



Szemantikus Web: egy rövid bevezetés

2006. március 18

Szemantikus Web: egy rövid bevezetés

Ez az előadás a [Magyarországi Web Konferencia](#) keretében hangzik el 2006. március 18-án, Budapesten.



A Szemantikus Web felé...

- A jelenlegi Weben az információk különböző formákban állnak rendelkezésre:
 - *természetes nyelveken (angol, magyar, kínai, holland, ...)*
 - *grafikákon, képeken, audió és videó formákban...*
 - *stb.*
- Emberek számára ez nem jelent igazán problémát...

A Szemantikus Web felé... (folyt.)

- A Weben gyakran van szükség adatok *kombinálására*
 - *a szállodai és az utazási adatok általában különböző forrásból származnak, habár együtt akarjuk őket használni*
 - *valamely kutatásnak különböző digitális könyvtárak anyagaira van szüksége*
 - *stb.*
- Ezt is könnyedén megtesszük; egy fogalomról másikra asszociálni nekünk egyszerű...

De...

- a gépek buták!
 - *résztleges információt nem tudnak használni*
 - *a képek értelmezése még mindig komoly kutatás tárgya*
 - *analógiákat nehezen tudnak automatikusan megtalálni*
 - *az adatok kombinálása is nehézkes*
 - ugyanaz-e az *<abc:alkotó>* mint az *<cba:író>*?
 - ...

Gyakorlati példa: keresés

- A legtöbbet emlegetett példa...
 - *a Google és társai csodálatos eszközök, de túl sok a hamis találat*
 - *segítséget jelenthet, ha az adatforrásokhoz valamilyen további (esetleg alkalmazásfüggő) leírást lehetne hozzárendelni*

Gyakorlati példa: utazásszervezés

- Egy automatikus utaztató rendszer, amely
 - *ismeri a szokásaimat, kívánalmaimat*
 - *a múlt alapján további tudást alakít ki rólam*
 - *a helyi információt össze tudja kombinálni távoli információkkal, mint például:*
 - légitársaság adataival
 - orvosi kérdésekkel, mint diétával, gyógyszerek hozzáférhetőségével
 - naptáradatokkal, állami vagy vallási ünnepek adataival
 - stb.
- A rendszer *távoli* információkat kombinál a Weben
- (lásd M. Dertouzos: Félkész forradalom)

Gyakorlati példa: adatbázisok integrációja

- Adatbázisok struktúrája, tartalma nagyon különböző lehet
- Sok alkalmazás alapul adatbázisok kombinációján:
 - *cégösszeolvadások*
 - *biokémiai, orvosi, genetikai adatok*
 - *kormányzati és adminisztratív adatok*
- Ezek az adatok legtöbbször a Weben vannak már (habár nem feltétlenül nyilvánosak)
- Az adatok, adatbázisok *szemantikáját* kell ismerni ahhoz, hogy kombinálhatók legyenek (az, hogy a szemantika hogy képződik le a konkrét adatbázisra, voltaképpen mellékes)

Mire van szükség?

- Az adatokat gépi kezelésre is elérhetővé kell tenni
 - *egyres esetekben az adat más adatokat ír le (mint a keresés esetén): ezek az ún. metaadatok*
 - *máskor magát az adatokat kell kombinálni, például a naptáram vagy utazási szokásaim esetén*
- Az adatokat össze kell tudni olvasztani, kombinálni, és mindezt a Web nagyságrendjén
- A gépeknek *következtetéseket is le kell tudnia vonni* az adatokról (például hogy a használt terminológia azonos szemantikát takar...)

Mire van szükség (technikailag)?

