VIKMA06

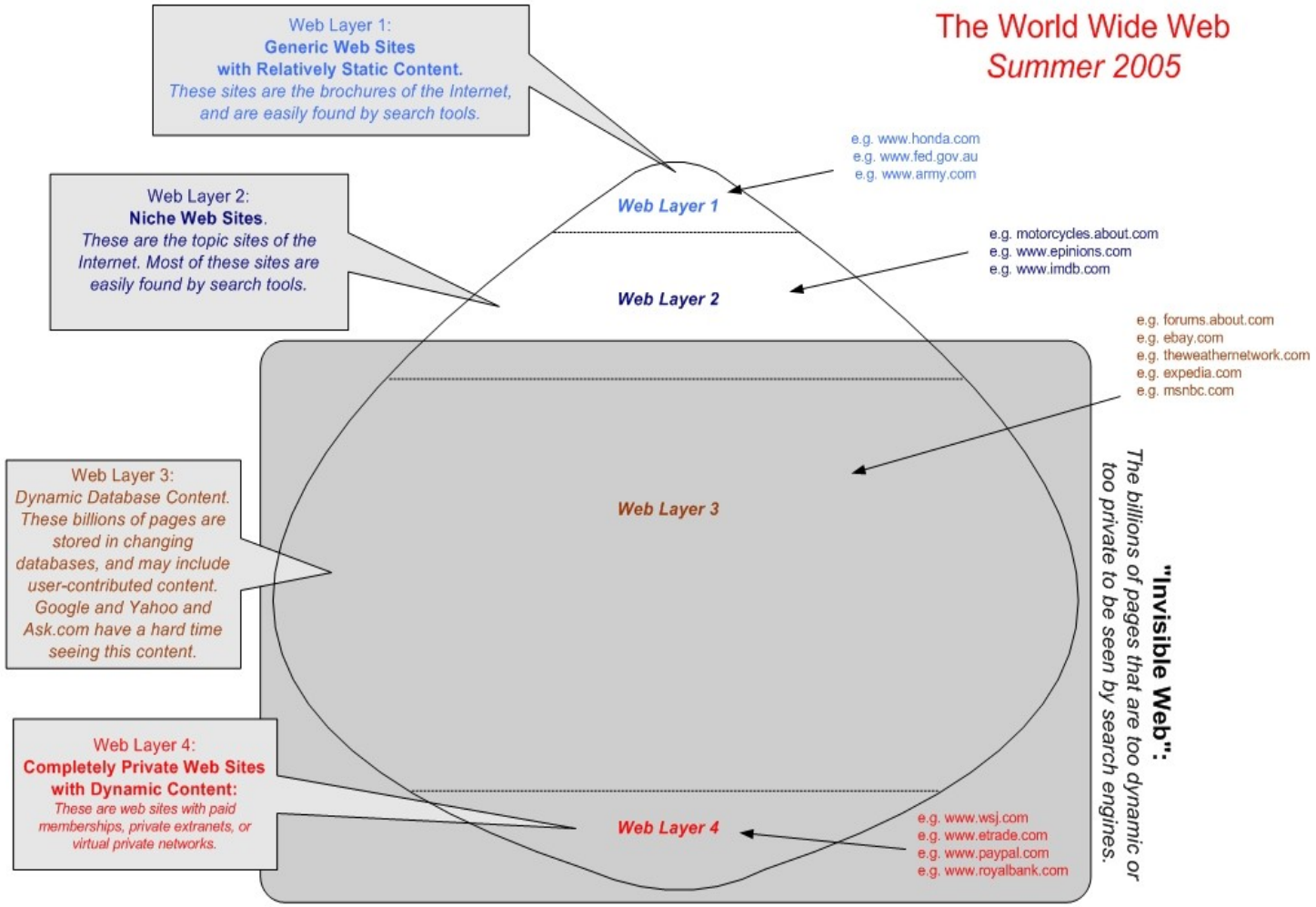
Rešeršní a studijně rozborová činnost

14. 5. 2010: Přednáška D8: Vyhledávání a internet

FF MU, jaro 2010

Mgr. Josef Schwarz
126172@mail.muni.cz

Neviditelný web



Typy „neviditelnosti“

- Nepřehledný web (Opaque web)
- Soukromý web (Private web)
- Vlastnický web (Proprietary web)
- Skutečně neviditelný web (Truly invisible web)

Nepřehledný web

Obsahuje soubory, které mohou být, ale z určitých příčin nejsou vyhledávači indexované.

