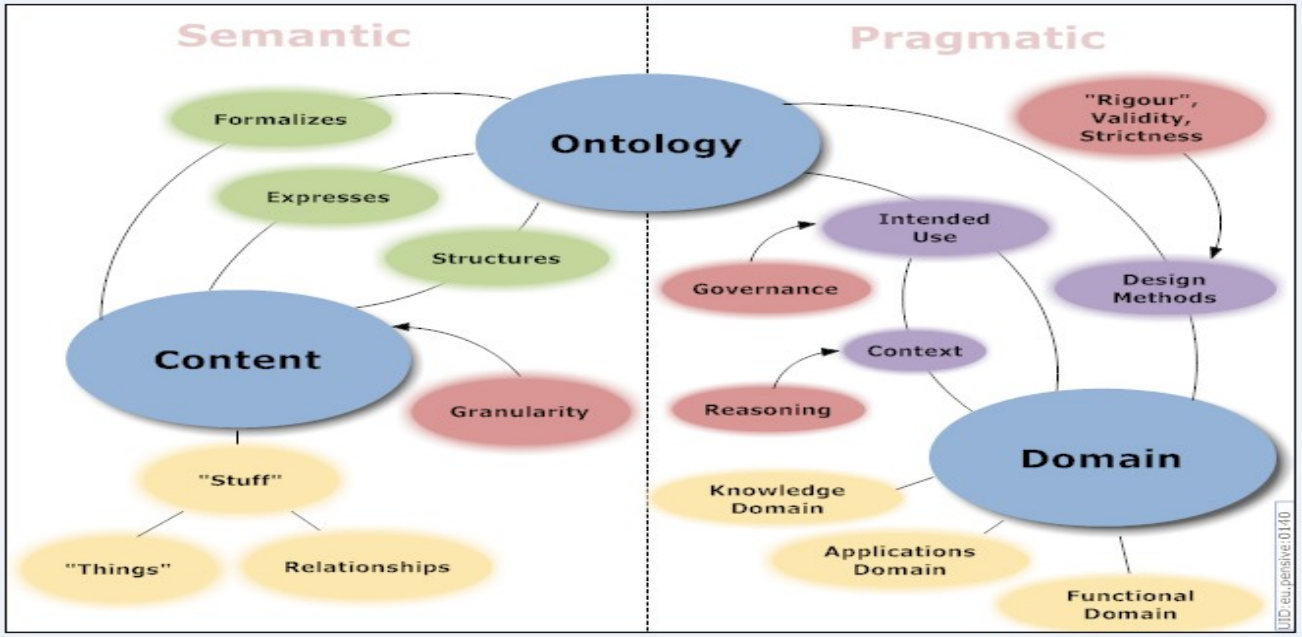
Sémantický web, ontologie, digitální knihovny



esf evropský sociální fond v ČR
EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Sémantický web

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

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**

Metadata v HTML

- Pomocí **<meta>** tagů:
  ```
  <meta name="keywords" content="HTML, CSS, XML" />
  ```
- Cíl: 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

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

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

RDF/XML

- Příklad: „Obloha má modrou barvu.“
  - Podmět: „obloha“ (http://fi.muni.cz/rdf/sky)
  - Vlastnost: „mít barvu“ (http://fi.muni.cz/rdf/sky/color)
  - 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

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

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

Inferenční enginy

- 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

Ukázky SPARQL

- „Vyhledej jméno a email všech lidí na světě“:

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
SELECT ?name ?email
WHERE {
    ?person a foaf:Person.
    ?person foaf:name ?name.
    ?person foaf:mbox ?email.
}
```

Ukázky SPARQL

- „Vyhledej všechna hlavní města Afriky“:

