

Tudásalapú információ integráció

(A Szemantikus Web megközelítés
és a másikk irány)

Tanszéki értekezéslet, 2008. május
14.

Miért van szükségünk ilyesmire? (Alkalmazások)

- **WWW:**
 - Keresés a weben (pl. összehasonlítás alapú vásárlás)
 - Portál építések több adatforrás felhasználásával
 - B2B, elektronikus piacterek
- **Tudomány és kultúra:**
 - Asztrofizika: égi jelenségek gyűjtése.
 - Kultúra: kulturális információs adatbázisok egységes elérése országhatárokon túl
 - Genetika: gén információk integrálása
- **Vállalati adatintegráció**
 - Egy átlagos KKV 49 adatbázist alkalmaz és IT költségvetésének 30%-át az adatintegrációra költi (US, 2002)

Információ integráció

Az információ integráció célja strukturált és félig-strukturált információforrások együttes kezelése, információigények egységes felületen történő kiszolgálása

Területek:

- Intézmények (közép- és nagyvállalatok, államigazgatás)
- Web

Megközelítési módok:

- Források kiegészítése (szemantikus web koncepció)
- Lekérdező eszközök fejlesztése (mediátor és információintegrációs rendszerek)

Miért nem elég:



- Keresőgépek szövegalapú keresést végeznek
 - Jól működik egyedi dokumentumokon
 - Nem tudnak integrálni több dokumentumból származó információkat
 - Nem képesek hatékony általánosításra
 - Nem tudnak dokumentumokat és adatbázisokat összekapcsolni

Összehasonlítás alapú vásárlás?

- Hasonló források
- De:
 - **Eltérő fókusz**
 - **Eltérő a spektrumuk az adatbázisoknak**
 - **Szolgáltatások**
 - **Új kihívás**
 - **Az “adattár” megközelítés nem működik**
 - **Kézi forrásleírás, kezelés korlátai**



Netbot



DealPilot.com, the Ultimate Comparison Shopping Engine: Price Comparison Results - Microsoft Internet Explorer

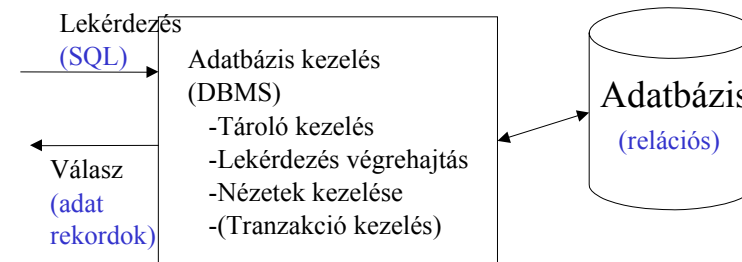
Click on one offer (1st column) to proceed to the respective shop!

Displaying Top Ten offers:

Total Price	Shop	Item Price	Discount	US Sales Tax	Shipping Costs	Shipping Time	Shipping Service	Delivery Time
US\$ 6.66	1000's of Discount Books, USA, WA	US\$ 3.91	83%	US\$ 0.00	US\$ 2.75	2-7 days	Book Rate	4-9 days
US\$ 16.43	Powells, USA, OR	US\$ 12.95	44%	US\$ 0.00	3.50	14-56 days	USPS 4th Class Book Rate	17-61 days
US\$ 18.95	Al Books, USA, NJ	US\$ 15.00	35%	US\$ 0.00	3.95	4-6 days	UPS Ground	9-11 days
US\$ 19.95	powells, USA, OR	US\$ 12.95	44%	US\$ 0.00	6.00	n/a	UPS Ground	n/a
US\$ 20.48	kingbooks.com, USA, WA	US\$ 17.02	26%	US\$ 0.00	2.44	14 days	USPS Book Rate	n/a
US\$ 20.67	Shopping.com, USA, CA	US\$ 15.64	32%	US\$ 1.12	3.90	3-7 days	UPS Regular Mail	n/a
US\$ 20.96	kingbooks.com, USA, WA	US\$ 17.02	26%	US\$ 0.00	3.94	3-7 days	UPS Ground	n/a
US\$ 21.46	Kingbooks.com, USA, WA	US\$ 17.02	26%	US\$ 0.00	4.44	2-4 days	USPS Priority Mail	n/a
US\$ 22.20	1bookstreet.com, USA, CA	US\$ 20.70	10%	US\$ 1.50	0.00	3-14 days	Parcel Service	6-21 days

