Networkshop 2009

Szemantikusan annotált tartalom létrehozása intelligens szövegfeldolgozó eszközök támogatásával

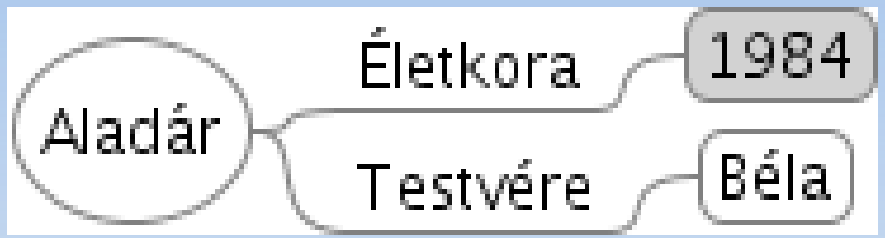
Héder Mihály
MTA SZTAKI

mihaly.heder@sztaki.hu

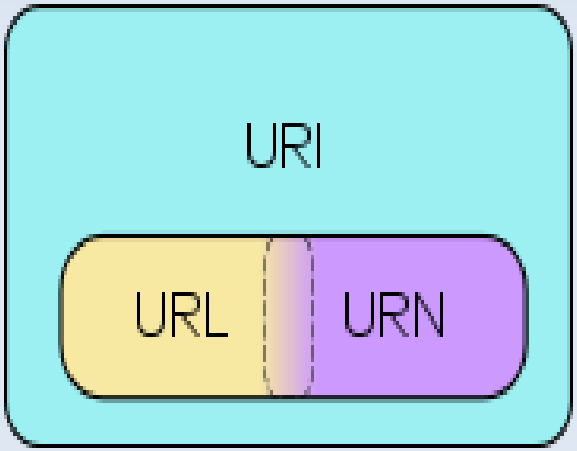
Problémafelvetés

RDF

- RDF Hármas:
  - Alany
  - Reláció vagy tulajdonság
  - Objektum vagy érték



- Az első és második tag egy Uniform Resource Identifier (URI)
- <Aladár>,<születési éve>,<1984>
- <Aladár>,<testvére>,<Béla>
- Reifikált állítások, kontextusok



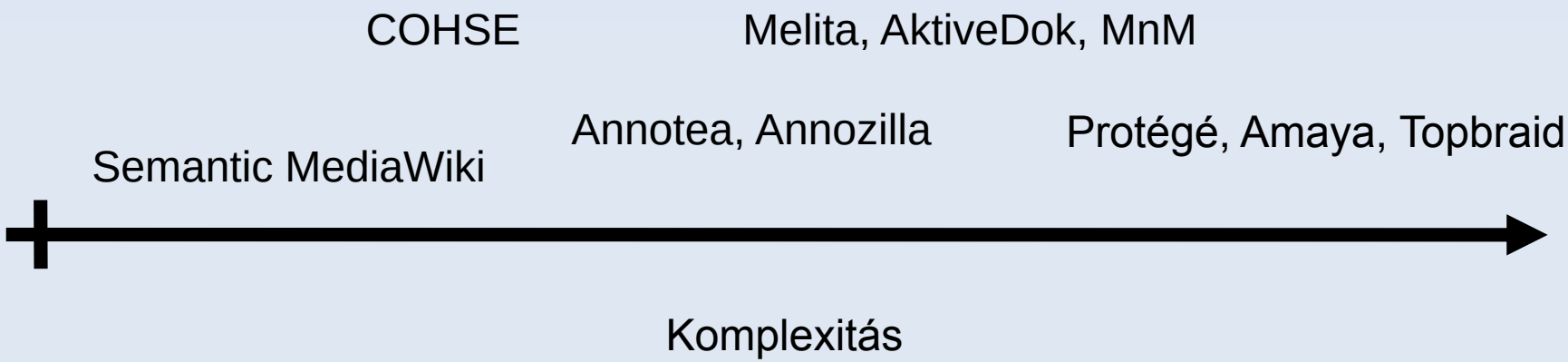
Szemantikus annotáció formátumok

- HTML Metadata
  <META name="author" content="M Héder">
- GRDDL
  grddl:transformation="glean_title.xsl
  http://www.w3.org/2001/sw/grddlwg/td/getAuthor.xsl"
- RDFa
   <h2 property="dc:title">The trouble with Bob</h2> <h3 about="#ch1" property="dc:creator">Alice</h3>