- Mindehhez szükség van:
 - az erőforrások egyértelmű elnevezésére: *URI*
 - az adatok összekapcsolására, leírására szolgáló általános modellre: *RDF*
 - az adatok a modell alapján való elérésére: *SPARQL*
 - a közös szóhasználat definíciójára: *RDFS, OWL, SKOS*
 - következtetési rendszerekre: *OWL, Rules*
- *A szemantikus Web célja egy olyan infrastruktúra létrehozása, amely lehetővé teszi a Weben lévő adatok integrálását, a közöttük levő kapcsolatok definiálását és jellemzését, illetve az adatok értelmezését*

RDF hármások

- Az adatok „összekapcsolásáról” beszéltünk... vagyis az adatokat („erőforrásokat”) *egymáshoz kell rendelni*
- Egy egyszerű hozzárendelés nem elegendő... a hozzárendelést *el kell nevezni*
 - *egy hozzárendelés a naptáramhoz nem ugyanaz mint az önéletrajzomhoz: az első ki kell hogy fejezze, hogy „naptaam”, míg a második azt, hogy „önéletrajzom”*
- Innen származnak az RDF hármások: *két erőforrás közötti címkézett kapcsolat*

RDF hármások (folyt.)

- Egy RDF hármás (s,p,o):

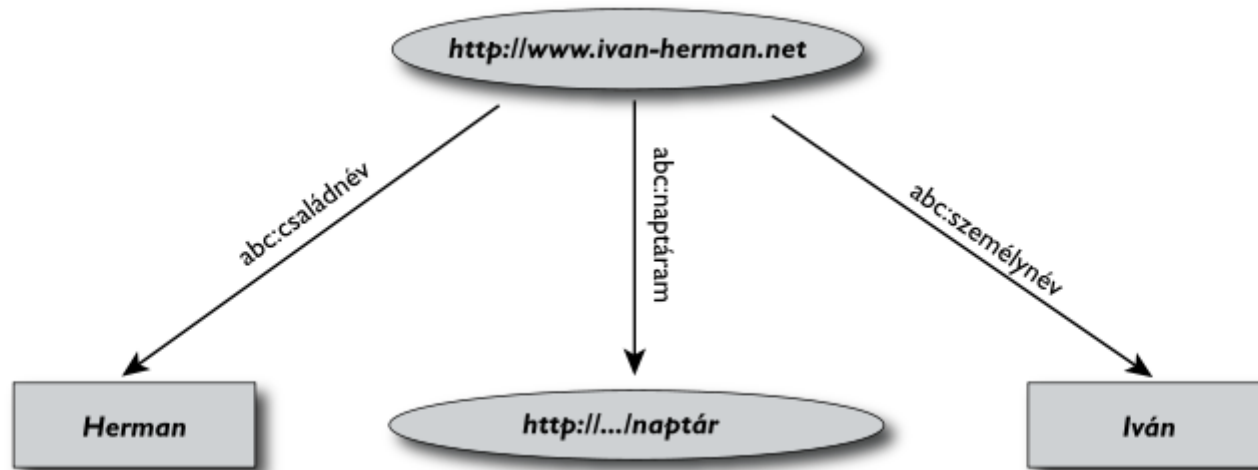
- „s”, „p” URI-k; „o” egy URI vagy egy literál
- jelentése: a „p” összekapcsolja az „s”-t az „o”-val
- az elnevezések/cimkék eszközei szintén a URI-k: <http://.../naptaram>
- íme a teljes hármás:

(<http://www.ivan-herman.net>, <http://.../naptaram>, <http://.../naptar>)

- *RDF* a hármások általános modellje: lényegében egy *irányított, címkézett gráf*

