Nagytömegű, strukturált szövegek online szolgáltatása

Király Péter

(Tesuji Magyarország Kft.)

NetworkShop 2006. Miskolc

Miskolci Egyetem
„B” terem, XIX. előadó
2006. fáprilis 20. (csütörtök) 10.[50]

http://www.tesujionline.com

Mi a cél?

- rendkívűl gyors szemantikus keresőmotor
- digitális tartalommegjelenítő rendszer nagy méretű rendezett vagy rendezetlen archívumok számára
- testreszabható indexelőmotor, melyben TEI dtd (próza, vers, dráma, szótár stb), HTML, PDF, MS Word és Excel dokumentumok indexelésére van lehetőség (további a dokumentumtípusok lehetőleg egyszerű felvétele).
- Open Source program, amely a Jakarta Lucene indexelő motorra és a Tomcat servlet containerre épül
- kizárólagosan Javaban írt webes alkalmazás
- szerver oldalon bármilyen Javát támogató operációs rendszerrel kompatibilis, míg kliens oldalon minden elterjedt böngészőn fut

apaclẽto

Adatmennyiségek

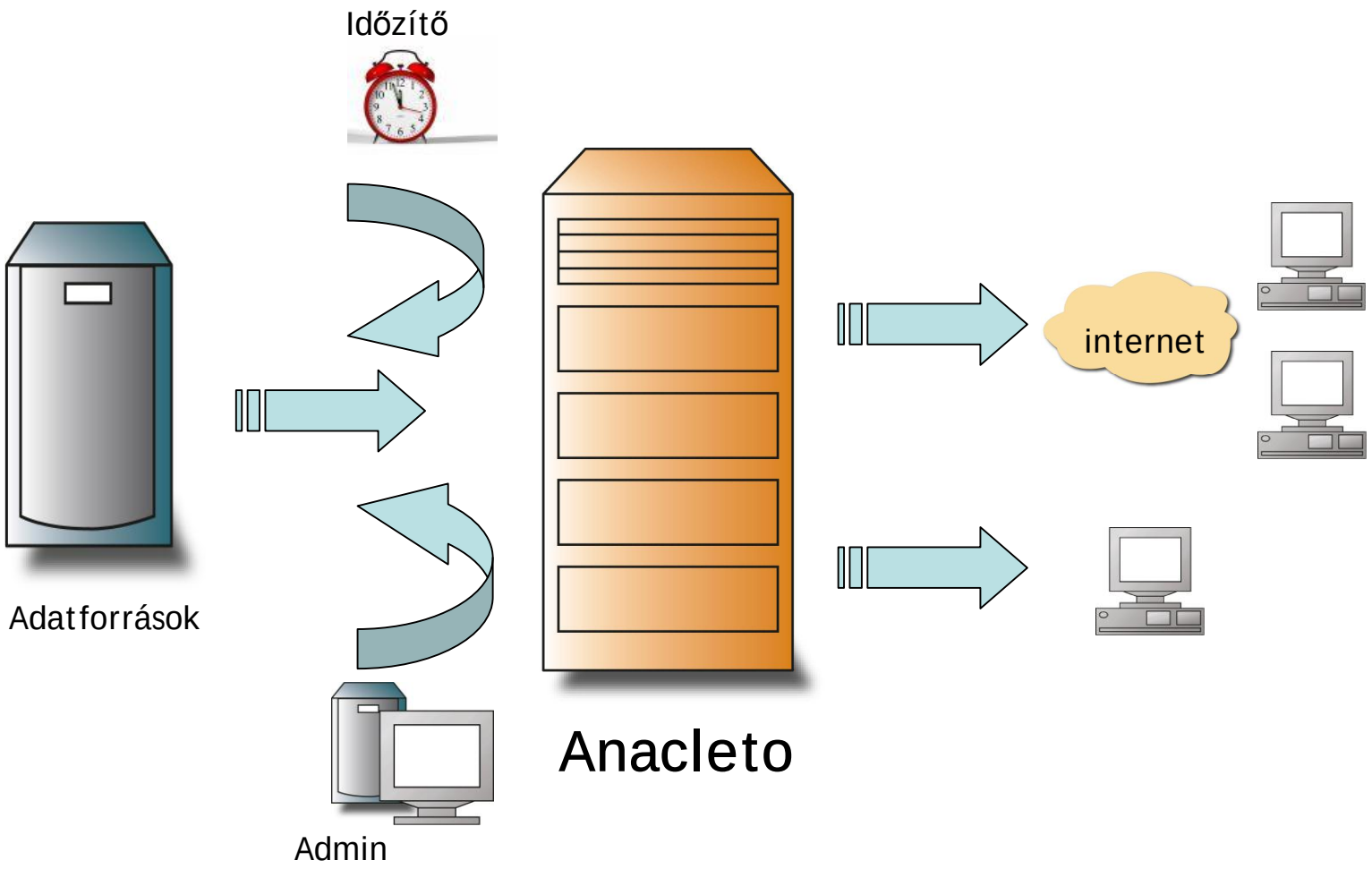
- Arcanum Adatbázis Kft.
  - 7 GB szöveg
  - 1,5 millió HTML oldal
  - leváltott program: NXT3
- Project Gutenberg Consortia Center
  - 8 GB szöveg
  - 100 000 dokumentum (egy könyv ~ 1 HTML lap)
  - leváltott program: HtDig
- Javaportal.it
  - Leváltott program: Exalead

