

Integrációs és ellenőrzési technikák

VIMIAC04, 2019. tavasz

Micskei Zoltán
Strausz György

Méréstechnika és Információs Rendszerek Tanszék

<https://www.mit.bme.hu/oktatas/targyak/vimiaco4>

Integrációs és ellenőrzési technikák

Hogyan építsünk információ gazdag megoldásokat?

- Információ/adat integráció
 - Webes környezetek
 - Intézményi környezetek
- Technikák: szemantikus technológiák

Hogyan fejlesszünk jó minőségű szoftvert?

- Kód gyakori integrálása és ellenőrzése
- Kód review, kód analízis, tesztelés...
- Technikák: continuous integration (CI), mocking, automatikus tesztelés...

Integrációs és ellenőrzési technikák

Kurzus menetrendje/követelmények

Előadások: hétfő 10.15 – 12.00., IB 025

Gyakorlatok: péntek (8.15-10.00., 10.15-12.00, 12.15 – 14.00)

6db gyakorlat, kéthetente, IB.413. terem

Gyakorlatok témái:

- *1-3. alkalmak: Szemantikus technológiák (adatmodellezés, integráció)*
- *4-6. alkalmak: Tesztek írása, CI beüzemelése, kód átvizsgálása...*

Integrációs és ellenőrzési technikák

Kurzus menetrendje/követelmények

Zárthelyi:

Zárthelyi időpontja: 2019. április 2., 8.15-10.00

*Pótzárthelyi időpontja: 2018. április 16., 8.15-10.00
(helyszínek később)*

*ZH értékelés: elfogadás 40%,
e feletti pontok 25%-a beszámít a vizsgába,
maximum 15% vizsgapont szerezhető)*

HF: *követelmény: elfogadás; 4 fős csapatok,
GitHub infrastruktúra használata*

Vizsga: *írásbeli*

Integrációs technikák - tartalom

- Bevezetés, probléma felvetés
- Szemantikus technológiák
 - Szemantikus web koncepció
 - RDF, SPARQL
 - RDFS, OWL
 - Linked Data
- Adatintegráció
 - Elosztott, webes környezetek
 - Mediátorok (lekérdezés átalakítások nézetek felhasználásával)
 - Intézményi környezetek, adattárházak

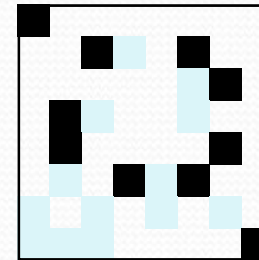
Integráció: Tartalom vs Forma

Szemantikus nézet



*A megbeszélés sikere az
érvelésen és a logikus
következtetésen múlik.*

Szintaktikus nézet



A _megbeszélés_sikere
_az_érvelésen_és_a_lo
gikus_következtetésen
_múlik.

Adatbázisok

name	manager	office	phone
John	Jill	MJH222	38086
Jane	Jerry	Cedar12	57493
Jill		MJH222	
Jerry		420-032	56777

Töredezettség

Horizontalis töredezettség

name	manager	office	phone
John	Jill	MJH222	38086
Jane	Jerry	Cedar12	57493

name	manager	office	phone
Jill		MJH222	
Jerry		420-032	56777

Vertikális töredezettség

name	manager
John	Jill
Jane	Jerry
Jill	
Jerry	

name	office	phone
John	MJH222	38086
Jane	Cedar12	57493
Jill	MJH222	
Jerry	420-032	56777

Heterogenitás

name	manager	office	phone
John	Jill	MJH222	38086
Jane	Jerry	Cedar12	57493
Jill		MJH222	
Jerry		420-032	56777

name	employee	location	telephone
John		MJH222	7238086
Jane		Cedar12	7257493
Jill	John	MJH222	
Jerry	Jane	420-032	7256777

Technológia heterogenitást csak-csak megoldjuk...

Miért van szükségünk információ integrációra? (Alkalmazások)

- **WWW:**
 - Összehasonlítás alapú vásárlás
 - Portál építések több adatforrás felhasználásával
 - B2B, elektronikus piacterek
- **Tudomány és kultúra:**
 - Genetika: gén információk integrálása
 - Asztrofizika: égi jelenségek gyűjtése.
 - Kultúra: kulturális információs adatbázisok egységes elérése országhatárokon túl
- **Vállalati adatintegráció**
 - Egy átlagos KKV 49 adatbázist alkalmaz és IT költségvetésének 30%-át az adatintegrációra költi (US, 2009)

Csak szöveg volna a weben?

- A web jelentős része valójában strukturált...
 - A legtöbb web szerver mögött adatbázisok állnak
 - Dinamikusan konvertálják az adatokat olvasható nyelvi formára
 - **<India, New Delhi> => The capital of India is New Delhi.**
 - Ha vissza tudnánk konvertálni lenne strukturált adatunk!
 - (ki)csomagolók, csomagolók tanulása, stb...
 - Dinamikus lapokat is fel tudunk deríteni...
 - Félig-strukturált web (kialakulóban)
 - Legtöbb lap részben strukturált (pl. XML)
 - XML a szabvány a szintaktikára, ismert problémák az értelmezéssel
 - Szolgáltatások
 - Utazási szolgáltatások, vásárlások támogatása
 - Érzékelők

Tőzsdei árfolyamok, hőmérsékletek, jegyárak...