- Microformat: hCalendar, hCard

Annotáló eszközök

- Nagyon sok van
- A kicsit komplextől a nagyon komplexig terjed a skála



Szemantikus Wikik

- Semantic Mediawiki, PHPWiki, IkeWiki, SWiM,MindWiki, Rhizome, SemperWiki, Confluence+wikidsmart
- '''Berlin''' is the capital of [[capital of::Germany]]
  and also its largest city; the city is now home to [[population::3,391,407]], down from a peak of 4.5 million before [[World War II]].
  It measures [[area::891.69 square kilometers]] and has the coordinates [[coordinates::52°31'N; 13°24'E]].
  Berlin is located in the north of [[located in::Germany]]
  [[Category:City]][[Category:sample pages]]

Szakértői eszközök

- Swedt (Eclipse), Apolda (GATE), Katia
- Annotea, Amaya, Annozilla (Firefox plugin)
- Mangrove, Melita, ActiveDoc, MnM
- S-CREAM, OntoMat, QBLS, Cohse, MagPie, Smore
- OntoGloss, WebKB, Protégé, TopBraid

Így visszük be a szöveget

```
[[Kép:Www.wikipedia.org screenshot.png|280px|jobbra|bélyegkép|A Wikipé
nyelvű változatokat]]

A '''Wikipédia''' egy többnyelvű, [[nyílt tartalom|nyílt tartalmú]], a
[[web]]es [[enciklopédia]]. A Wikipédiát a [[Wikimedia Alapítvány]] üz
nonprofit alapítvány –, szerkesztését pedig önkéntes közösség végzi.

A [[Wikipédia:Névjegy|Wikipédia]] magában foglalja a különböző nyelvi
Wikipédia|magyar Wikipédiát]]. Az angol változat 2007. szeptember 9-ér
világ legnagyobb enciklopédikus műve.
A 260 különböző nyelvű változatban összesen (az angollal együtt) több
szócikk|szócikk]] olvasható és szerkeszthető, és több mint 13 mil
<ref>[http://meta.wikimedia.org/wiki/List_of_Wikipedias#Grand_Total Al
Grand Total]</ref>

A ''Wikipédia'' név a [[wiki]] és az [[enciklopédia]] szavakból ered.
```

See below for help in editing this page.