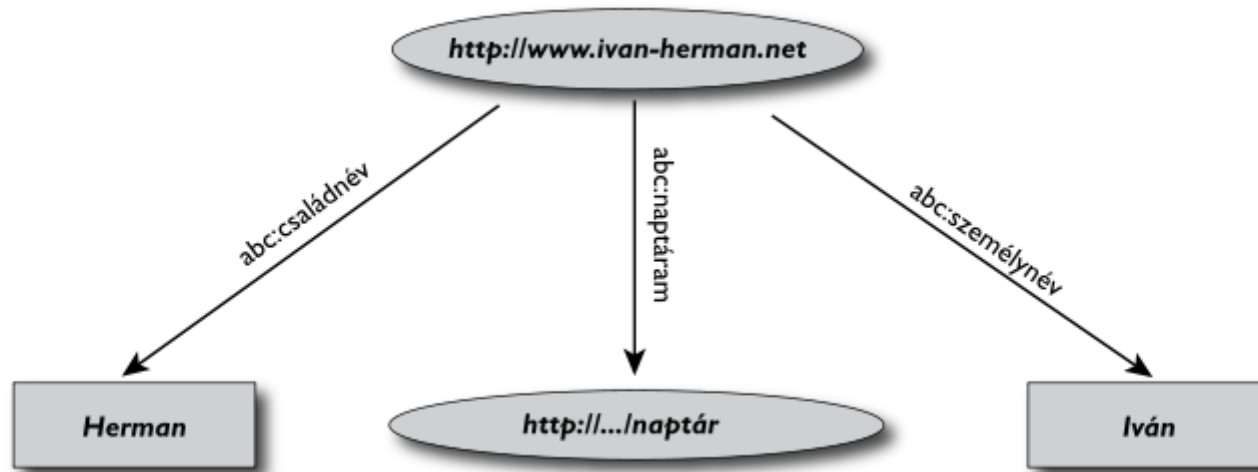
- *gépileg olvasható formátumokkal (RDF/XML, Turtle, n3, RXR, ...); RDF/XML a „hivatalos”, XML alapú formátum*

Egy egyszerű RDF példa (RDF/XML)



```
<rdf:Description rdf:about="http://www.ivan-herman.net">
  <abc:családnév>Chart</abc:családnév>
  <abc:naptáram rdf:resource="http://.../napár"/>
  <abc:személynév>Iván</abc:személynév>
</rdf:Description>
```

Egy egyszerű RDF példa (Turtle)



```
<http://www.ivan-herman.net>  
  abc:családnév "Chart";  
  abc:naptáram  <http://.../naptár>;  
  abc:személynév "Iván".
```

RDF hármások (folyt.)

- *Bármely* URI használható; vagyis egy XML fájl-ba is lehet címezni, nemcsak a teljes anyagra, pld:
 - `http://www.example.org/file.xml#xpointer(id('naptár'))`
 - `http://www.example.org/file.html#naptár`
- Az angol terminológia:
 - „*triplets*”, „*triples*”, vagy „*statement*”
 - magyarul: „hármás”, vagy „állítás”
 - „*subject*”, „*predicate*” vagy „*property*”, „*object*”
 - magyarul: „alany”, „állítmány” vagy „tulajdonság”, és „tárgy”