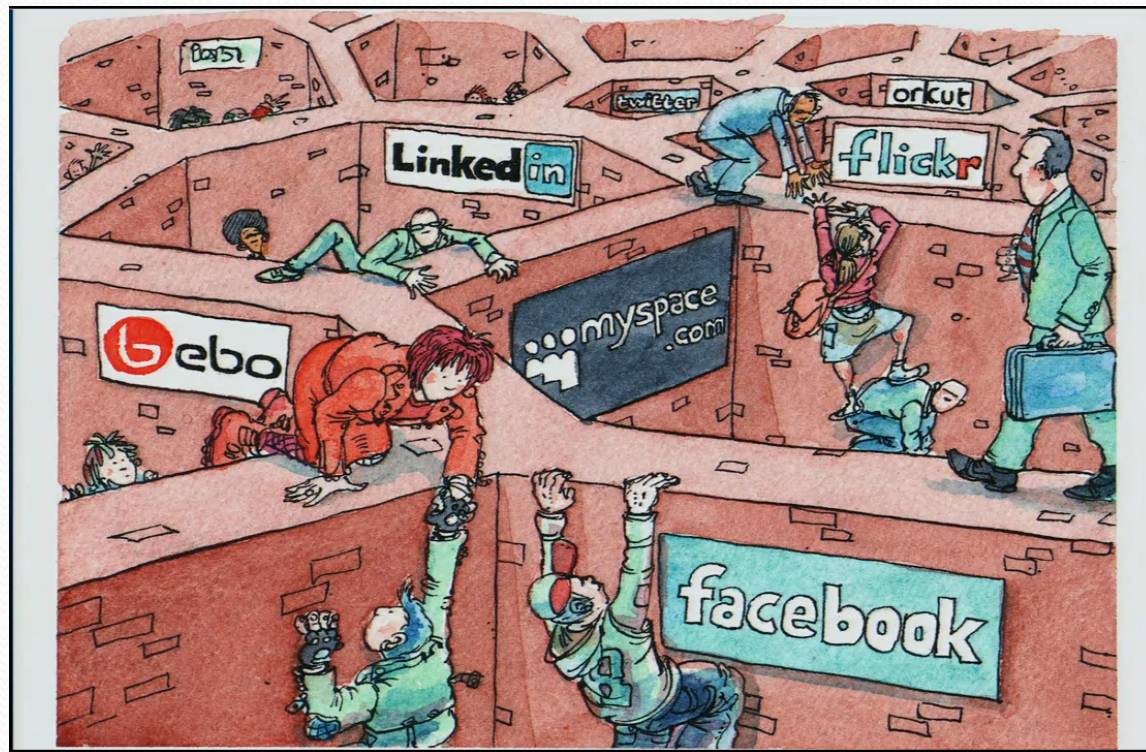


Miért nem elég:

- 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
- Az információ integráció célja strukturált és félig-strukturált információforrások együttes kezelése

Ennél azért már több is elérhető:

- Szövegalapú keresés -> tudásháló (knowledge graph)
<https://www.google.com>
- Wikipedia -> DBPedia
- Profil hálók



Példa: tudás hálók

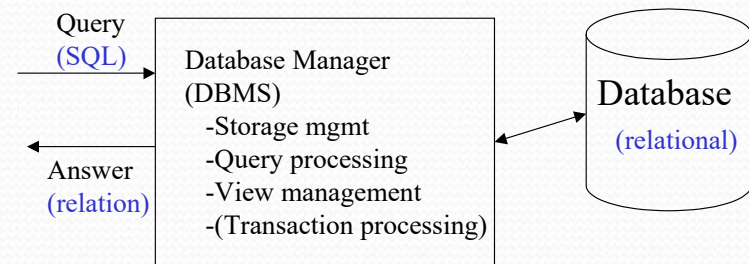
Name	Instances	Facts	Types	Relations
DBpedia (English)	4,806,150	176,043,129	735	2,813
YAGO	4,595,906	25,946,870	488,469	77
Freebase	49,947,845	3,041,722,635	26,507	37,781
Wikidata	15,602,060	65,993,797	23,157	1,673
NELL	2,006,896	432,845	285	425
OpenCyc	118,499	2,413,894	45,153	18,526
Google's Knowledge Graph	570,000,000	18,000,000,000	1,500	35,000
Google's Knowledge Vault	45,000,000	271,000,000	1,100	4,469
Yahoo! Knowledge Graph	3,443,743	1,391,054,990	250	800

