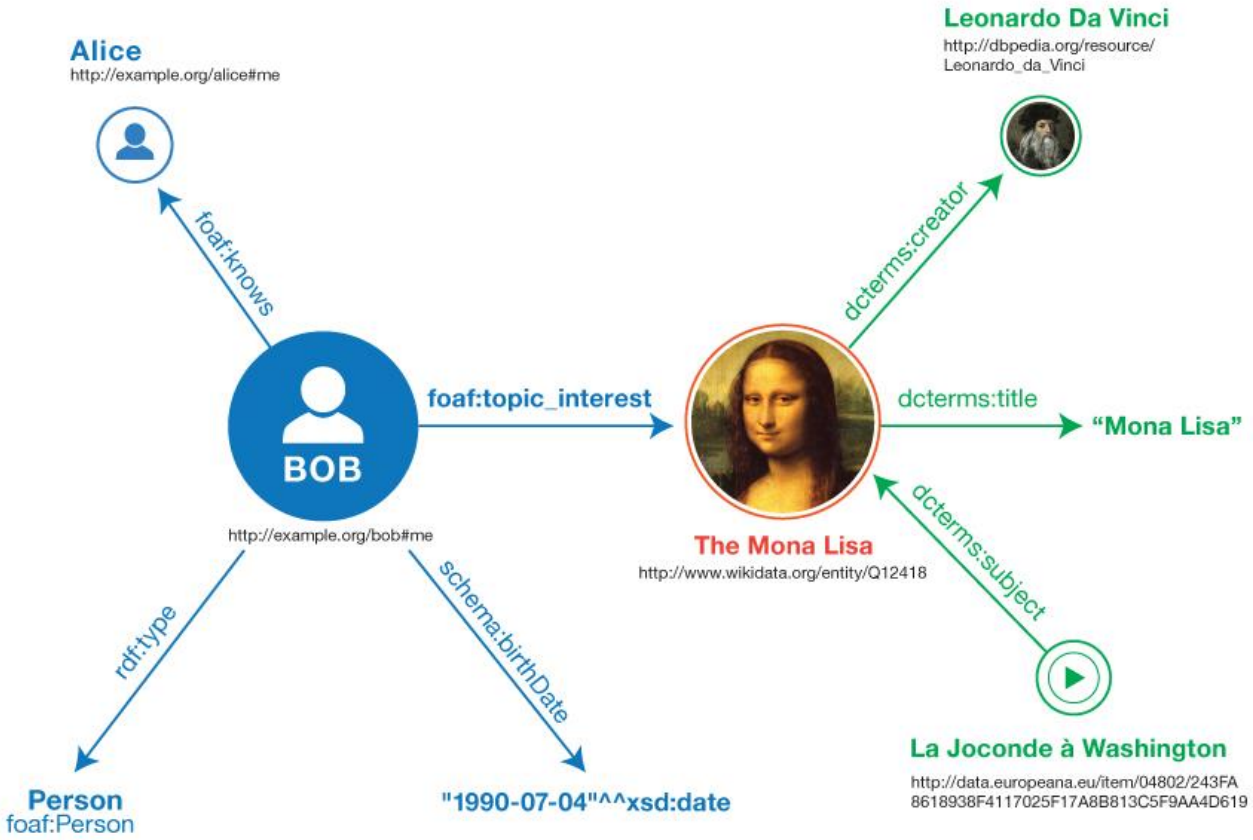
Princip RDF (1)



- Každému zdroji na webu přiřadí trojici:
  - Subject (subjekt, podmět)
  - Predicate (predikát, vlastnost)
  - Object (objekt, předmět)
- Při definici subjektů a predikátů je typicky potřeba definovat **URI** (Unique Resource Identifier) pro jednoznačné přiřazení významu

Princip RDF (2)

- RDF dokumenty lze ukládat do **triplestore** databází (databáze optimalizované pro RDF trojice) nebo serializovat pomocí XML (formát **RDF/XML**)



Příklad - RDF/XML

- Příklad: „Obloha má modrou barvu.“
  - Podmět: „obloha“
  - Vlastnost: „mít barvu“
  - Předmět: „modrá“ („blue“)
- Serializace ve formátu RDF/XML:

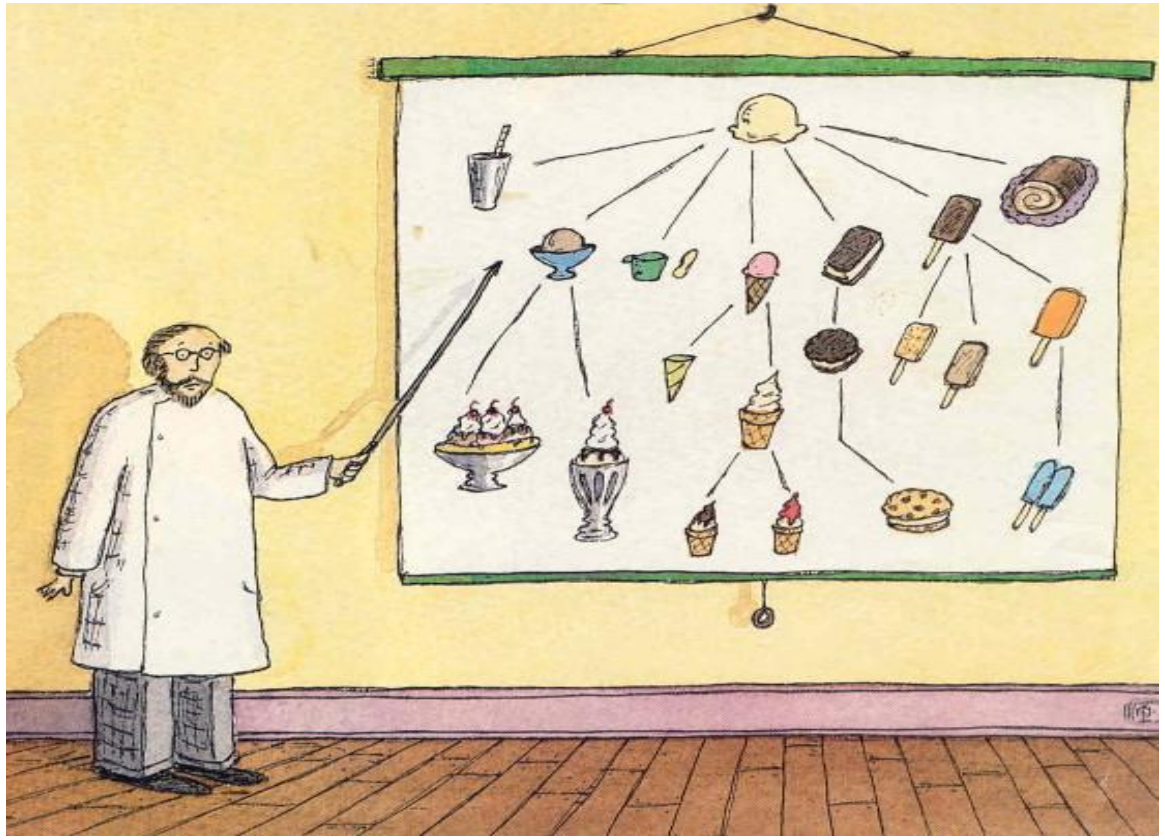
```
1: <?xml version="1.0"?>
2:
3: <rdf:RDF
4:     xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
5:     xmlns:sky="http://fi.muni.cz/rdf/sky/">
6:     <rdf:Description rdf:about="http://fi.muni.cz/rdf/sky">
7:         <sky:color>blue</sky:color>
8:     </rdf:Description>
9: </rdf:RDF>
```

Triplestores

- Databáze optimalizované pro ukládání RDF trojic (subjekt, predikát, objekt)
- Mnoho implementací v různých jazycích
  - C, C#, PHP, Java, Perl
- Postaveny buď nad existujícím relačním databázovým strojem (MySQL, PostgreSQL, MS SQL, Oracle), nebo vyvinuty kompletně od začátku přesně pro svůj účel (vyšší efektivita)