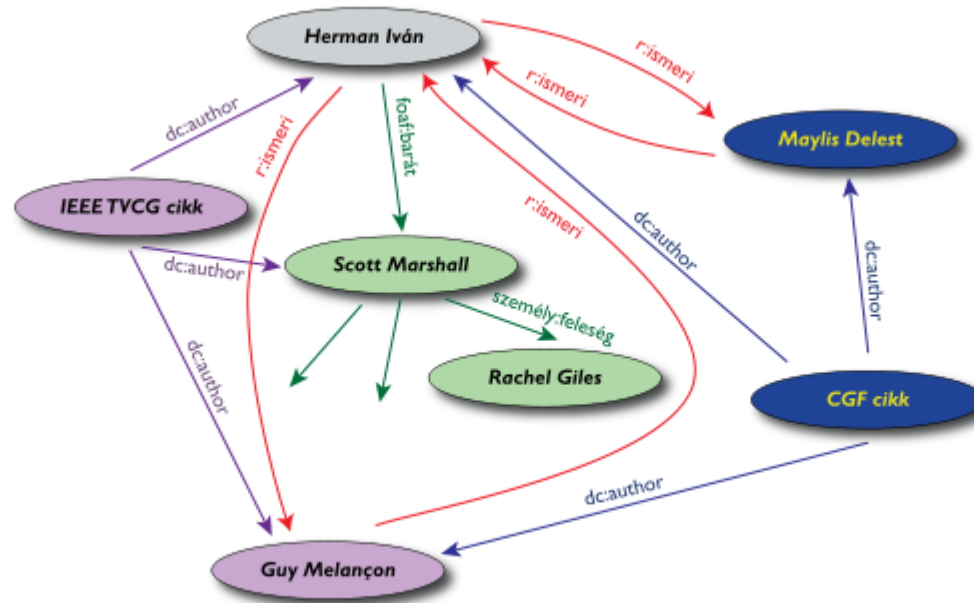
A URI-k alapvető szerepe

- *Bárki* kreálhat (meta)adatot *bármely* Web–erőforrásról
 - *pld. ugyanazt az XML–alapú állományt le lehet írni egymástól eltérő terminológiákkal*
 - *az URI-k teszik lehetővé adatok egymáshoz kapcsolását*
- *A URI-k ágyazzák az RDF-et a Webbe*
 - *így lesz a „Szemantikus Web”... „Szemantikus Web”*

URI-k: összevonás

- Könnyűvé válik az adatok (logikai) összevonása
- Az összevonás megtehető az *azonos* URI-k alapján
 - *egy gráfban: azonos URI-val rendelkező csomópontok egymással azonosíthatóak*
- Ez az összevonás az RDF–modell *nagyon* fontos jellemzője
 - *a leírásokat különböző személyek, csoportok hozhatják létre, de ...*
 - *...az alkalmazás egységként kezelheti őket*
 - *egyike azon területeknek, ahol az RDF–modell sokkal könnyebben használható, mint az XML*

Példa az összevonásra...



Az RDF nem elegendő...

- A kapcsolatok létrehozása és programból való használata működik, feltéve, hogy a program *tudja*, hogy milyen terminológiát használhat!
- Például használtuk a következő fogalmakat:
 - *naptáram, családnév, személynév, ...*
- Ismertek-e ezek? Korrektek-e? (A probléma egy kicsit hasonló egy adatbázis rekordtípus definiálásához)

Megoldandó kérdések

- Mely terminológiák, szavak használhatók? Ismert-e a terminológia?
- Korrekt módon használjuk-e a tulajdonságokat? Van-e értelmük az adott erőforrások esetén?
- Lehet-e következtéseket levonni? Például:
 - *„ha »A« »B«-től balra van, »B« »C«-től balra van, akkor balra van-e »A« »C«-től?”*
 - *nekünk nyilvánvaló, de egy programnak nem ...*
 - *... vagyis: levonhatják-e a programok ezeket a következtetéseket?*
- Ha valaki más definiál egy állításhalmazt: ugyanaz-e, mint a mienk?

Ontológiák

- A Szemantikus Webnek szüksége van *ontológiákra*:

„egy adott tudásterület leírására használt fogalmak és összefüggések definíciója”

- Szükség van egy *Webontológia nyelvre*, amellyel definiálni lehet:
 - *az adott kontextusban használható fogalmakat*
 - *a tulajdonságokra érvényes korlátozásokat*
 - *a tulajdonságok logikai jellemzőit*
 - *a fogalmak és tulajdonságok ekvivalenciáját (vagy különbözőségét)*
 - *stb*
- Az erre szolgáló specifikációk: RDFS (RDF Sémák) és OWL (Webontológia Nyelv)