```
PREFIX abc: <http://example.com/exampleOntology#>
SELECT ?capital ?country
WHERE {
  ?x abc:cityname ?capital ;
     abc:isCapitalOf ?y .
  ?y abc:countryname ?country ;
     abc:isInContinent abc:Africa .
}
```

Sociální sítě

- propojená skupina lidí, kteří se navzájem ovlivňují
- **Sociální software** (**socioware**) - software, který umožňuje tvořit komunity pomocí počítačových propojení.
- **Virtuální komunita, e-komunita**

  Periferní (tj. lurker – *číhající*) - externí, nestrukturovaná účast

  Příchozí (tj. nováček) – nově příchozí je vpuštěn do komunity a může se plně účastnit diskuze

  Zasvěcenec (tj. stálý člen) – plně uznaný účastník

  Strážce hranic (tj. vůdce) – podporuje členství a zprostředkovává interakce

  Odchozí (tj. starý) – proces opouštění komunity kvůli novým vztahům, novým místům, novým vyhlídkám

Sociální sítě

- **Facebook**
- **MySpace** – sdílení hudby a videa
- **Orkut –** sdílení multimédií, chatování a hledání ztracených přátel.
- **Classmates (**Spolužáci.cz)
- **Blackplanet -** síť určená pro Afroameričany a jejich přátele
- **Hi5, Friendster, Bebo, …**

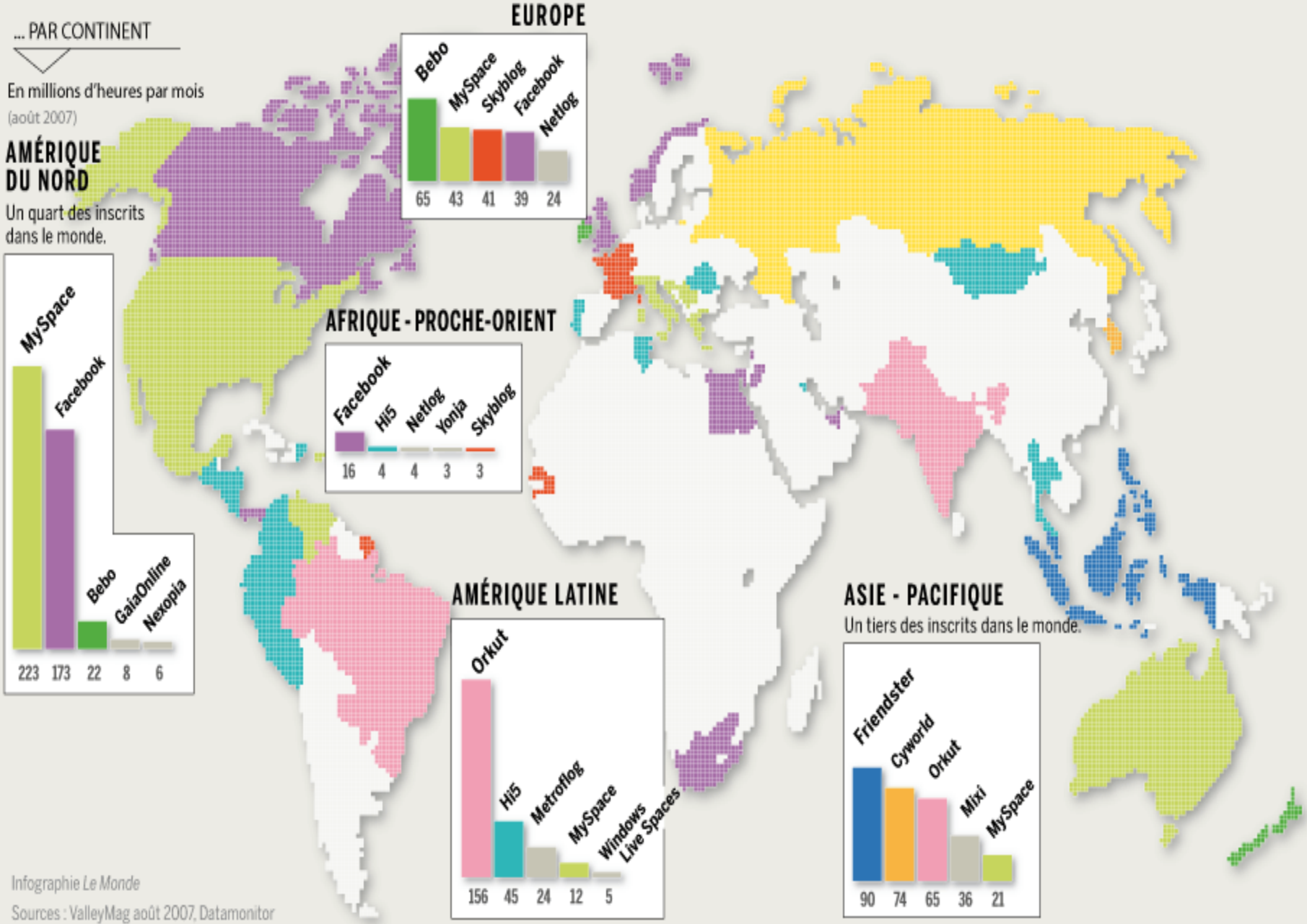
LES SITES DE RÉSEAUX SOCIAUX LES PLUS FRÉQUENTÉS...

... PAR PAYS

Nom du site	MySpace	Facebook	Bebo	Cyworld	Skyblog	Hi5	Friendster	Orkut	Live Journal
Nationalité de l'entreprise :	Etats-Unis	Etats-Unis	Etats-Unis	Corée du Sud	France	Etats-Unis	Etats-Unis	Etats-Unis	Russie

... PAR CONTINENT

En millions d'heures par mois
(août 2007)



Infographie *Le Monde*

Sources : ValleyMag août 2007, Datamonitor

Sociální aspekty

- Internet jako svět, kde se adolescenti učí komunikovat, zacházet s jazykem, vyjádřit sami sebe.
- Internet jako příležitost pro sociálně handicapované.
- Internet jako příležitost pro vznik nedorozumění.
- Internet jako příčina odloučení z běžného života, ze společnosti

Internetová generace

- Roste internetová generace?
- Percepce virtuální komunikace závisí výrazně na věku.
- Existuje skupina lidí, která upřednostňuje virtuální komunikaci před komunikací tváří v tvář. Tato skupina:
  - se na internetu méně bojí komunikovat s autoritami.
  - se domnívá, že na internetu lépe vyjadřuje své pocity.