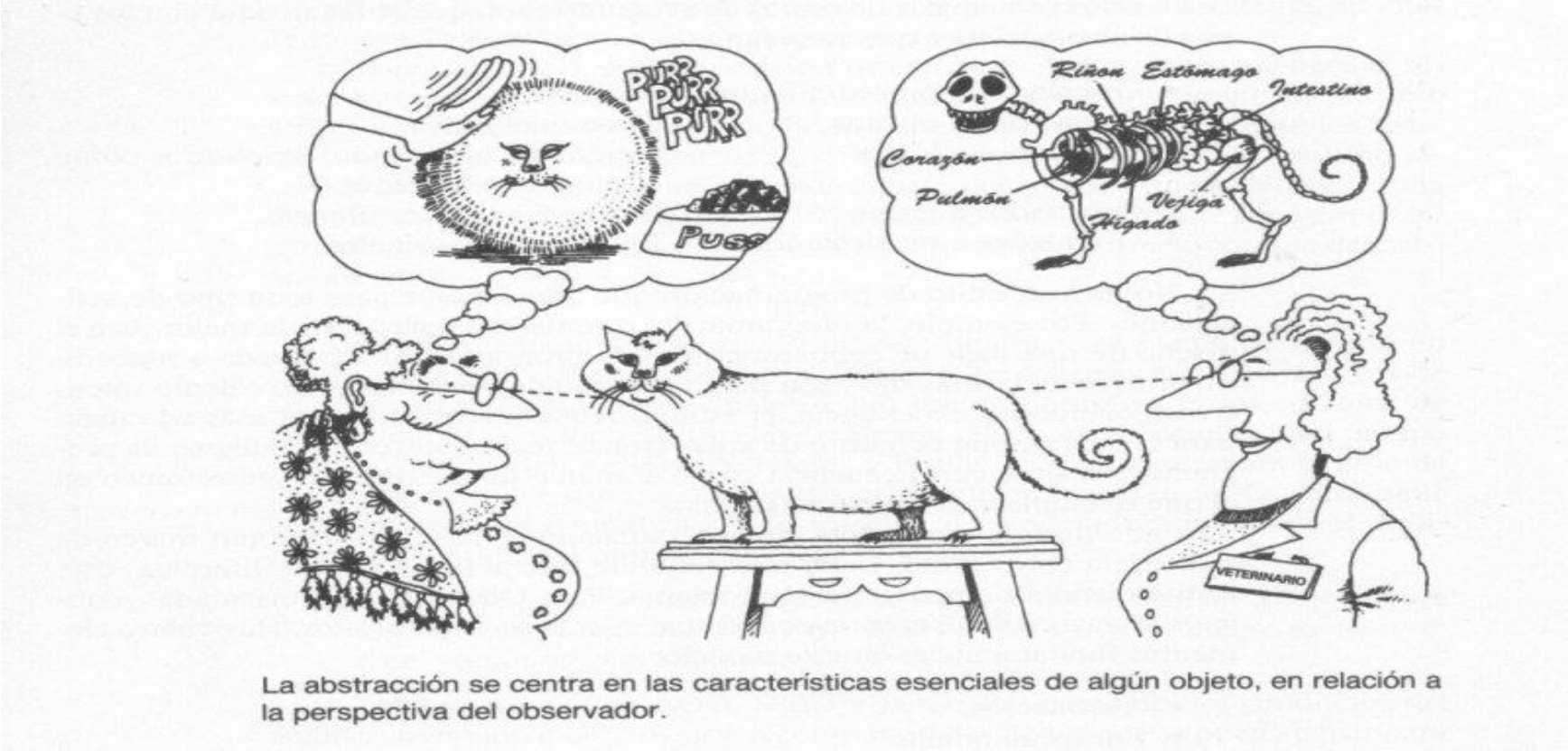
Ontologie (1)

- Model pro popis světa složeného z typů, vlastností a vztahů
- Je explicitní a formalizovaná
- Využití v sémantickém webu pro přiřazení významu datům (tj. pro tvorbu metadatového modelu)
- Při tvorbě ontologií je snaha o co nejpřesnější podobnost mezi objekty reálného světa a vlastnostmi modelu



Ontologie (2)

- Opravdu každý vidí svět stejně?