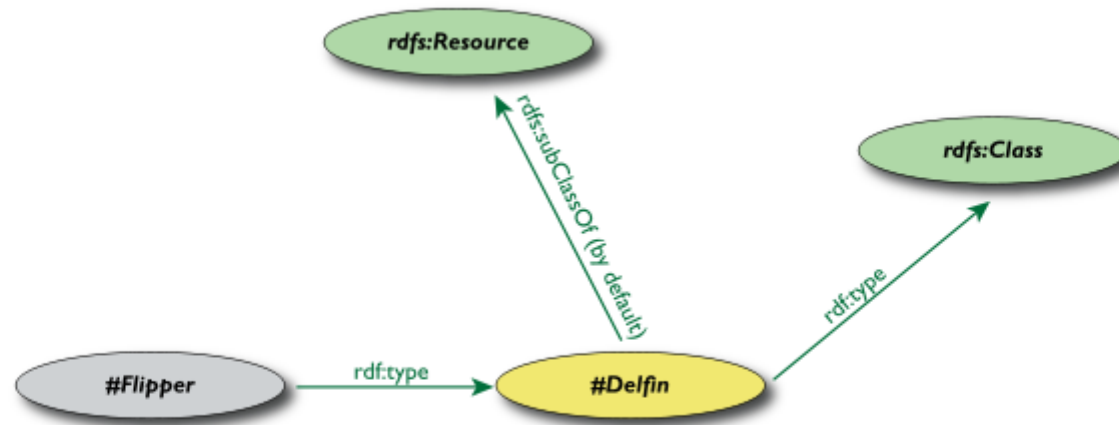
Osztályok, erőforrások ...

- Gondoljunk az ismert, tradicionális ontológiákra:
 - ismerjük az „*emlős*” fogalmát (valahonnan)
 - „*minden delfin emlős*”
 - “*Flipper egy delfin*”
 - *stb.*
- Az RDFS definiálja az *erőforrás* és az *osztály* fogalmát::
 - az *RDF* számára minden egy „*erőforrás*”
 - egy *osztály szintén egy erőforrás, de egyben...*
 - ...*más erőforrások („egyedek”)* lehetséges összessége
 - „*emlős*”, „*delfin*”, ...

Osztályok, erőforrások ... (folyt.)

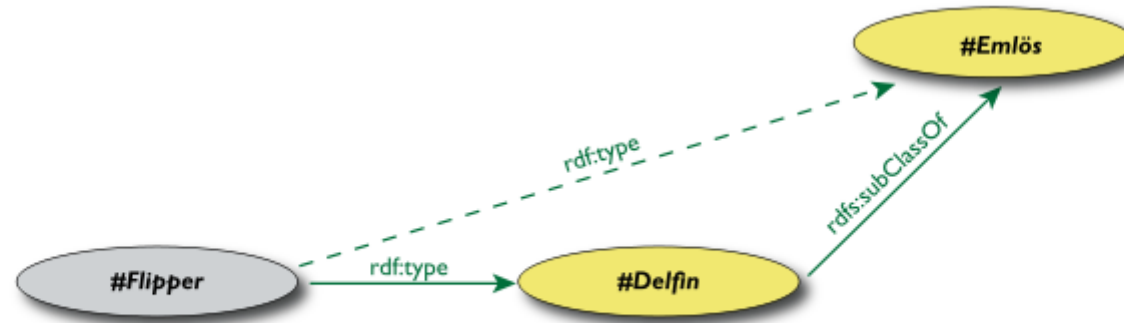
- Az erőforrások és az egyedek között relációk létesíthetők:
 - „típus” („typing”): vagyis egy egyed egy adott osztályhoz tartozik („Flipper egy delfin”)
 - „alosztály” („subclassing”): az egyik osztály egyedei automatikusan a másiknak is egyedei („minden delfin emlős”)
- Az *RDFS* ezeket a (tradicionális) fogalmakat formalizálja

Classes, Resources in RDF(S)



- Az RDFS definiálja a `rdfs:Resource`, `rdfs:Class`, `rdf:type`, `rdfs:subClassOf` fogalmakat
 - (ezek mind speciális, az ábrán névterekkel rövidített URI-k)