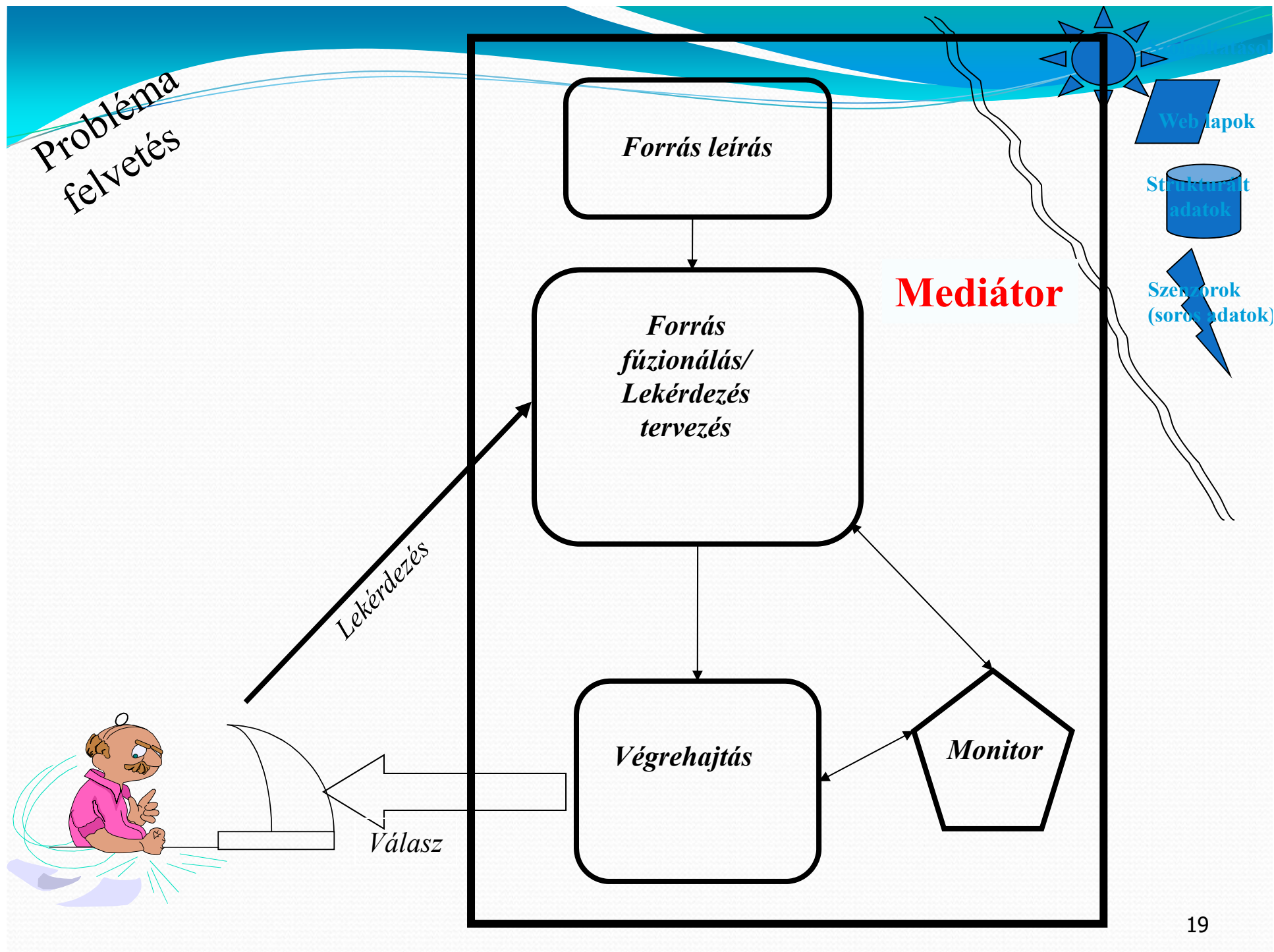
Forrás: Knowledge Graph Refinement: A Survey of Approaches and Evaluation Methods, Editor(s): Philipp Cimiano, Universität Bielefeld, Germany
 Solicited review(s): Natasha Noy, Google Inc., USA; Philipp Cimiano, Universität Bielefeld, Germany; Semantic Web 0 (2016) 1–0 IOS Press