Miért nem csak adatbázisok elosztott adatbázisok

- **Közös séma hiánya**
 - Források heterogén sémákkal (és fogalmakkal, ontológiákkal) rendelkeznek
 - Félig-strukturált források
- **Régi források**
 - Nem relációs sémák
 - Eltérő elérési módok
- **Független források**
 - Nincs közös adminisztráció
 - Nem kezelt forrás tartalmi átfedések
- **Nehezen előre jelezhető viselkedés**
 - Lekérdezés végrehajtás bonyolult
- **Általában csak olvashatóak**
 - Ez lehet szerencsés is
 - Terjednek a tranzakció kezelési megoldások a



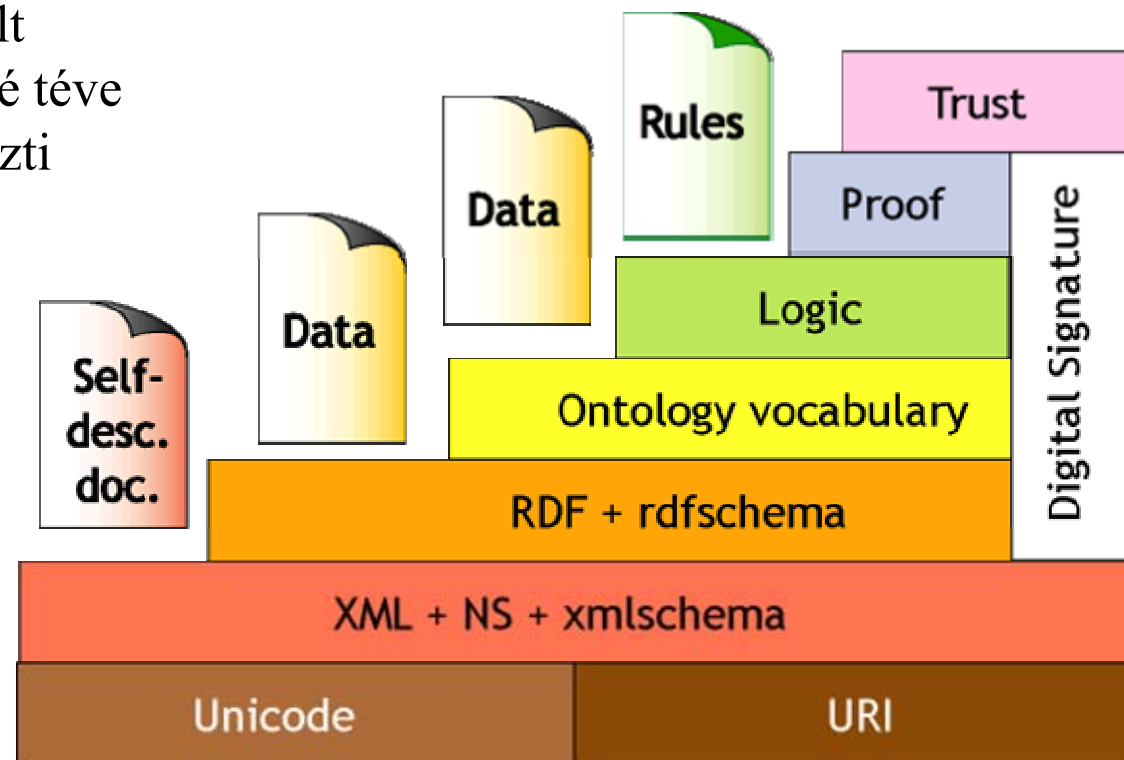
Az egyik irány

A szemantikus web koncepció

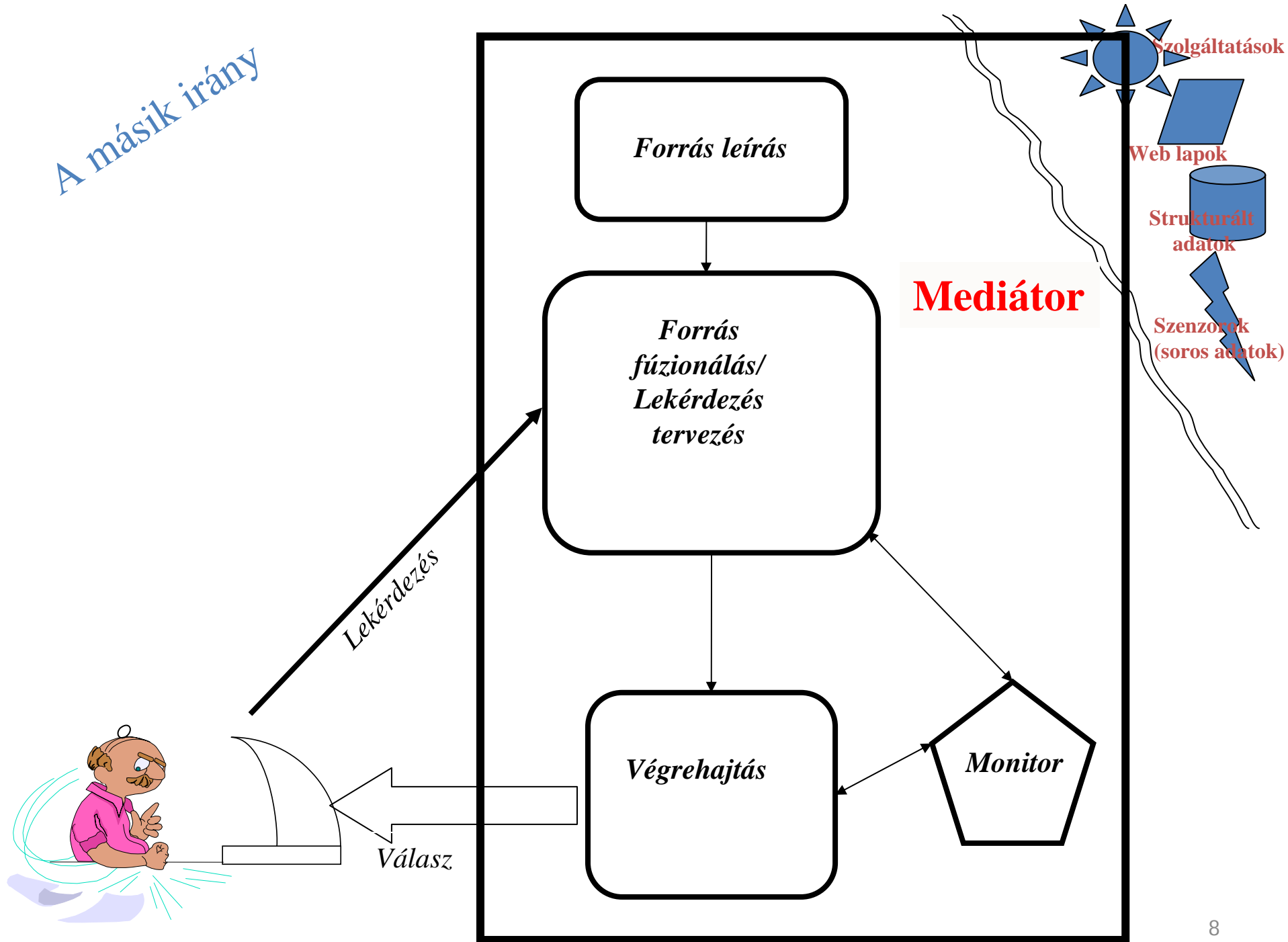
„A szemantikus web egy kiterjesztése a jelenlegi webnek, amelyben az információknak jól definiált jelentést adhatunk, lehetővé téve a gépek és felhasználók közti jobb együttműködést.,,

„A szemantikus web elérhetővé teszi a tudást, mint a web a hypertextet.”