Következtetett tulajdonságok



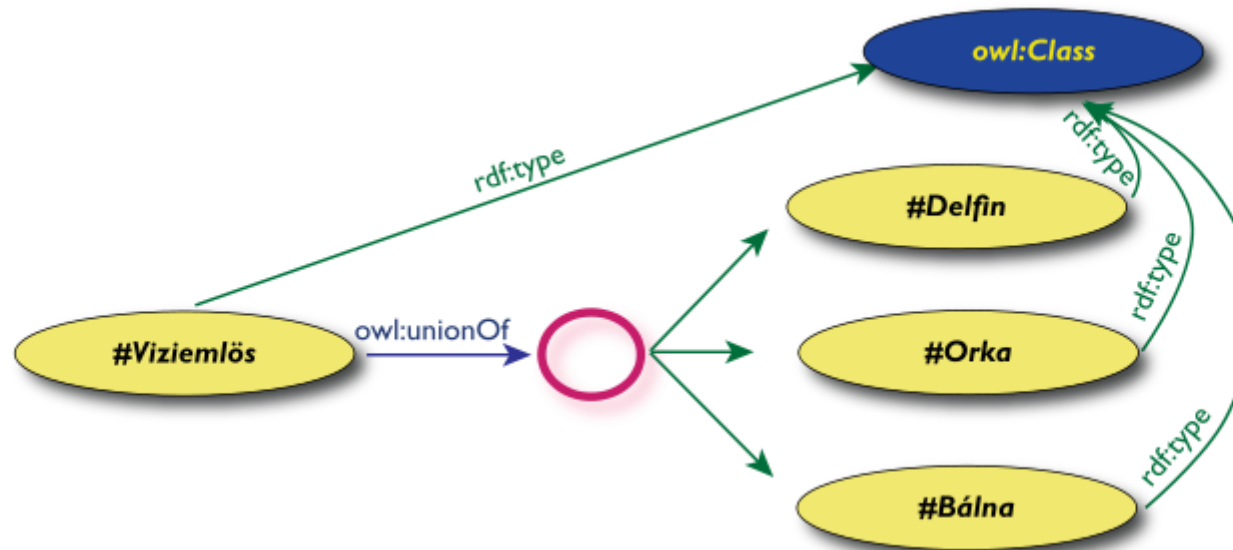
- **(#Flipper rdf:type #Emlős)** *nem* része az eredeti RDF adathalmaznak...
- ...de ki lehet *következtetni* az RDFS szabályokból
- Jobb RDF környezetek ezt az állítást is tartalmazzák

RDFS és OWL

- Az RDFS az alapelveket definiálja
- Az OWL hozzáad bonyolultabb lehetőségeket, mint például:
 - osztályok konstrukciója (a meglévő osztályokból kiindulva)
 - a tulajdonságok logikai jellemzése (pld. tranzitivitás, szimmetria, függvény)
 - stb.

Osztályok úniója

- Lényegében egy halmazelméleti únió (lehetne metszet, komplement, stb):



Az OWL további lehetőségei

- Az ontológiák nagyon nagyok lehetnek:
 - *nagy figyelmet kell fordítani a karbantartásukra*
 - *több részből (modulból) állhatnak*
 - *a részeknek különböző eredetük lehet melyeket integrálni kell*
- Ezek *Webontológiák*. Vagyis
 - *az alkalmazások több, egymástól különböző ontológiát használhatnak, vagy...*
 - *... ugyanazon ontológiát, de különböző nyelveken*
 - *vagyis a terminológiák ekvivalenciája fontos kérdéssé válhat*
- OWL lehetőséget ad az osztályok/tulajdonságok ekvivalenciájára, verziókontrollra, stb.

Példa: kapcsolat az angol és a magyar között



De: az ontológiák bonyolultak!

- Nehéz egy teljes ontológiarendszert implementálni
 - és egyes alkalmazások számára felesleges is lehet
- Innen az egyre bonyolultabb specifikációk „réteges” modellje, különböző megkötésekkel
- De: az RDFS, OWL-Lite és OWL-DL *kiszámítható*, míg ez nem igaz OWL Full-ra



A munka folytatódik... (a W3C-ben vagy azon kívül)

Lekérdezések

Ma már milliós(!) nagyságrenben használnak RDF hármassokat: „Query Language and Protocol for RDF (SPARQL)” egy alkalmas lekérdezőnyelv

(Logikai) szabályok

Vannak logikai kapcsolatok, amelyek nem írhatók le OWL-ben sem, további logikákra van szükség (pld. Horn–logika)

Bizalom

Például: „megbízhatok-e ezen és ezen állítások létrehozójában?”