Miért nem csak

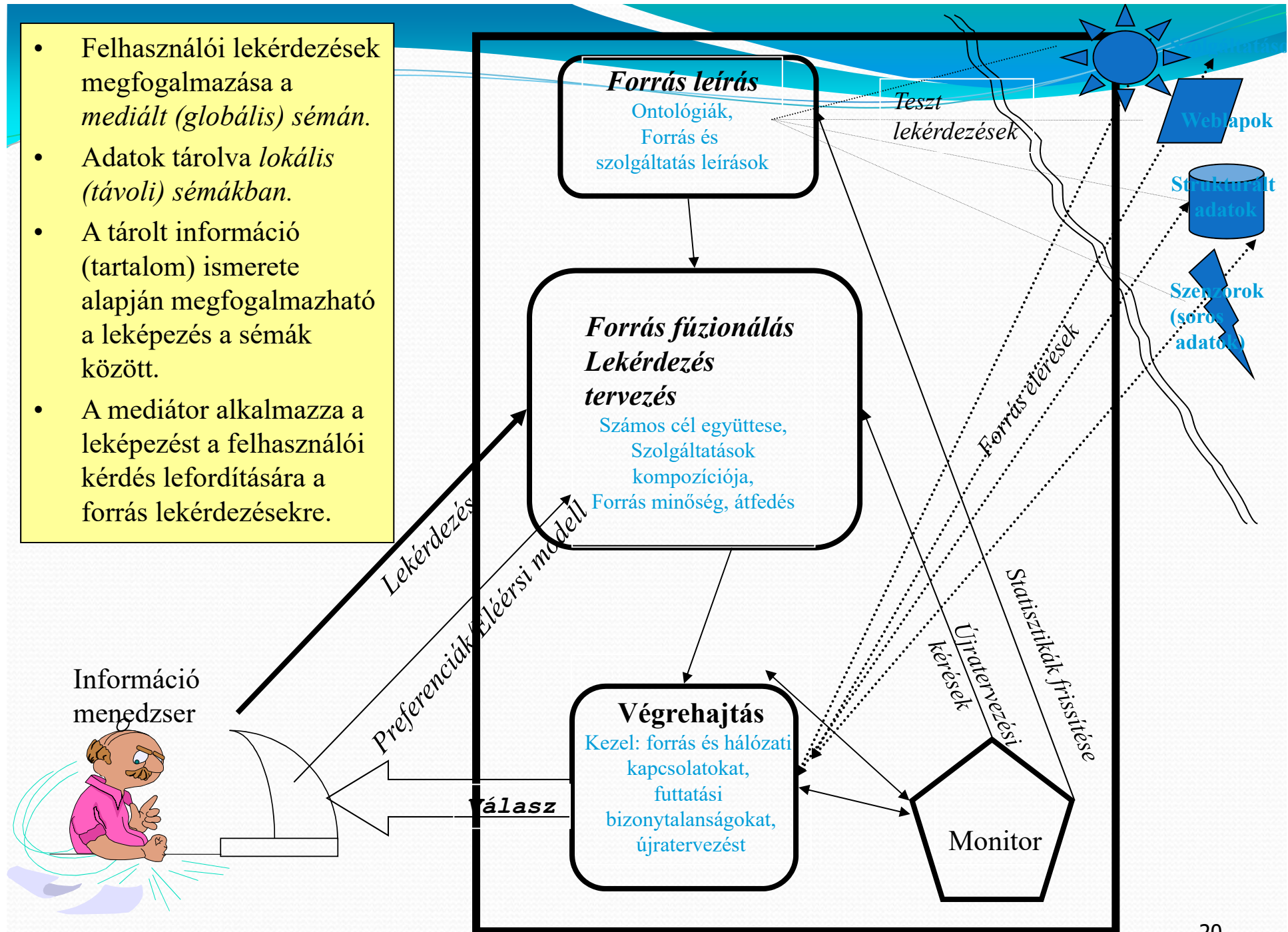
adatbázisok
elosztott adatbázisok

- 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 előrejelezhető viselkedés
 - Lekérdezés végrehajtás bonyolult
- Általában csak olvashatóak
 - Ez lehet szerencsés is
 - Bár terjednek a tranzakció kezelési megoldások a weben





- Felhasználói lekérdezések megfogalmazása a *mediált (globális) sémán*.
- Adatok tárolva *lokális (távoli) sémákban*.
- A tárolt információ (tartalom) ismerete alapján megfogalmazható a leképezés a sémák között.
- A mediátor alkalmazza a leképezést a felhasználói kérdés lefordítására a forrás lekérdezésekre.



Mesterséges intelligencia

Tanulás/bányászás

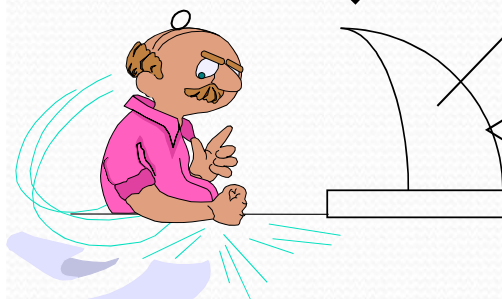
- Forrás felkutatás
- Forrás statisztikák
- Wrapper tanulás

Tudásreprezentáció

- Ontológiák
- Metaadatok
- Következtetés
- Lekérdező nyelvek

Automata tervezés

- Nyelvek tervezése
- Szolgáltatások kompozíciója
- Reaktív tervezés/
terv monitorozás



Preferenciák

Lekérdező

Elérési modell

Válasz

Forrás leírás

Ontológiák,
Forrás és
szolgáltatás leírások

Teszt
lekérdezések

Forrás fűzionálás Lekérdezés tervezés

Számos cél együttese,
Szolgáltatások
kompozíciója,
Forrás minőség, átfedés

Forrás elérések

Végrehajtás

Kezel: forrás és hálózati
kapcsolatokat,
futtatási
bizonytalanságokat,
újratervezést