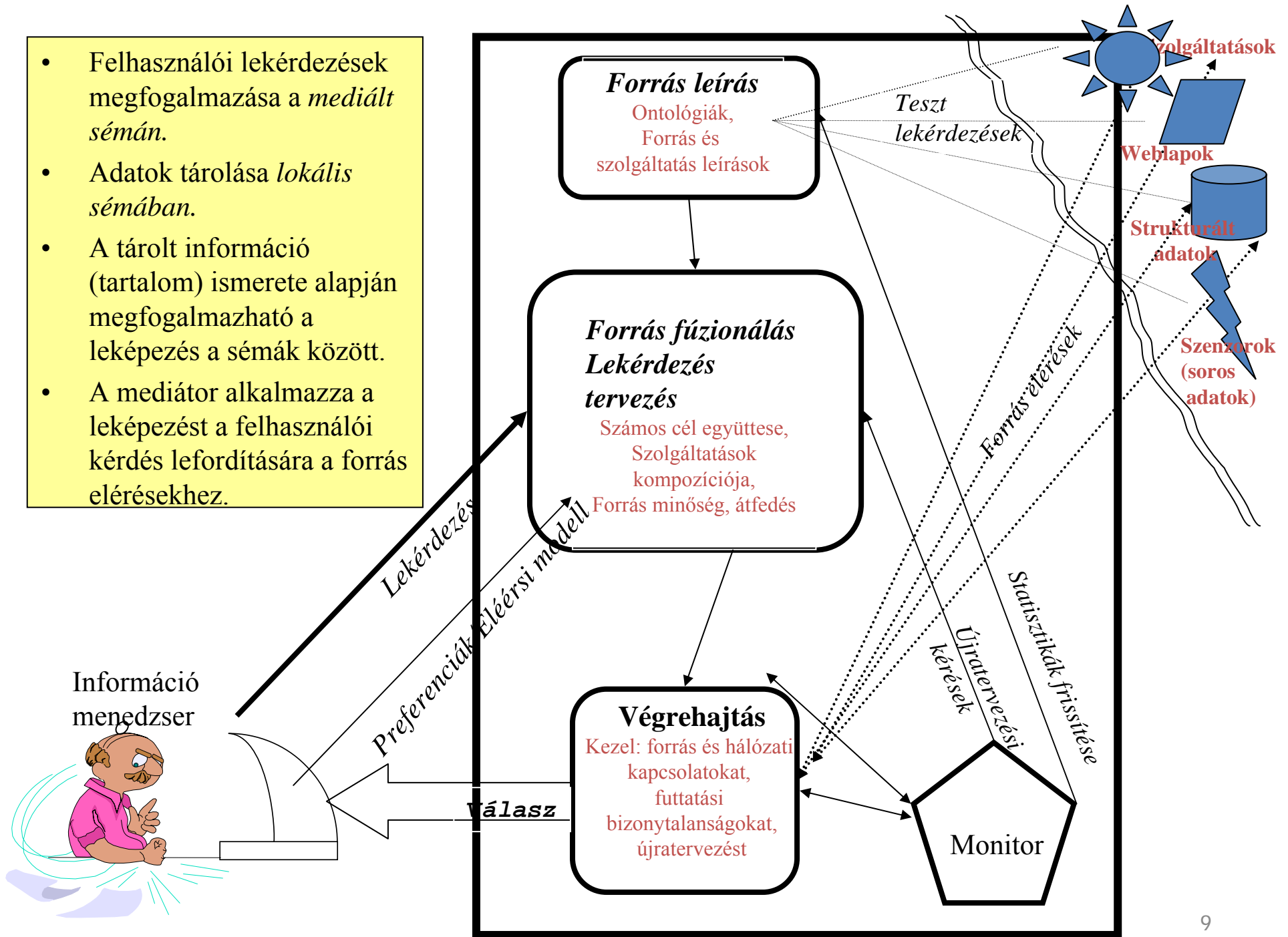
Tim Berners-Lee



A másik irány

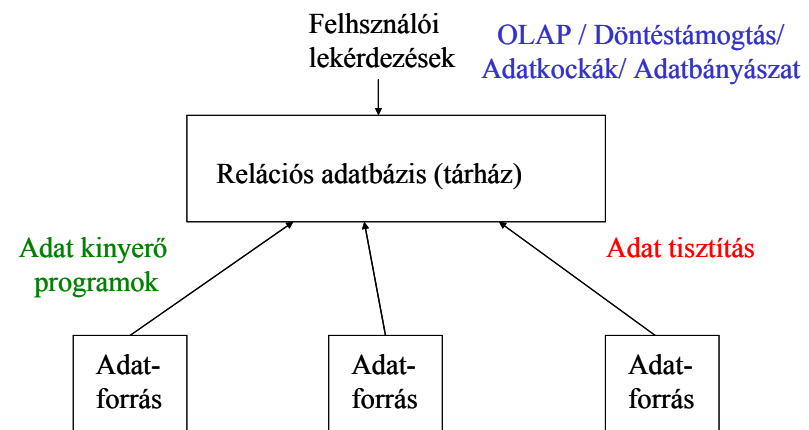
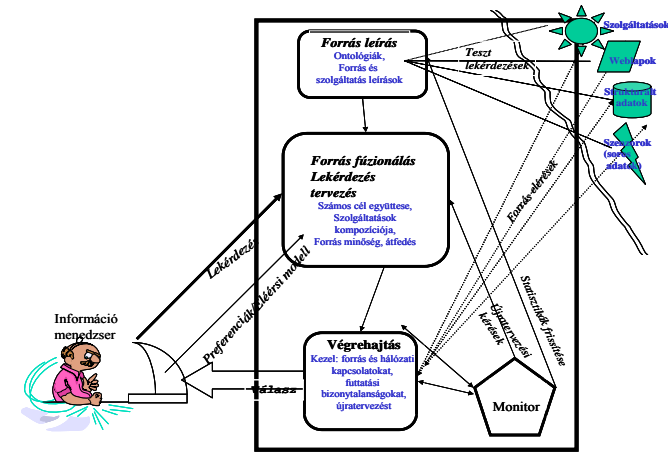


- Felhasználói lekérdezések megfogalmazása a *mediált sémán*.
- Adatok tárolása *lokális sémában*.
- A tárolt információ (tartalom) ismerete alapján megfogalmazható a leképezés a sémák között.
- A mediátor alkalmazza a leképezést a felhasználói kérdés lefordítására a forrás elérésekhez.



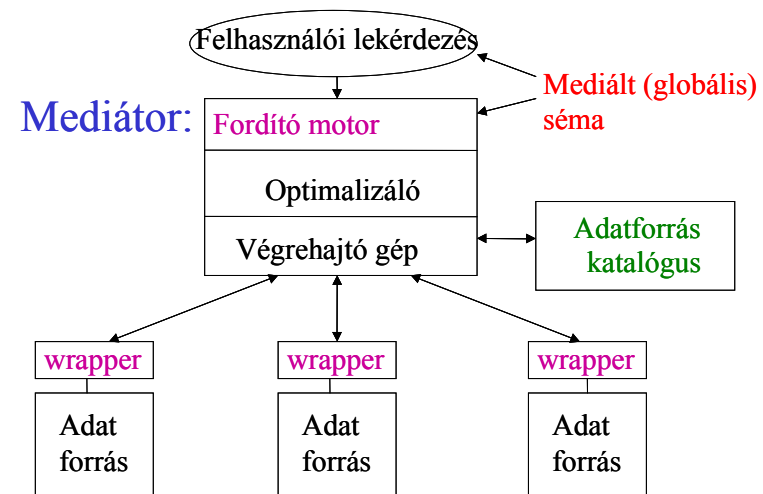