anacleto

Open source összetevők

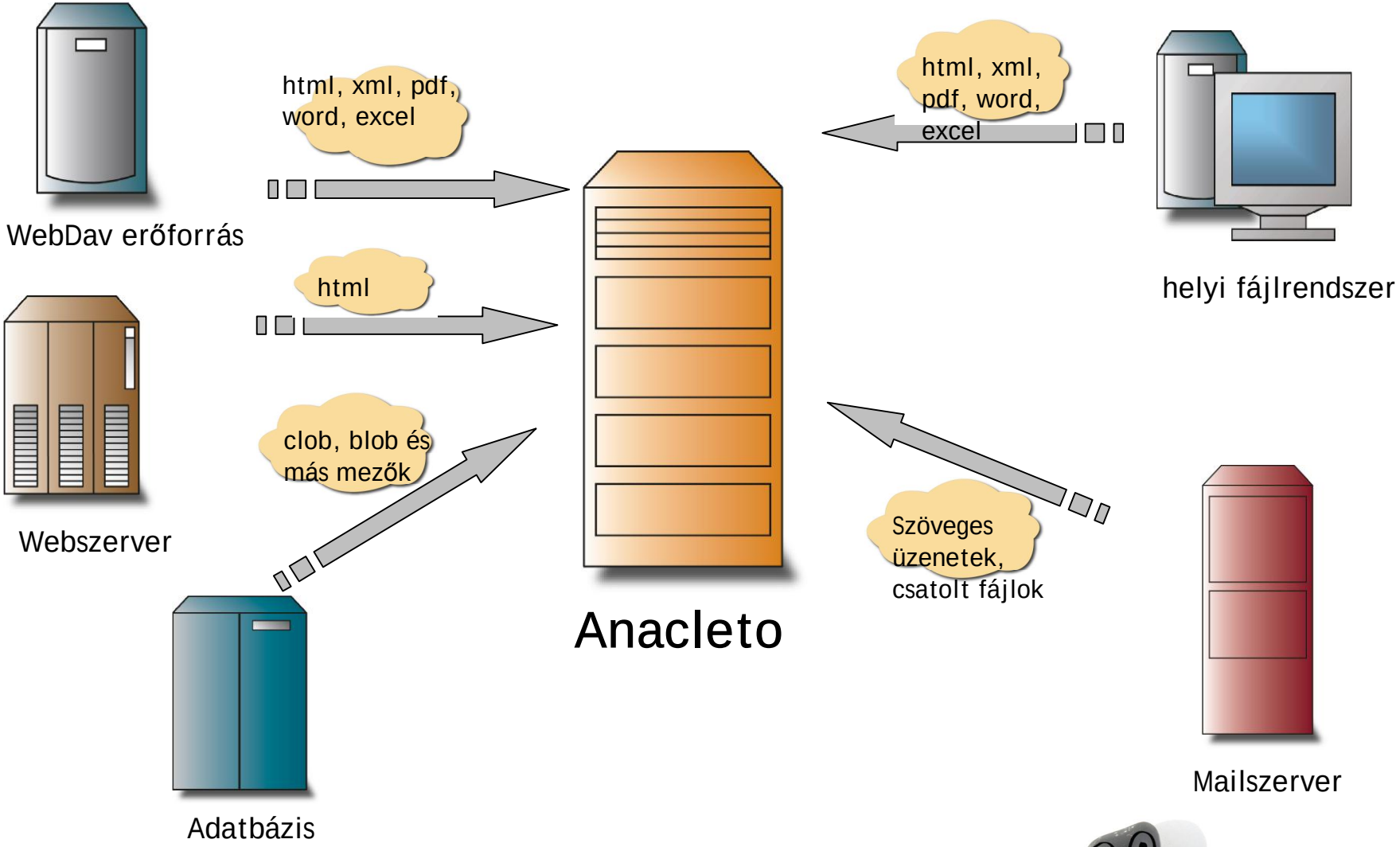
- MVC felépítés: Struts
- Indexelés: Lucene
- Fordítás, telepítés: Ant
- Szerver: Tomcat (más szervlet konténerekkel is)
- egyéb: Jakarta commons, Junit, JDTL, stb.
- IDE: Eclipse, MyEclipse
- Verziókövetés: CVS
- Hibajelentés: Bugzilla
- Dokumentáció: Wiki, OpenOffice