Újratervezési
kérdések

Statisztikák frissítése

Monitor

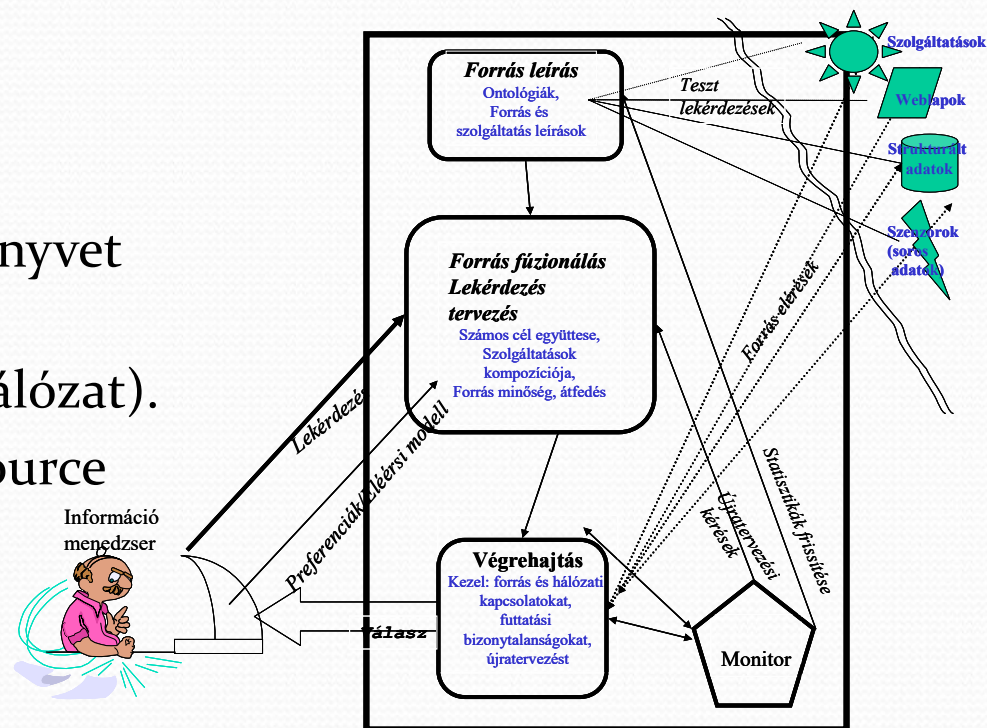
Webapok

Strukturált
adatok

Szenzorok
(sörös
adatok)

Forrás leírások

- Minden meta-adat információt tartalmaz
 - Forrás tartalom logikai leírása (könyvek, új autók).
 - Forrás képességek (pl. SQL lekérdezés feltehető)
 - Forrás teljesség (minden könyvet tartalmaz).
 - Fizikai jellemzők (forrás, hálózat).
 - Statisztikák az adatokról Source reliability
 - Tükör források
 - Frissítési frekvencia.

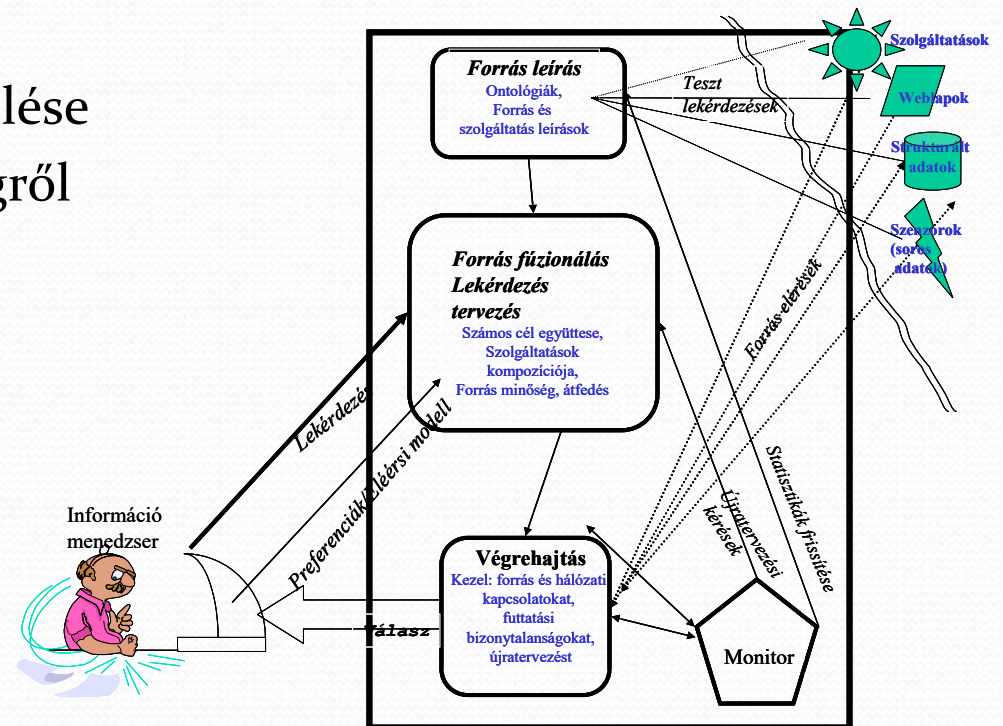


Forrás elérések

- Hogyan kapunk n-eseket
 - Számos forrás strukturálatlan adatokat ad
 - Néhány inherensen strukturálatlan, mások természetes nyelvi köntösben vannak
 - Vissza kell csomagolni az adatokat
 - Wrapper építés/információ kinyerés
 - Kézi munka/fél-automatikus