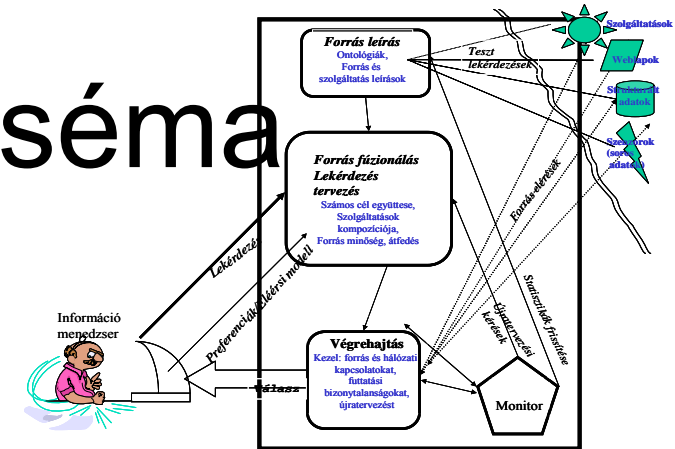
Kis forrás szám melletti integráció

- **Általában ad-hoc programozás:** speciális eset megvalósítása minden esetre, sok konzultáció.
- **Adattárházak:** minden adat periódikus feltöltése az adattárházba.
 - 6-18 hónap bevezetési idő
 - Operációs és döntéstámogatási RDBMS elválasztás. (nem csak adatintegrációra megoldás).
 - Teljesítmény jó,
 - adat lehet, hogy nem friss;
 - Rendszeres adattisztítás szükséges.