Důvody:

- hloubka indexování (depth of crawling)
- frekvence indexování (zprávy, inzerce, ceny akcií)
- maximální počet viditelných výsledků
- odpojené stránky

Soukromý web

Obsahuje stránky, které by robot dokázal zaindexovat, ale správce webu to znemožňuje.

- stránky chráněné heslem
- soubor robots.txt
- metatagy „noindex“, „nofollow“

Vlastnický web

Část webu, ke které je přístup pouze po splnění určitých podmínek.

- stránky vyžadující souhlas s podmínkami pro vstup
- stránky dostupné po zaplacení poplatku

Skutečně neviditelný web

Stránky, které roboty neindexují kvůli svým technickým omezením.

- problém s formáty (sfw, exe apod.)
- dynamicky generované stránky
- relační databáze (Oracle, MS SQL Server, IBM DB2)

Přednosti hlubokého webu

- specializovaný obsah – komplexnější informace
- sofistikovanější uživatelské rozhraní
- větší důvěryhodnost
- oborovost

Přístup k hlubokému webu

- metavyhledávače
- specializované vyhledávače, katalogy, adresáře
- oborové (předmětové) vyhledávače, katalogy, adresáře
- referenční zdroje
- weby knihoven
- digitální a virtuální knihovny
- oborové databáze
- weby organizací
- knihy (archivy, e-books)
- blogy

Výběr vyhledávačů hlubokého webu

- Complete Planet
  - adresář více než 70 000 databází a specializovaných vyhledávačů
- GoshMe
  - vyhledávač vyhledávačů a databází (registrace)
- IncyWincy
  - vyhledávač hlubokého webu (vyhledávače a databáze)
- BUBL LINK
  - polytematický (DDC) katalog vybraných internetových zdrojů
- ResourceShelf
  - blog s informacemi a novinkami o informačních zdrojích (připravovaný informačními profesionály)

Sémantický web

klasický x sémantický web

- Tvořen tak, aby jeho obsahu porozuměl pouze člověk
- Citlivý na použitou terminologii
- Nalezených dokumentů je obvykle příliš mnoho nebo naopak příliš málo (případně žádné)
- Výsledkem vyhledávání je pouze jedna stránka

- Rozšíření klasického webu
- Obsah ve strojově přístupné formě
- Vyhledávání podle klíčových slov nahrazeno zodpovídáním dotazů
- Dotaz je možno zodpovědět na základě extrakce informací z více stránek

Klasická podoba webu

<h1>Agilitas Physiotherapy Centre</h1>
Welcome to the home page of the Agilitas Physiotherapy Centre.
Do you feel pain? Have you had an injury? Let our staff
Lisa Davenport, Kelly Townsend (our lovely secretary)
and Steve Matthews take care of your body and soul.
<h2>Consultation hours</h2>
Mon 11am - 7pm
Tue 11am - 7pm
Wed 3pm - 7pm
Thu 11am - 7pm
Fri 11am - 3pm<p>
But note that we do not offer consultation
during the weeks of the
State Of Origin games.