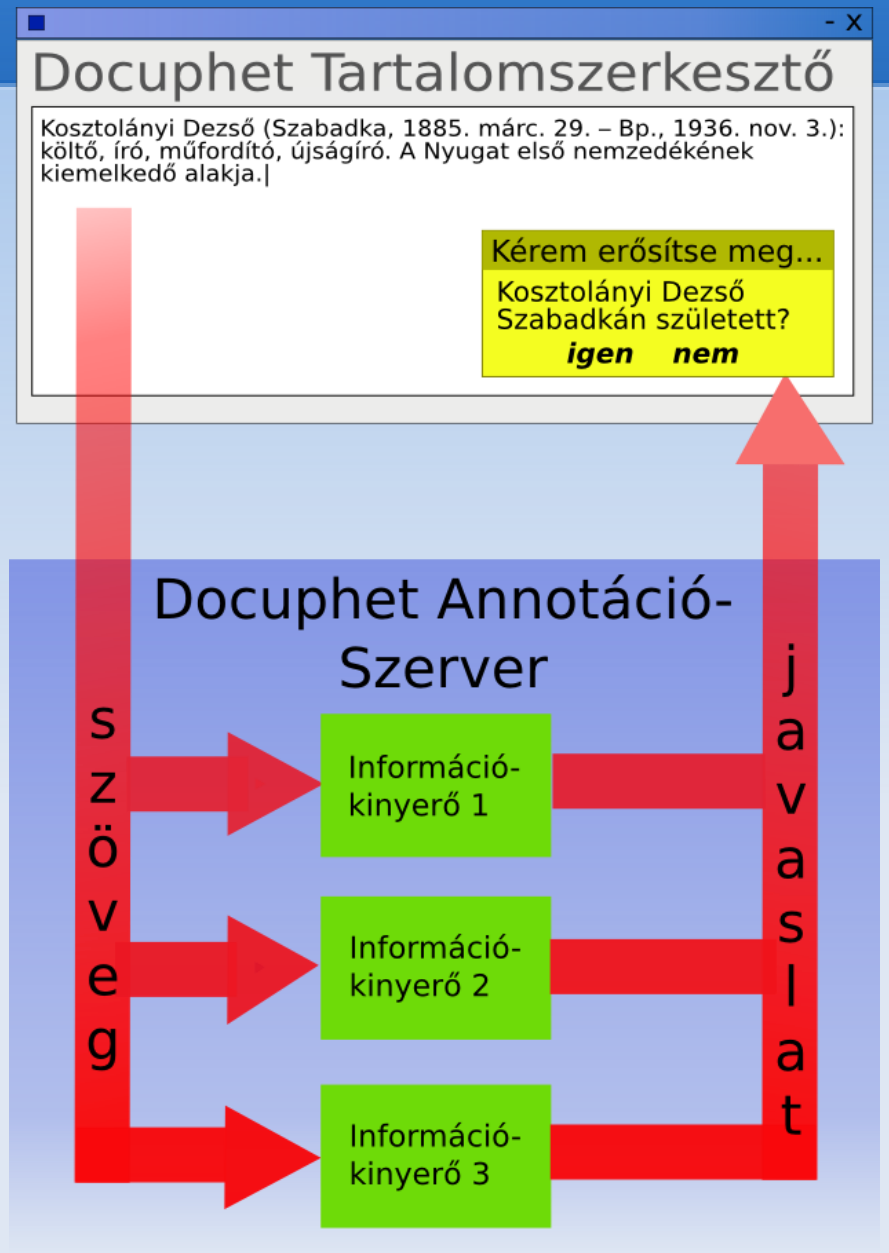
```
-- [[Main.XYAnonimus][Ano]] - 2008.10.5
---++[[Sandbox.FooBar][Here]]

   * Bullet
      * in
         * an other
            * bullet

:scull:

---+++title..

%ATTACHURL%/F3.jpg
```

Editing "Cikk emlékeztetők"

```
== Olvasandó cikkek ==
 * [http://xmlns.com/foaf/spec/#sec-glance]
 * [http://relaxng.org/compact-tutorial-20030326.html]
 * [http://en.wikipedia.org/wiki/Comparison_of_document_markup_lan
 * [http://en.wikipedia.org/wiki/List_of_document_markup_languages
 * [http://www.w3.org/TR/ruby/]
 * [http://www.xml.com/pub/a/2007/06/01/xquery-the-server-language
 * [http://www.xml.com/pub/a/2007/07/12/xquery-and-data-abstractio
 * [http://www.xml.com/pub/a/2002/10/16/xquery.html XQuery] A máso
 * [http://www.w3.org/TR/xquery-use-cases/ XQuery use cases]
 * [http://www.xml.com/pub/a/2007/02/14/introducing-rdfa.html RDFa
 * [http://www.xml.com/pub/a/2007/04/04/introducing-rdfa-part-two.h
 * [http://www.xml.com/semweb/ xml.com szemantikus web]
 * [http://www.xml.com/pub/a/2007/03/14/a-relational-view-of-the-se
```

Docuphet - Áttekintés

1. A szöveg bevitele
2. A szöveg küldése AJAX-szal
3. A szöveg elemzése
4. Javaslatok megfogalmazása
5. Felhasználói megerősítés
6. Annotáció elhelyezése