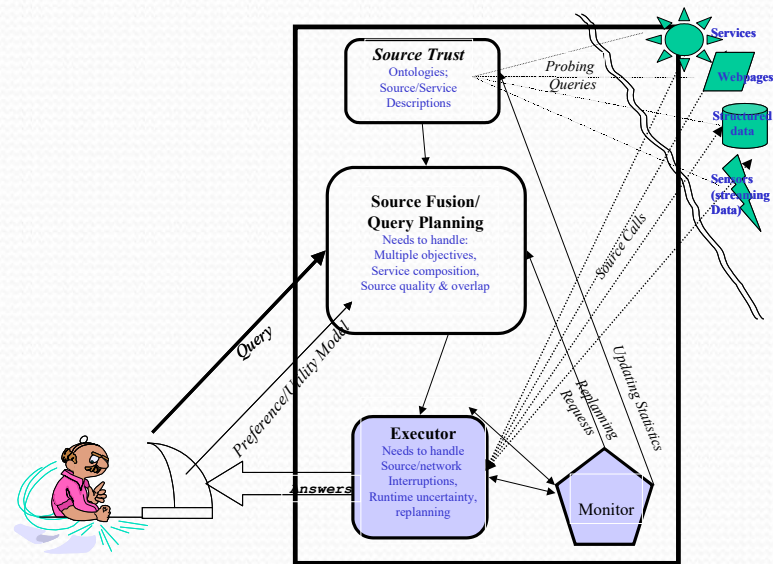
Forrás fúzió/ lekérdezés tervezés

- Feldolgozza a felhasználói lekérdezést és előállítja a végrehajtási tervet
 - Költség és hatékonyság közti optimalizáció
 - Forrás elérési korlátok kezelése
 - Információ a forrásminőségről



Monitoring/ Végrehajtás

- Lekérdezési terv alapján elvégzi a feladatot a forrásokon
 - Forrás késleltetések kezelése
 - Hálózati, tranziens kimaradások
 - Forrás elérési korlátok
 - Szükséges lehet újratervezések elvégzése

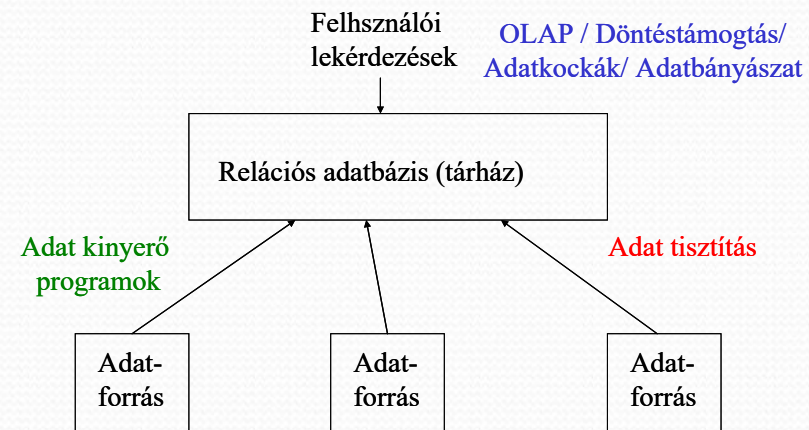
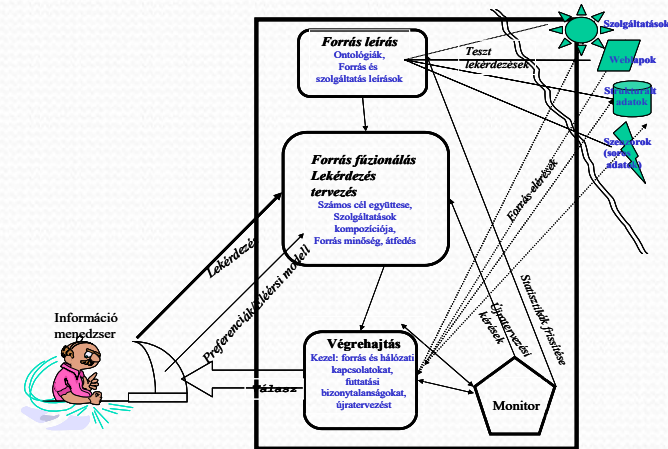