A működés vázlata



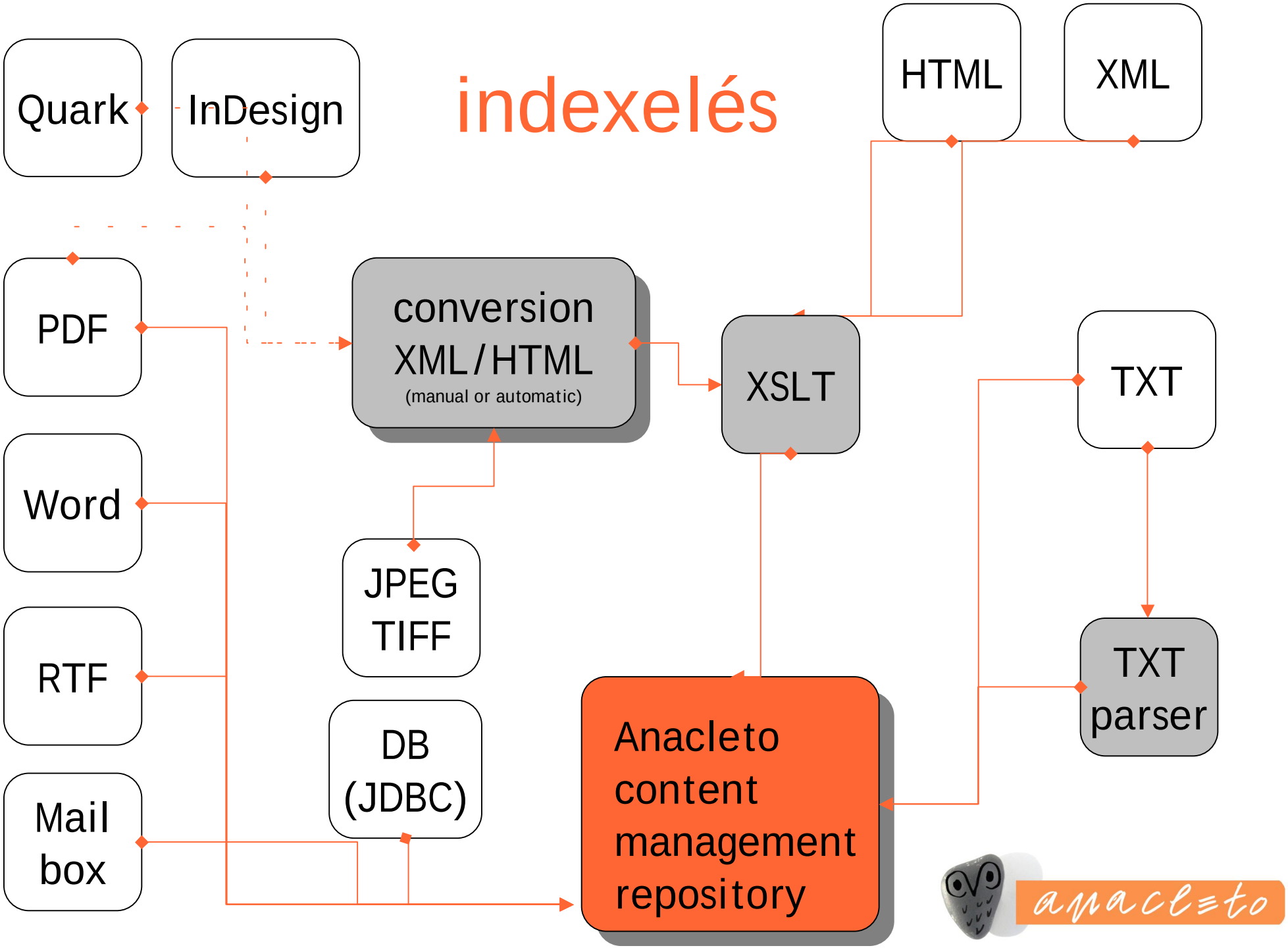
anacleto

Adatforrások



anacleto

indexelés


Quark
InDesign
HTML
XML
PDF
conversion
XML/HTML
(manual or automatic)
XSLT
TXT
Word
JPEG
TIFF
RTF
TXT
parser
DB
(JDBC)
Mail
box
Anacleto
content
management
repository
anacleto

HTML indexelése I.

Folio

<RD:"Könyv"><JD:"01"><PN:"könyv">Das erste Buch Mose (Genesis)</PN>
<RD:Fejezet><PN:"fejezetszám">1. Mose 1</PN>
<RD:Vers><JD:"01:1.1"><JD:"1. Mose 1.1"><PN:"versszám">1. Mose 1.1</PN>
<RD><PN:"SZÖVEG">Am Anfang schuf Gott Himmel und Erde. *</PN>

<RD:"Könyv"> = 'level' HTML: H1....H6, jelöletlenül a '<p>'-nek felel meg

<PN:"SZÖVEG"> = mező, szemantikus jelölőelem HTML: span class="..."

<JD:"01:1.1"> = kereshető ugrópont, HTML: <a name="..."> vagy újabban id="..."

anacleto

HTML indexelése II.

HTML

```
<H1 class="Konyv"><a name="JD_01"></a><span class="id" type="field" title="id">01</span><span class="biblebook" type="field" title="biblebook">Das erste Buch Mose (Genesis)</span></H1>

<H2 class="Fejezet"><span class="chapter" type="field" title="chapter">1. Mose 1</span></H2>

<H3 class="Vers"><a name="JD_01_1_1"></a><span class="id" type="field" title="id">01:1.1</span><a name="JD_1__Mose_1_1"></a><span class="verse" type="field" title="verse">1. Mose 1.1</span></H3>

<p class="rd"><span class="text" type="field" title="text">Am Anfang schuf Gott Himmel und Erde. *</span></p>
<p class="rd"><a href="showDocument.do?name=deutsch492#JD_22_38_4" class="link_Hiv"> * Hiob 38,4;</a></p>
```