La abstracción se centra en las características esenciales de algún objeto, en relación a la perspectiva del observador.

Kategorie ontologií

- **Individua** (instance a objekty)
- **Třídy** (množiny, kolekce, pojmy, typy, druhy)
- **Atributy** (aspekty, stavy, vlastnosti, charakteristiky a parametry, kterých mohou objekty/třídy nabývat)
- **Relace** (způsoby, jakými k sobě mohou třídy a individua navzájem patřit)
- **Funkční výrazy** (komplexní struktury nad relacemi)

Kategorie ontologií

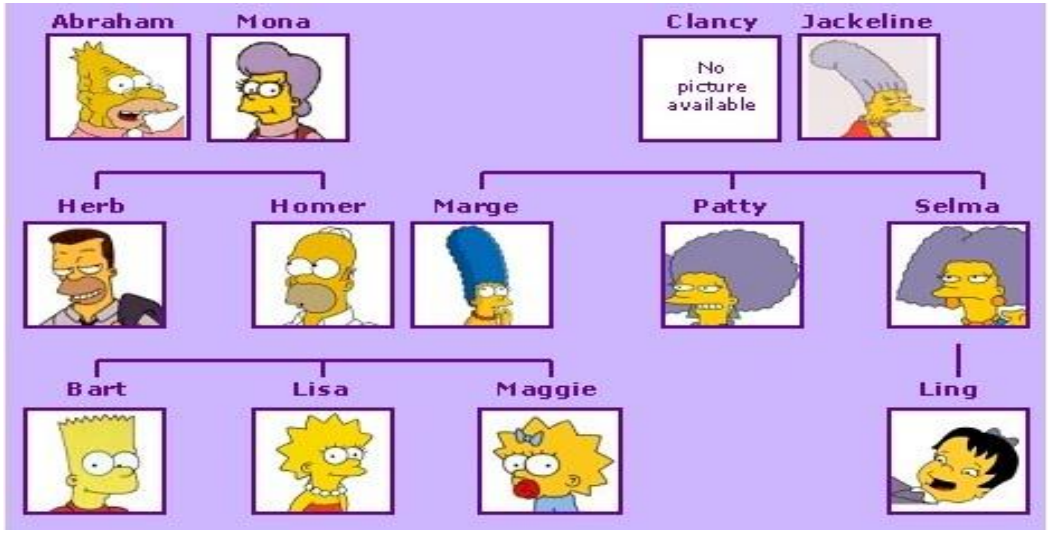
- **Restrikce** (formální popis platného vstupu)
- **Pravidla** (Příkazy ve formě if-then (příčina-následek) popisující logické inference, které mohou být odvozeny z výroků v dané formě)
- **Axiomy** (výroky (vč. pravidel) v logické formě, které dohromady skládají kompletní teorii, kterou ontologie popisuje. Nemusí obsahovat pouze apriorní znalosti, ale také odvozené teorie z jiných axiomů)
- **Události** (změny atributů a relací)

Inference znalostí

- Pojem **inference**
  - 1) dobře navržená logická heuristika pro odvozování nových znalostí
  - 2) odvozená znalost
- **Inference znalostí** - odvozování nových znalostí na základě existujících (známých) znalostí (inferencí)
- Využití v sémantickém webu při **strojovém vyhledávání** nových znalostí

Inference pomocí definovaných pravidel

if hasFather(?x, ?y) → ?y is Man
and if hasSister(?y, ?z) → ?z is Women
↓
then hasAunt(?x, ?z)