Lekérdezések: SPARQL

- Az alapvető ötlet: *gráfminták* megadása:

```
SELECT ?név
WHERE {
    ?x abc:naptáram    ?y.
    ?x abc:személynév ?név.
}
```

- Vagyis, körülbelül: „add meg mindazoknak a nevét, akiknek a naptára a Weben van”
- A specifikáció még nem teljes, de már nagyon sok implementáció és alkalmazás létezik

SW alkalmazások

- Sok-sok alkalmazás van alakulóban:
- A legtöbb alkalmazás még mindig „centralizált”, a decentralizált alkalmazások száma még nem nagy
- Érdeemes például a [Semantic Technology Conference](#) sorozatot figyelemmel kíséni
 - *nem egy tudományos konferencia, inkább üzleti jellegű*
 - *az idén (múlt héten) óriási édeklődés volt a konferencia iránt, több mint 600 résztvevővel (pld.: IBM, Nokia, Cisco, BellSouth, GE, Walt Disney, Oracle, Microsoft, ...)*

Példa: portálok

- A Vodafone „Live Mobile Portal”
 - *RDF*–alapú keresőrendszer (pl. telefonhangok, játékok, képek)
 - letöltéshez szükséges lapkeresések száma 50%-kal csökkent
 - telefonhangok letöltése 2 hónap alatt 20%-kal nőtt
- SwordFish a Sun-nál: szintén egy *RDF*–alapú keresés a [White Paper Collections](#) és [System Handbook collections](#) lapokon
- A Nokia nemrégiben nyitott egy hasonló [fejlesztői portált](#)



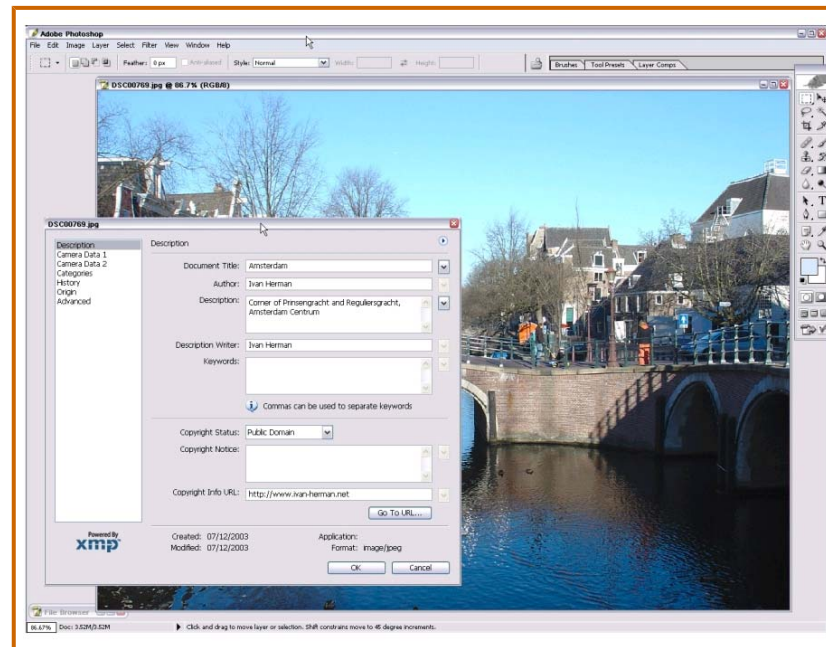
Példa: ontológia–alapú keresés: GoPubMed

- pubmed.org-ra alapozott keresés
- A keresés eredményeit újrendezi speciális ontológiák segítségével
- Extra keresési kulcsokat generál az ismert terminológia alapján
- Szép példa az *alkalmazásfüggő ontológiák* jelentőségére