Am Anfang schuf Gott Himmel und Erde. *

text

<span
class="text" a class neve 'text', erre lehet CSS stílust illeszteni és ez lesz a mező neve
type="field" típusa field, vagyis ez egy mező, amit le kell indexelni
title="text" a title minden HTML tagben opcionális attribútum, a gyorstippnek
>

HTML in

XSL

```
<xsl:attribute-set name="text" >
   <xsl:attribute name="indexed"          >true</xsl:attribute>
   <xsl:attribute name="stored"          >false</xsl:attribute>
   <xsl:attribute name="tokenized"        >true</xsl:attribute>
   <xsl:attribute name="termVectorStored">false</xsl:attribute>
</xsl:attribute-set>
```

ha a span type attribútuma 'field'

```
<xsl:template match="span[@t
```

saját xslt extension

```
  <xsl:element name="in:index" use-attribute-sets="text">

    <xsl:attribute name="field" ><xsl:value-of select="@class" /></xsl:attribute>

    <xsl:apply-templates/>

  </xsl:element>

</xsl:template>
```

a mező nevét a class attribútumból kell kinyerni

a H1...Hn tagek címek, még ha nincsenek is megjelölve

```
<xsl:template match='H1|H2|H
```

```
<xsl:attribute-set name="keyword" >
   <xsl:attribute name="indexed"          >true</xsl:attribute>
   <xsl:attribute name="stored"          >true</xsl:attribute>
   <xsl:attribute name="tokenized"        >true</xsl:attribute>
   <xsl:attribute name="termVectorStored">true</xsl:attribute>
</xsl:attribute-set>
```

```
  <xsl:element name="in:index
```

Le kell indexelni úgy is, hogy a 'bárhol' keresésre is e

```
    <xsl:attribute name="field">

    <xsl:element name="in:index" use-attribute-sets="keyword">

      <xsl:attribute name="field">title</xsl:attribute>
```

rekurzív indexelés (az XSLT erőssége)

```
      <xsl:apply-templates/>

    </xsl:element>

  </xsl:element>
```

Kép indexelése

XMP adat a képi állományban


```
<x:xmpmeta xmlns:x="adobe:ns:meta/" x:xmptk="3.1.1-111">
  ...
  <dc:description><rdf:Alt><rdf:li xml:lang="x-default">OSZK Quart. Lat. 16.</rdf:li></rdf:Alt>
</dc:description>
  <dc:creator><rdf:Seq><rdf:li>Latsny Adam</rdf:li></rdf:Seq></dc:creator>
  <dc:title><rdf:Alt><rdf:li xml:lang="x-default">Catalogus librorum, dissertationum et manuscriptorum
variorum ad rem Hungaricam praecipue facientium ex bibliotheca, quae Vitebergae est, Hungarorum
congestus ab Adamo Latsny Turotzensi. … Vitebergae Saxonum … an[no] …
1755.</rdf:li></rdf:Alt></dc:title>
```


Kinyerés és feldolgozás

Description : OSZK Quart. Lat. 16.

Creator : Latsny Adam

Title : Catalogus librorum, dissertationum et manuscriptorum variorum ad rem Hungaricam praecipue facientium ex bibliotheca,
quae Vitebergae est, Hungarorum congestus ab Adamo Latsny Turotzensi.… Vitebergae Saxonum … an[no] … 1755.