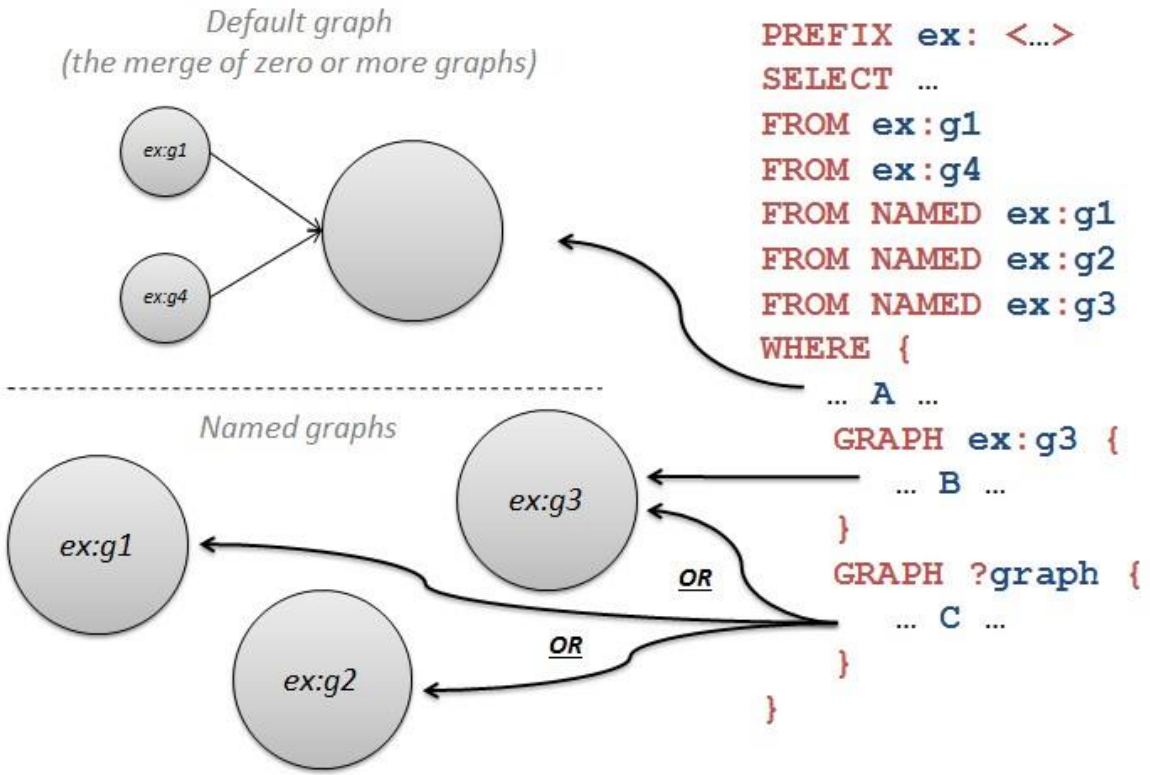


Inferenční stroje

- Počítačové programy, které zkouší odvodit odpověď **z báze znalostí** (knowledge base, množina axiomů/výroků/faktů/znalostí/popř. inferencí)
- Data v bázi znalostí musí být uložena takovým způsobem, aby stroj/engine dokázal odvodit a porozumět jejich významu, tj. musí být explicitně vyjádřena jejich **sémantika** (samotná data musí být doplněna o **metadata**)

SPARQL [„spa:kl“]

- Jazyk / protokol pro inferenci znalostí z RDF dokumentů
- Umožňuje provádět dotazy nad RDF trojicemi (triplestore databázemi)
- Podobná syntax jako SQL
- Výhoda SPARQL
  - Dotazy jsou díky přítomnosti URI v RDF formátu globálně jednoznačné



Příklad SPARQL

- Dotaz ke zjištění jmen všech osob pocházejících ze Stockholmu má následující formát