Méretek figyelembe vétele

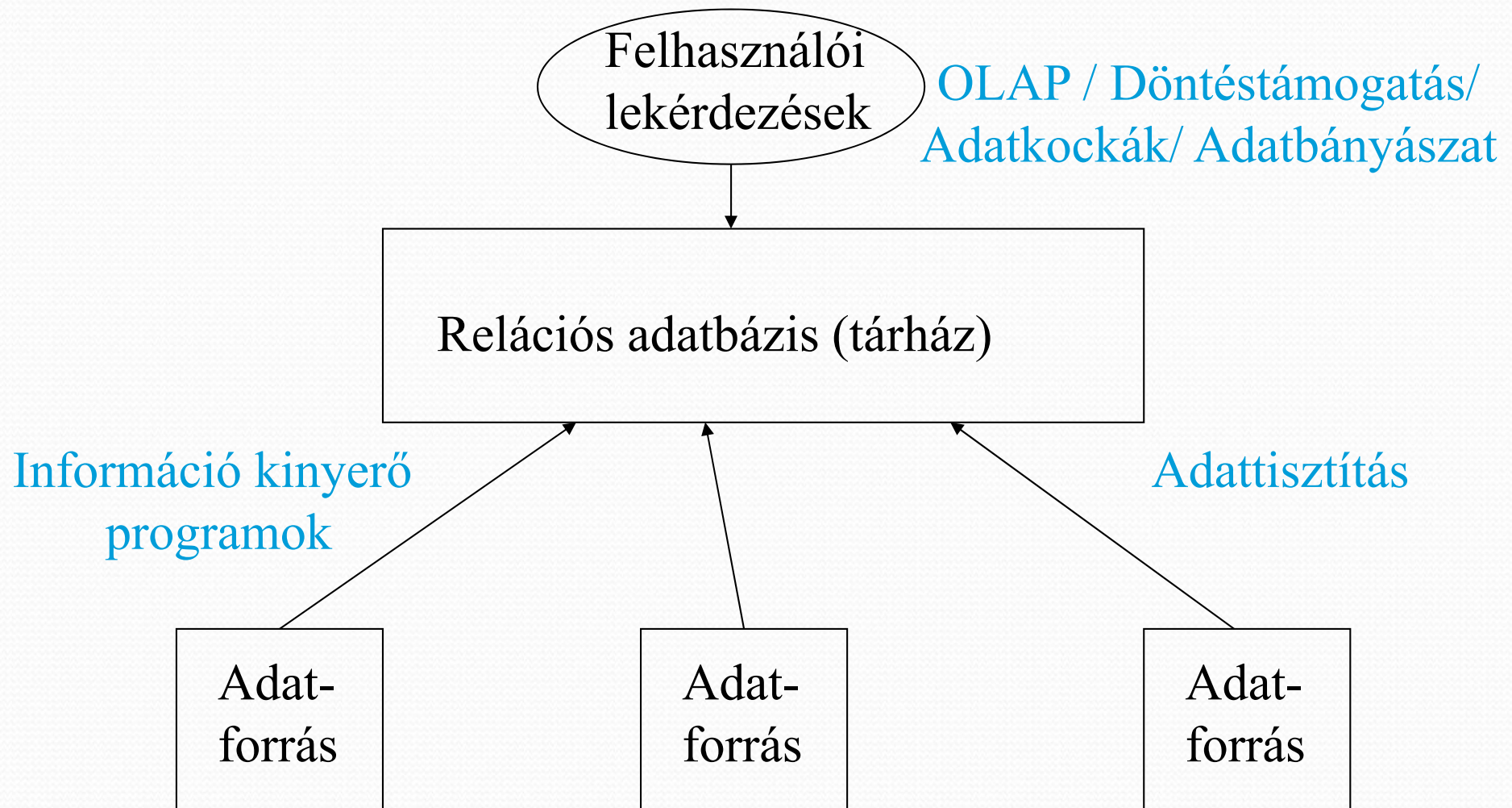
- Hány forrást kell elérni?
- Mennyire autonómok ezek?
- Van ismeretünk a forrásokról?
- Strukturáltak az adatok?
- Csak lekérdezés lehetséges vagy módosítás is?
- Követelmények: pontosság, teljesség, teljesítmény, inkonzisztenciák kezelése
- Zárt vagy nyílt világ feltételezés?

Kis forrás szám melletti integráció

- **Általában ad-hoc programozás:** speciális eset megvalósítása minden esetre, sok konzultáció.
- **Adattárházak:** minden adat periodikus 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.