  - udává, že vyjadřování na internetu je pro ni snazší než v běžném hovoru.

Preferují lidi virtuální komunikaci?

- „Raději potkávám lidi na internetu než osobně“ – souhlasí 11,6% populace
- Není rozdíl mezi muži a ženami (pouze velký rozdíl u nejmladších)
- Opět rozdíly podle věku – čím mladší, tím spíše preferují virtuální komunikaci (24% 12-15 let, 14,1% 16-20 let, 12,4% 21-30 let, 9,3% průměr 31 let a více)

Časový rámec

- Typ komunikace – synchornní x asynchornní atd.
- 2 protichůdné pohledy:
  - Více času -*„Na internetu bych třeba měla i čas na rozmyšlenou, neříkám, že teď nemám, ale asi bych nad tím více přemýšlela. Třeba až odtud odejdu, tak si řeknu že jsem to tady mohla podat jinak.“*
    - Lepší schopnost formulovat, vyjádřit se
  - Méně času *„V realitě si to můžu třeba trošku v hlavě zesumírovat, třeba můžu být chvilku ticho... Tam je ale ta rychlost, tam musím reagovat bezprostředně.“*
    - Omezená slovní zásoba
- Tolerance delší latence
- 1 hodina reálná ~ 3-4 hodiny virtuální

Časový rámec – trocha statistiky

- 61,8% adolescentů souhlasí „Na internetu se často baví s více různými lidmi zároveň“
- 50,1% adolescentů souhlasí: „Na internetu si často zároveň povídám o úplně odlišných tématech s více lidmi najednou.“
- „Vyjadřovat se na internetu je snazší než v běžném hovoru“ – souhlasí 32,2% populace, 12-20 let více souhlasí dívky (48,7% ku 42,8%)

Prostor a prostorové uspořádání

- Sdíleným prostorem je typicky monitor
- Absence neverbálního – dalekosáhlé důsledky
  - „Nevyplyne to tak, jak když komunikuješ z očí do očí. Z očí do očí člověka líp poznáš, z osoby něco vyzařuje ale z textu na netu ne.“
  - Princip projekce
- Výhoda nebo nevýhoda?
- Multiplicita komunikace (viz dále)

Multiplicita komunikace

- Souběžná synchronní online komunikace s více lidmi na více témat
- *„Dobří netisti zvládají tak 5 lidí. A to ještě tak, že jednoho mají vjednom okně, druhýho v druhým okně a tři na icq. Takže to je taky jiný, jako bys seděl zároveň ve třech hospodách. Ve třech různých hospodách.“*
- Motivace ? *„...možná, že ti to dohromady dá větší efekt, než kdybys seděl jenom na jednom místě. Né efekt toho kdybys seděl na třech místech, ale trošku větší než kdybys seděl na jednom místě.“*