Web s explicitními metadaty

- *XML + XML schéma*
- *RDF + RDF schéma*

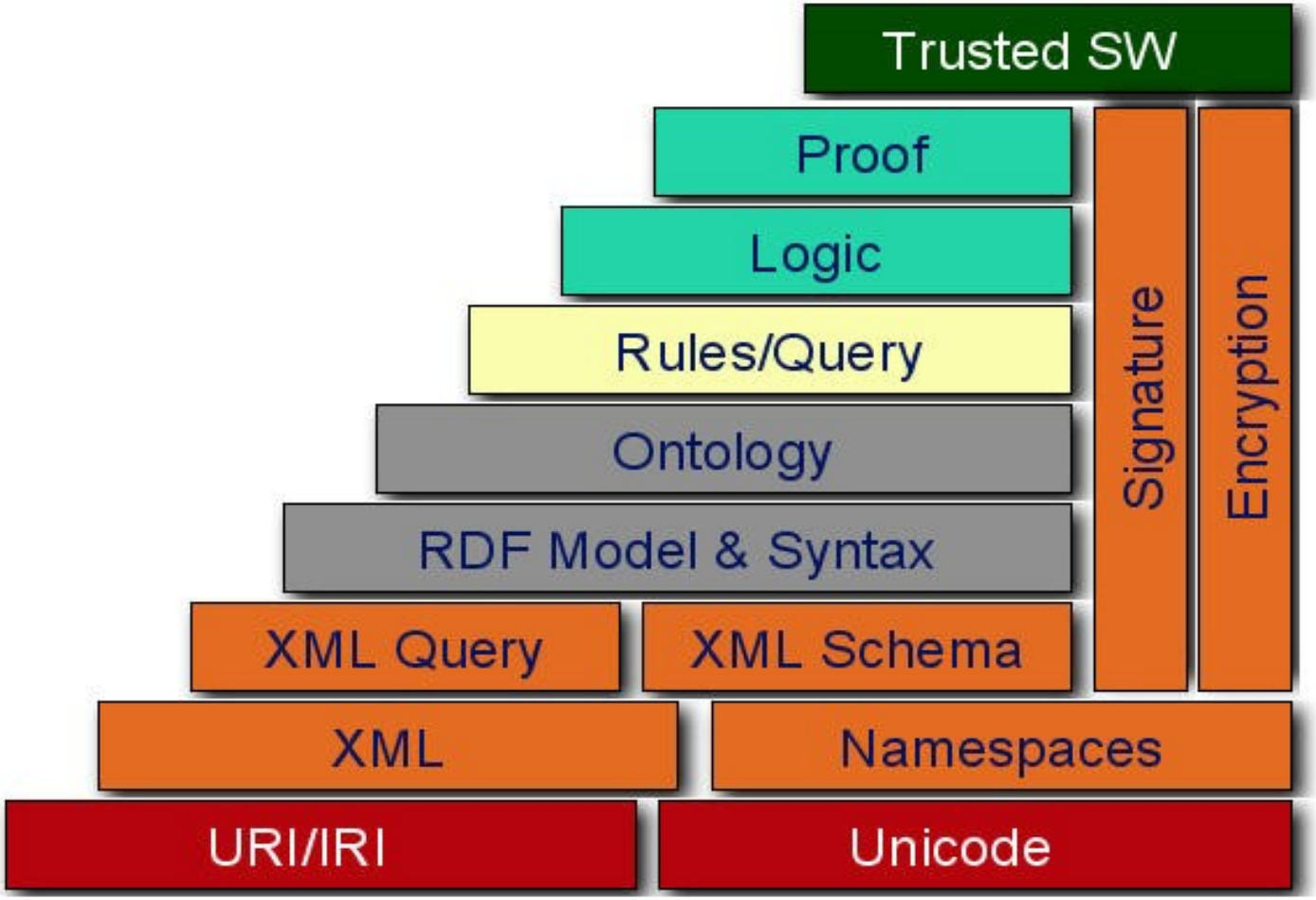
```
<company>
<treatmentOffered>Physiotherapy</treatmentOffered>
<companyName>Agilitas Physiotherapy
   Centre</companyName>
<staff>
<therapist>Lisa Davenport</therapist>
<therapist>Steve Matthews</therapist>
<secretary>Kelly Townsend</secretary>
</staff>
</company>
```

Sémantický web

- Základní složky (předpoklady) SW
  - strukturace dokumentů
  - vyjádření sémantiky - ontologie
  - vyhledávací nástroje - agenti

  - standardy
    - syntaktická složka
      - URI
    - strukturální složka
      - XML
    - sémantická složka
      - RDF + RDFS (schéma RDF)
      - OWL, OIL

Vrstvy sémantického webu



Sémantický web – příklady řešení

- W3C
- příklad aplikace RDF
  - energetika
- Výzkum
  - The Open University London, Knowledge Media Institut
    - Magpie
  - Stanford Knowledge Systems Laboratory
    - DAML (agenti)
  - EU, 5. rámcový program
    - On-to-knowledge

Pokročilé webové vyhledávání

- Pomocné metanástroje:
  - Thumbshots Ranking
- Vyhledávače
  - Ask.com
  - FyberSearch
  - Answers
  - Last.Fm