Legyenek-e mezők?

	mező nélkül	mezővel
konverzió	Félig vagy teljesen automatizálható	bonyolult, félautomatikus eljárások emberi tudást VAGY nagyon különleges szoftvereket igényel
	egyszerűbb módszer	Bonyolultabb módszer
keresés	teljes szövegű keresés	teljes szövegű keresés
		Mező szerinti keresés (cím, szerző, stb.)
	“részecske”	“részecske” a címben “molekula” a képaláírásban “MIT” az összefoglalóban
találatok	kisebb relevancia a szöveg bármely részén (a szövegnek nincsenek kitüntettett fontosságú részei)	nagyfokú relevancia a szöveg meghatározott részén

Karakterkészletek

Belső adatábrázolás

A Java mindent UTF-ben tárol, a külső forrásból érkező adatokat az esetek többségében automatikusan konvertálja

Különleges esetekre bőségesen léteznek '3rd party' konverterek

pl. az OszK katalógusa 'ANSEL' nevű, az Unicode-hoz hasonló, de attól számos helyen eltérő készletet használ

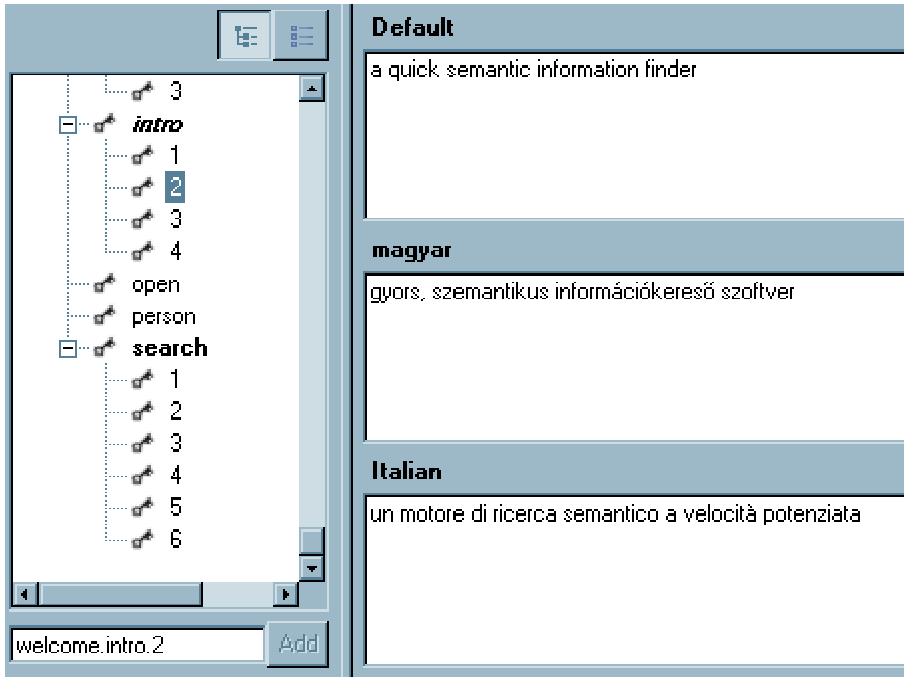
Webfelület

UTF-8. A bemenő adatokat az Anacleto ellenőrzi és adott esetben konvertálja

I18n — „nemzetköziesítés”

A Java alaptulajdonsága a 'Locale' (helyi beállítások) figyelembevétele. Vannak lokalizált (standard név=lokalizált érték párokból álló) tulajdonság fájlok, amiket a rendszer a böngésző nyelvi beállítása alapján választ ki. Jelenleg az Anacletoban angol, magyar és olasz lokalizáció érhető el.

```
<li><bean:message key="welcome.intro.2" /></li>
<li><bean:message key="welcome.intro.3" /></li>
<li><bean:message key="welcome.intro.4" /></li>
```



Hierarchia

polcok
- könyvek
  - dokumentumok
    - dokumentumok ('ad infinitum')

Minden elemhez tartozhat:
- saját indexséma
- dokumentumtípus-meghatározás (ami alapján az Anacleto kezelni fogja)
- saját stíluslap
- polcok és könyvek esetében: címoldal

A dokumentum lehet virtuális is, pl. egy XML fájlt virtuálisan feloszthatunk több apró részre, ami azt jelenti, hogy a tartalomjegyzékben, találati listán, dokumentum-nézetben külön-külön látjuk a fizikailag egyazon fájlban lévő részeket.
Eredetileg nem hierarchikus dokumentumokból (pl. egy adatbázis tábla) is lehet hierarchikusat csinálni (pl. valamilyen elven való csoportosítással, kezdőbetűkkel stb.)