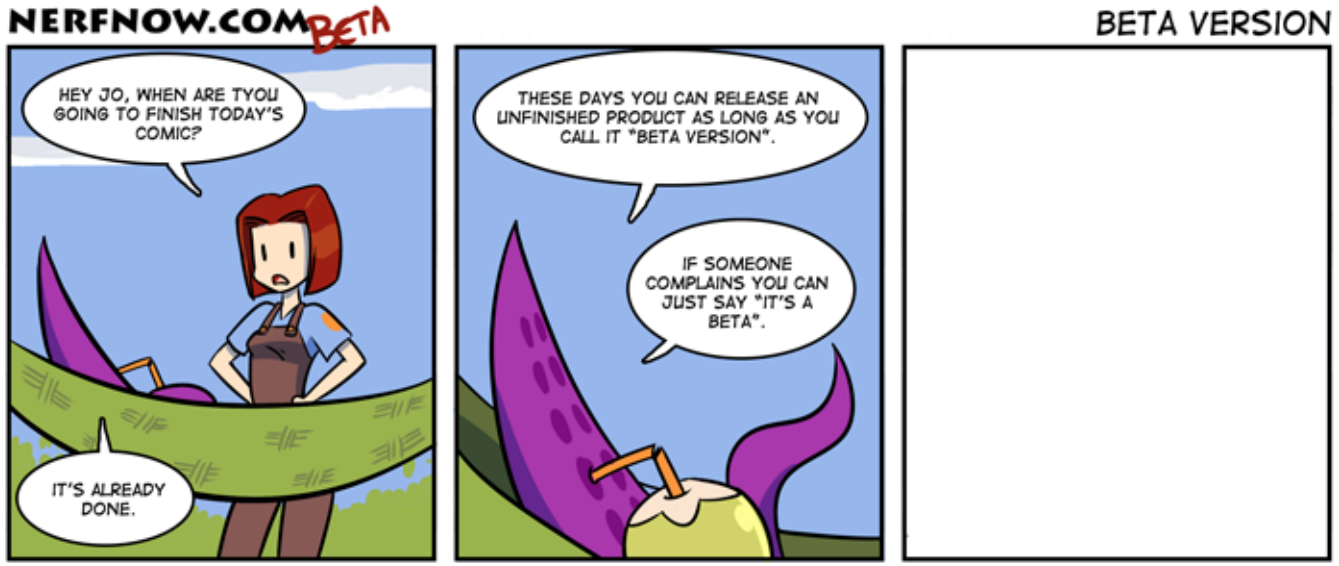
```
SELECT ?name FROM <example>
WHERE {
        ?x example:name ?name;
        ?x example:city „Stockholm“
}
```

Crowdsourcing

- Co to vlastně je?
- Jeff Howe v červnu 2006
  - Článek pro časopis Wired
  - „Outsource work to the crowd“
- Rozdíl mezi crowdsourcingem a obyčejným outsourcingem je v tom, že úloha nebo problém jsou zadané veřejnosti

Příklady crowdsourcingu

- Wikipedia
- Beta verze her
- Crowdfunding
  - Kickstarter, HitHit, Startovač
- Amazon's Mechanical Turk
  - Hledání ztracených osob
  - Sociologické výzkumy
- Netflix Prize
- InnoCentive
  - Velké firmy zveřejní problém
  - Honorář za jeho řešení



Výhody crowdsourcingu

- Lze ušetřit náklady
  - Oproti zaměstnaní vlastních zaměstnanců
  - Nákupu profesionálních služeb
- Platí se za výsledek
- Organizace může zachytit více talentovaných lidí, kteří jsou pro ni ochotni pracovat
  - Lze vytipovat potenciální pracovníky bez náročného výběrového řízení

Problémy crowdsourcingu

- Levná pracovní síla produkuje méně spolehlivé výsledky
  - V porovnání s profesionály
- Použití u těžších úkolů je riskantní
- Nutné organizovat mnoho pracovníků
  - Časově náročné
  - Řízení týmu místo řešení problému
- Možná soutěživost mezi pracovníky
- Příživníci
- Nemáte žádnou smlouvu s pracovníky
  - Mohou kdykoliv odejít
  - Znovupoužití svých výsledků/nápadů

Konkrétní příklad

- Chci se pustit do práce na automatickém vyhledávání objektů v obrázku a generování jejich popisu
- Mám (téměř) neomezené zdroje
- Problém
  - Potřebuji data k natrénování algoritmu
- Co udělám?
- Vytvořím hru Google Image Labeler
  - Dva lidé vkládají klíčová slova
  - Získávají body při shodě
    - Různě vysoké skóre dle specifičnosti klíčového slova
  - Některá klíčová slova mohou být odhalena