Formalizálás

- Információkeret - felismerés
- Az információkeret fogalma
  Egy RDF <alany,predikátum,tárgy>
  hármas, amely legalább egyik tagja ismert, a többit változónevek helyettesítik. Az ismeretlenek RDF típusa lehet ismert.
- Példa:
  <X(személy),született, D(dátum)>
- Feladat:
  Információkeret-példányokat felismerni azonosítani a szövegben
- (+Szöveges kérdést formalizálni)

Információkeret-felismerés példa

Ybl Miklós (Székesfehérvár, 1814. ápr. 6. – Bp., 1891. jan. 22.): építész. A bécsi polytechnikum elvégzése után 1832-től Pollack Mihály, 1836-tól Koth Henrik irodájában dolgozott

<X(személy),született, D(dátum)>

<Ybl Miklós,született, 1814. április 6.>

Igaz-e, hogy Ybl Miklós születési dátuma 1814. április 6.?

Felismerési stratégiák: névelemek

- JNER névelemfelismerő keretrendszer
- Bemenet: tokenszekvencia
- Reguláris kifejezések, katalógusok, egyéb programok
- Példa:

token n:”/^[ ]*\p{Upper}.*/”,n+1:(keresztnév kat.)

token n:PastYears.class,n+1 (hónap kat), n+2 : Days.class

- Információkeret:
  <#bekezdés1, kapcsolatos, Ybl Miklós>

Felismerési stratégiák: Mondathatár

- Feltevés: többismeretlenes információkeret egy mondaton belül
- JSentence szabályalapú mondathatár-felismerő.

Felismerési stratégiák: kategória

- Bármely (pl. hitec3) kategorizálómotor
- Információkeret:
  <#, kategória, életrajz>

Problémák, megfontolások

- Az annotációkat létrehozni könnyebb, mint karbantartani
- Mozgatási, szerkesztési problémák
- Mikor kell frissíteni?

Felmerülő probléma: változások kezelése

- Eredeti szöveg:
  A policisztás ovárium szindróma definíciója - szöveg szöveg szöveg szöveg szöveg szöveg
  szöveg szöveg szöveg szöveg szöveg szöveg szöveg szöveg
  szöveg szöveg szöveg szöveg szöveg szöveg
- Új szöveg:
  A policisztás ovárium szindróma korábbi definíciója szöveg szöveg szöveg szöveg szöveg szöveg
  szöveg szöveg szöveg szöveg szöveg szöveg szöveg szöveg
  szöveg szöveg szöveg szöveg szöveg szöveg
- Általában: Elírás javítása, paragrafusok felcserélése, copy-paste... Mi történjen az annotációkkal?

Robusztus Annotációk (Bodain-Robert)

- Robusztus horgonyok
- Részletek elrejtése: elsősorban a szöveg
- Tetszőleges ontológia
- Tetszőleges granularitás
- Frissítések kezelése

Megerősítéses annotációval kapcsolatos javaslatok

- Ugyanazt a kérdést soha ne tegyük fel kétszer
- Minden (elutasított/elfogadott) javaslatot tárolni kell (Hogyan értelmezzük az elutasított javaslatokat?)
- Ne tegyünk fel egyszerre sok kérdést
- Vezessünk be annotáció típusokat a frissítés szükségessége szerint:
  - egyszerű
  - összetett

Annotáció-típusok

- Egyszerű:
  Tartalma csak az annotált szövegrésztől függ.
  Csak akkor kell felülvizsgálni, ha ez változik, pl.:
  <Ybl Miklós, szófaj, személynév>

- Származtatott
  Függ másik (egyszerű vagy származtatott) annotációtól, vagy egyéb feltételtől. Módosítani kell minden esetben, ha a függőségi gráf bármely eleme módosult.