- *„Ten pocit je supr, jenom musím furt psát ... převládá pocit, že nestíhám.“*
- *„Je to opojný, cítíš se prostě skvěle ... Seš středem pozornosti ... to je stav kdy ti většina/všichni píšou až to nezvládáš.“*

Digitální knihovny

- Kolekce vědomostí uložené v digitálním formátu a přístupné pomocí počítačů
- Systémy pro získávání informací
- Přístupné on-line nebo lokálně na CD/DVD (encyklopedie)
- Obsah digitálních knihoven může vzniknout digitalizací (např. skenování, OCR) nebo může být vytvořen přímo v digitální podobě (elektronické dokumenty)

Vyhledávání v digitálních knihovnách

- Vyhledávání v obrovských objemech dat (vysoké nároky na **výkon**)
- **Distribuované vyhledávání** (klient dotáže několik serverů naráz, které pak vyhledávají **simultánně**)
- V současné době lze prohledávat data většinou pouze **full-textem** + podle omezeného množství metadat (název, autor dokumentu, apod.)
- Cíl: umožnit vyhledávání podle dat, které spolu souvisí svým významem (pomocí RDF metadat)

Elektronické publikování

- Publikace elektronických dokumentů a tvorba digitálních knihoven
- Nemusí být nutně on-line (lze publikovat i na fyzických nosičích CD/DVD/Flash paměti, …)
- Nemusí jít výhradně o „nové znalosti“, lze digitalizovat i existující dokumenty
- V oblasti publikování vědeckých článků se publikuje prakticky výhradně v elektronické formě (rychlejší recenze od vědecké komunity)

Business modely e-publishingu

- On-line reklama
- Open-access (volné zveřejnění obsahu)
- Pay-per-view (placený obsah)
- Print-on-demand (tisk na objednávku)
- Subscriptions (předplatné)
- Self-publishing (vydání vlastním nákladem)

Nejdůležitější formy e-publishingu

- **CD/DVD** (i jiné fyzické nosiče)
- **E-book** (elektronická kniha, větš. zpoplatněno)
- **E-journal** (elektronický časopis, zpoplatněno)
- **Online newspaper** (webové noviny, větš. zdarma)
- **Online magazine** (webový časopis, větš. zdarma)
- **Sdílení souborů** (web, FTP, P2P sítě)
- **Podcast** + sdílení multimédií na webu (YouTube)
- **Groupware** (nástroje pro spolupráci určité skupiny lidí)