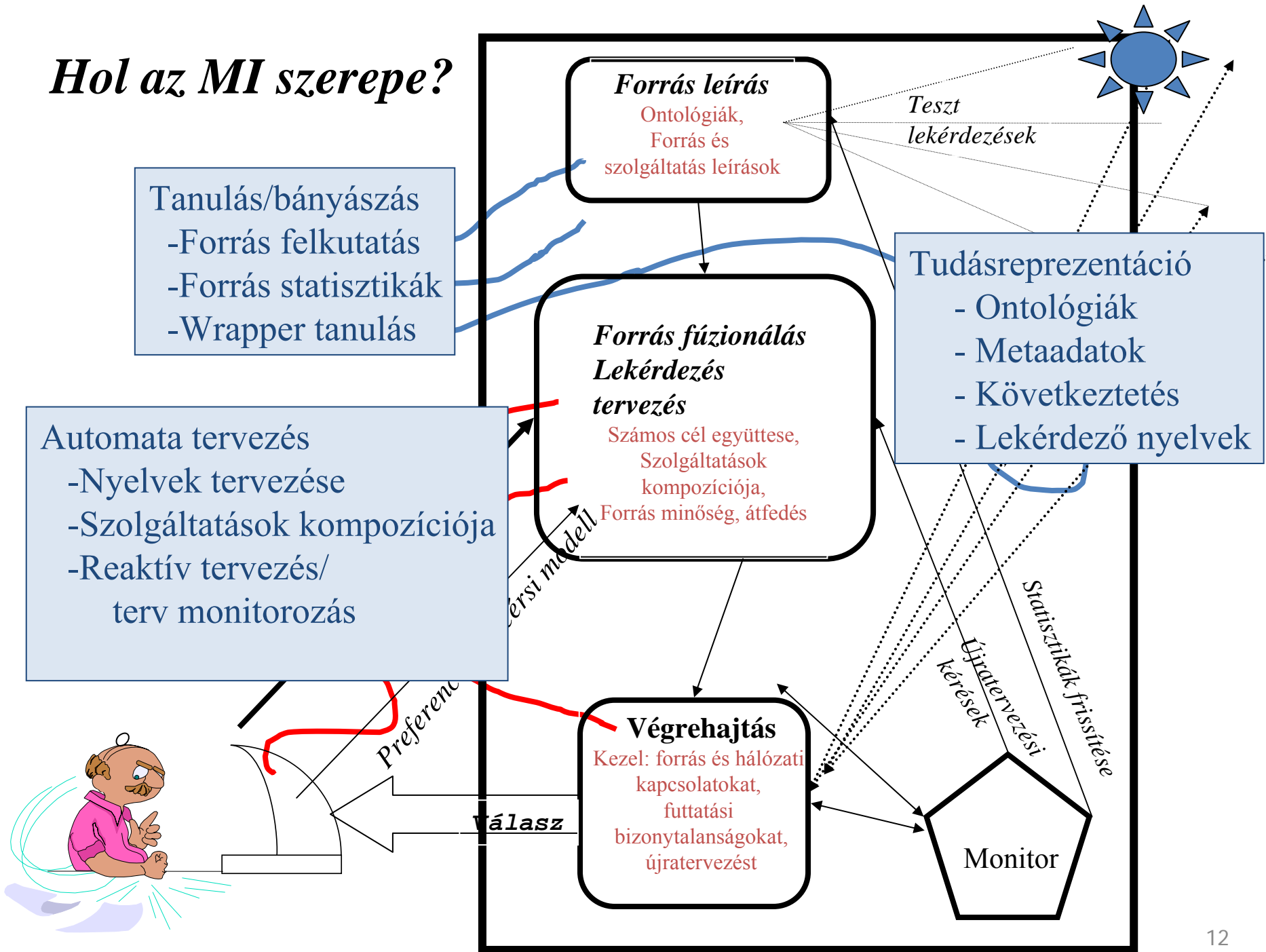


Virtuális integrációs séma

- Adatok a forrásokban maradnak
- Lekérdezés végrehajtásakor:
 - Releváns források meghatározása
 - Lekérdezés szétválasztása forrásokra vonatkozó lekérdezésekre.
 - Válaszok begyűjtése a forrásokból, és megfelelő kombinálása a válasz előállításához.
- Friss adatok
- A megoldás skálázható



Hol az MI szerepe?



Köszönöm a figyelmet!