Többismeretlenes, származtatott információkeretek

- Névelemfelismerés + keretfelismerés
- Egy mondaton vagy paragrafuson belül
- Példa:

Alkalmazási példák

- BioBase
  - nevek, alany
  - születési adatok
  - foglalkozás
  - stb...
- FlatBase
  - hely, utcaszinten
  - típus
  - fűtés típus
  - stb...

Összefoglalás

- A szöveges kérdések miatt a lehető legszélesebb felhasználói célcsoport
- Nem építünk ontológiát, fix keretek, korlátozott domain
- Korpusz hiányában nehéz mérni a pontosságot/felidézést
- Az alapfunkciók (névelemek, semantic role labeling) általánosan is használhatóak

Továbbfejlesztés

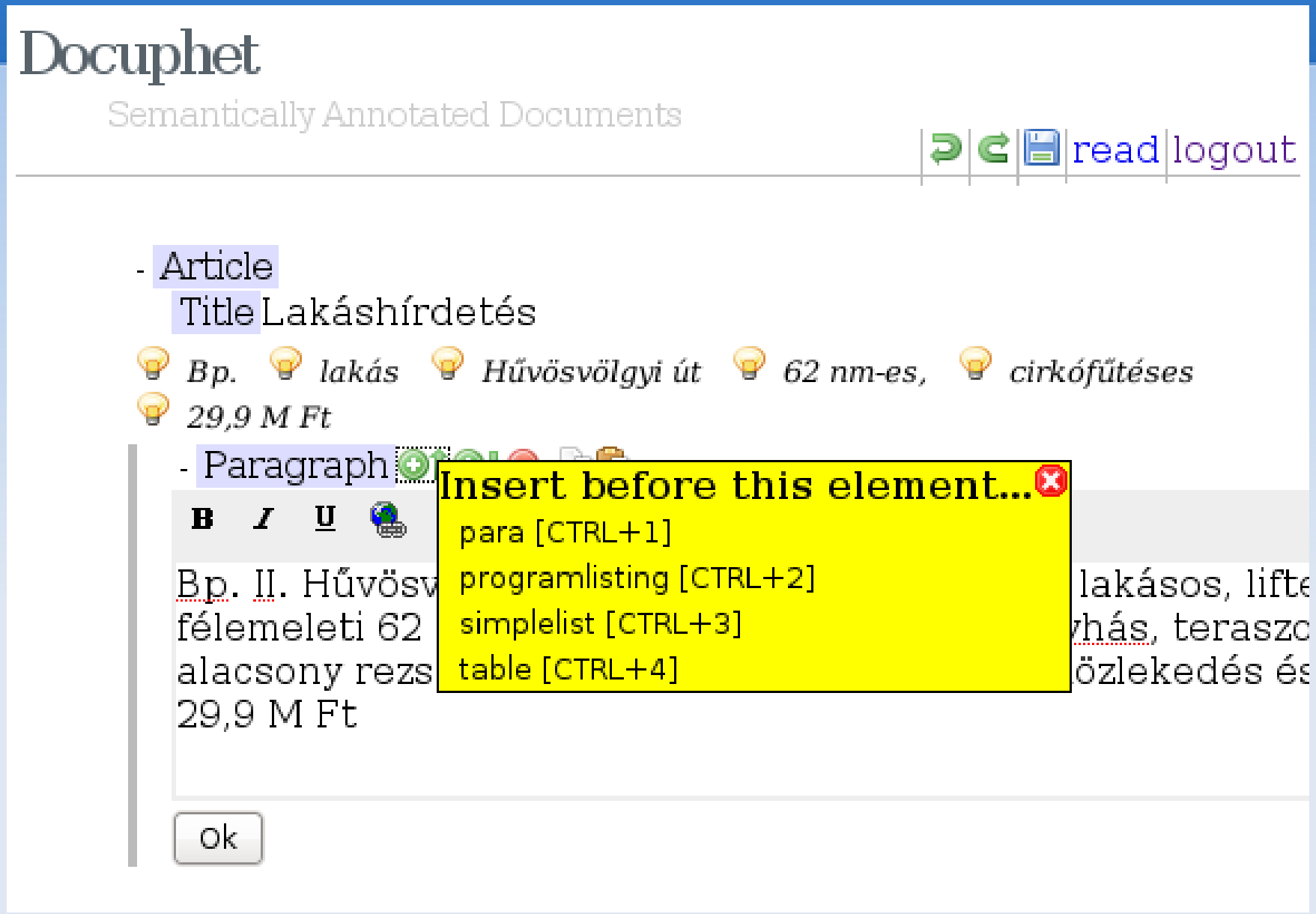
- Több domain feldolgozása, FrameNet adatbázis felhasználása
- Wikipedia szerkesztő
  -alternatív szerkesztő
  -infoboxok kitöltése