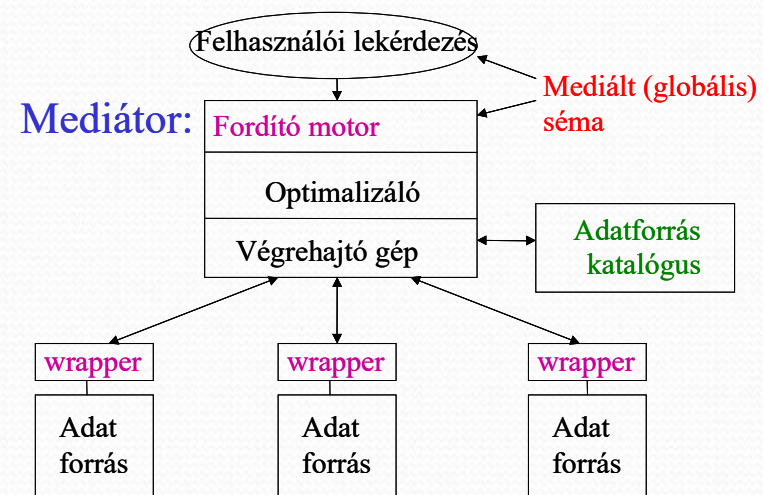
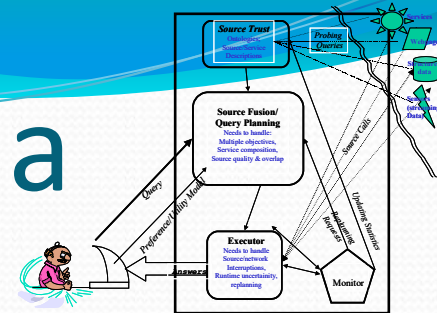


Integrátor séma



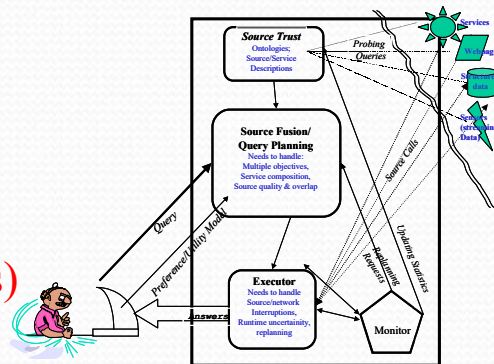
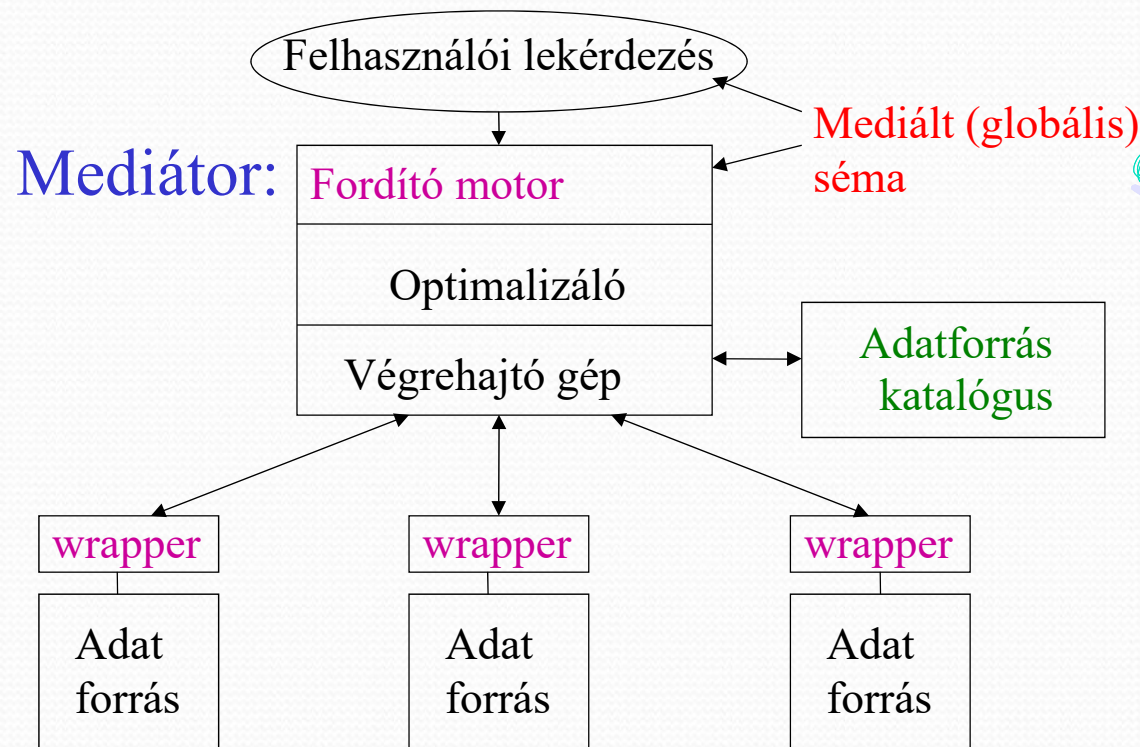
Virtuális integrációs séma

- 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



Garlic [IBM], Hermes[UMD];Tsimmis,
InfoMaster[Stanford]; DISCO[INRIA];
Information Manifold [AT&T];
SIMS/Ariadne[USC];Emerac/Havasu[ASU]

Virtuális integrátor architektúra



Források: relációs adatbázisok, weblapok, szövegek.