The screenshot displays the GoPubMed interface. At the top, the search bar contains the query 'tinnitus' and a 'Go' button. Below the search bar, the results are organized into two columns. The left column, titled 'Induced Gene Ontology', shows a hierarchical tree of GO terms. The right column, titled 'Results for "tinnitus" and GO term "cellular process"', displays a list of search results. Each result includes a title, a snippet of text, and a list of associated GO terms with their respective percentages. The first result is titled 'Ear problems in swimmers...' and lists terms like 'perception of sound (100%)', 'pH reduction (100%)', 'middle ear morphogenesis (83%)', and 'inner ear morphogenesis (79%)'. The second result is titled 'Self-reported nonmusculoskeletal responses to chiropractic interventions...' and lists terms like 'reproduction (100%)', 'respiratory gaseous exchange (100%)', 'digestion (100%)', 'perception of sound (100%)', 'circulation (100%)', and 'strategic response (73%)'.

Adobe XMP

- Az Adobe eszközök RDF–alapú metaadatot adnak a képekhez, rajzokhoz, stb.
- Az **eszköz** mindenki számára rendelkezésre áll!



Baby CareLink

- Koraszülött kisbabák kezelésére szolgáló információközpont
- Egy OWL–alapú webszolgáltatás
 - egymástól nagyon eltérő adatokat kombinál (orvosi, biztosítási, jogi, stb.)
 - a felhasználó komplex kérdéseket tehet fel, és — adott esetben — bővítheti a tudásbázist

The screenshot displays the CST Baby CareLink website. The main heading is "Product Map". Below it, a navigation menu lists "Components": Neonatal Intensive Care, Neonatal Care Management Program, After the NICU, Healthy Beginnings / First Year of Life, and High Risk Pregnancy. The main content area features a "Product Map" diagram with categories: Prenatal Care, Newborn Intensive Care, Infant Care, Clinician Tools, Healthy Beginnings, High-Risk Pregnancy, Neonatal Intensive Care, After the NICU, First Year of Life, and Care Manager Tools. A "Did You Know?" box states: "7.6% (300,000) of all births in the U.S. each year are low birthweight (< 2500 gms, 5 pounds, 8 ounces)." The footer includes "© 2004 Clinician Support Technology - One" and "459-3226 USA".

Sok-sok eszköz áll rendelkezésre

- Ontológia–szerkesztők:
 - *Protege 2000 (Stanford Univ.)*, *SWOOP (Univ. of Maryland)*, *Orient (IBM)*
- Programozási rendszerek:
 - *Jena (Java)*, *RDFLib (Python)*, *Redland (C, Tcl, Java, PHP, Perl, Python)*, *SWI-Prolog*, ...
- Adatbázisok (sql-re vagy kizárólag hármásokra alapozódva):
 - *Kowari*, *Gateway*, *3Store*, *Jena's Joseki*, *Oracle Database 10g*, ...
- RDF és OWL ellenőrzők:
 - *W3C's RDF Validator*, *BBN OWL Validator*, *Pellet OWL Reasoner* ...
- Érdemes [a W3C RDF–fejlesztői lapját](#) vagy [Dave Beckett's lapjait](#) figyelemmel kísélni

Információk magyarul

- A teljes RDF– és OWL–szabvány rendelkezésre áll magyarul is
 - lásd a [W3C Magyar Iroda fordításjegyzékét](#)
 - a fordítás Pataki Ernő munkája
- Könyvek magyarul:
 - Gottdank Tibor, *Szemantikus Web, ComputerBooks, Budapest, 2005*
 - Szeredi Péter, Lukács Péter, Benkő Tamás, *A szemantikus világháló elmélete és gyakorlata, TypoTex, Budapest, 2005*

További információk

Ez az előadás elérhető a Weben (XHTML vagy PDF):

<http://www.w3.org/2006/Talks/0318-Budapest-IH/>

<http://www.w3.org/2006/Talks/0318-Budapest-IH/Overview.pdf>

A cikk PDF változata szintén a Weben van:

<http://www.w3.org/2006/Talks/0318-Budapest-IH/cikk.pdf>

Semantic Web honlap

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

Elérés, információ a W3C-ről a W3C Magyar Irodáján keresztül:

<http://www.w3c.hu/>

Email címem:

ivan@w3.org