Köszönöm a figyelmet!
Kérdések?

Adatbázisos és szöveges adat

- A (relációs) adatbázisok adatainak általában van valamilyen szemantikája, mert lehet tudni, hogy mi az oszlopok jelentése - lásd még: mélyhálós keresés
- A szöveges adattal kapcsolatban alapvetően nem állnak rendelkezésre szemantikus adatok
- De a dokumentumokat el lehet látni annotációkkal

DCE - Screenshot



Frame Semantics

- Szemantikus keretek illesztése a mondatokra
- Tervez keret:
  frame(TERVEZ),
  inherit(ALKOT),
  frame_elements(TERVEZŐ (=ALKOTÓ), ÉPÜLET(=MŰ)),
  scenes(TERVEZŐ tervez ÉPÜLET)
- FrameNet projekt (Berkeley), NewsPro Projekt (MTA NYI, Szegedi TE, Morphologic, Magyar Gallup I.)
- Ez az válhat az információkeretek felismerésének általános módjává

Felismerési Stratégiák
Semantic Role Labeling

- A szöveg nyelvtani elemzése, a szemantikus szerepek felismerése:

  Ybl alany,cselekvő személy,Agent az Operaházat tárgy,Object 1879-ben időhatározó,Date tervezte.állítmány,
  tervez ige, múltidő, egyes szám

BioBase

Docuphet

Semantically Annotated Documents

read | logout

- Article

  Title A cikk címe

1814 ápr. 6. *1891 jan. 22.* *Ybl Miklós* *Székesfehérvár* *1814* *1814 ápr. 6.* *1891* *1891 jan. 22.* *Pollack Mihály* *Koth Henrik* *Batthyány Lajos* *Károlyi György*

  - Paragraph

Ybl Miklós (Székesfehérvár, 1814. ápr. 6. – Bp., 1891. jan. 22.): építész. A bécsi polytechnikum elvégzése után 1832-től Pollack Mihály, 1836-tól Koth Henrik irodájában dolgozott. 1840-től a müncheniak.-n, majd Itáliában képezte tovább magát Hazatérve Pollack Mihály fiával, Ágosttal társuk; közösen építették át gr. Batthyány Lajos ikervári kastélyát, majd Károlyi György és Ede megbízásából építette azok fóti és radványi kastélyát, a kaplonyi és fóti templomot. Első nagy alkotásai a keleti elemekkel tűzdelt romantikus

FlatBase

Docuphet

Semantically Annotated Documents

read | logout

- Article
  Title Lakáshírdetés

Bp. *lakás* *Hűvösvölgyi út* *62 nm-es,* *cirkófűtéses* *29,9 M Ft*

  - Paragraph

Bp. II. Hűvösvölgyi út elején, 1996-ban épült, 8 lakásos, liftes társasházban, félemeleti 62 nm-es, 2 szobás, nagy étkezőkonyhás, teraszos, cirkófűtéses, alacsony rezsijű lakás garázzsal eladó. Kiváló közlekedés és infrastruktúra. Ir.ár: 29,9 M Ft

Szemantikus Web

- Miről szól?
  - A gépek (alkalmazások) legyenek képesek érteni egymás adatait
  - Okosabb keresés
  - Következtetések - ontológiák segítségével