A keresés

Egyszerű keresés, az archívumban szereplő összes kifejezésre kereshetünk

Mező szerinti keresés, az indexelés során mezőket lehet meghatározni, amelyeket külön és más keresésekkel kombinálva is lehet keresni (*szerző: Iván és tartalom: Duna*)

pontos kifejezés *(„Kossuth Lajos”)*

hasonlóságon alapuló keresés *(Kossuth, Kosuth stb.)*

közelségi keresés *(Kossuth és Lajos szavak 5 szó távolságra, bármely sorrendben)*

Google szintaktika alkalmazása a keresőkérdések egymáshoz való viszonyának leírásához (‘és’, ‘vagy’, ‘nem’)

A keresés hatókörének szűkítése az archívum egy megadott részére *(„Kossuth a 2000-es évfolyamban”)*

Ékezetek helyettesítésen alapuló indexelése szabadon beállítható a forrásdokumentumban helyesen ‘Košice’ amit a ‘Kosice’ keresőkérdés is megtalál

Szóelemzéses keresés és ontológia (tezaurusz) beépíthető (‘alma’ megtalálja az ‘almát’, ‘almának’ alakokat, az ‘úszásnem’ a ‘gyorsúszás’, ‘pillangó’ alakokat)

apaczeto

Felület I. Tartalomjegyzék

```
--≤ Arcanum szövegtár
  |-≤ lexikonok, szótárak
  |  |-± Pallas nagy lexikona
  |  |-± A Magyar Nyelv Értelmező Szótára
  |  |-± Czuczor-Fogarasi: A magyar nyelv szótára
  |  |-± Finály Henrik: A latin nyelv szótára
  |  |-± Magyar Életrajzi Lexikon
  |  |-≤ Pecz: Ókori lexikon
  |  |  |- Előszó az első kötehez
  |  |  |- Előszó a második kötethez
  |  |  |- Előszó az új kiadáshoz
  |  |  |-± A
  |  |  |-± B
  |  |  |-± C
  |  |  |-± Cs
  |  |  |-± D
  |  |  |-≤ E
  |  |  |  |- Ebora
  |  |  |  |- Eboracum
```

hierarchia böngészése

dinamikus

egyszerre 50 elem (a szerver és a felhasználó kímélése céljából, pl. a Pallasban akár 10 000 gyermek eleme is van egy-egy vaskosabb betűnek

oda vissza szinkonizálható a dokumentummal

keresésre szűkített tartalomjegyzék

```
--< (383) Arcanum szövegtár
   |-< (223) lexikonok, szótárak
   |  |-+ (45) Pallas nagy lexikona
   |  |-+ (3) Czuczor-Fogarasi: A magyar nyelv szót
   |  |-+ (26) Finály Henrik: A latin nyelv szótára
   |  |-+ (2) Magyar Életrajzi Lexikon
   |  |-< (143) Pecz: Ókori lexikon
   |  |  |-< (15) A
   |  |  |  |- (1) Academia
... alatt álló 12 szent olajfa ( stb. Itt tanított
Plato és tanítványai, a kiket aztán ...
   |  |  |  |- (1) Acheron
... folyón kellett átvonulniok az árnyaknak; v. ö.
Plato Phaedonjában a leirást. Valószínüleg az a ...
   |  |  |  |- (1) Aeschines
... és védőbeszédek irásával foglalkozott. A
Plato-féle dialogusokhoz csatolt, A.-nek ...
```

csak azok az ágak szerepelnek, amikre van találat

a találat kontextusa (KWIC)

a keresőkérdés ki van világítva

navigáció

előző/következő dokumentum

előző/következő dokumentum a találati listán

előző/következő találat a dokumentumon belül

vissza a találati listára

találat kijelzése (pl. 62/383)

a tartalomjegyzék szinkronizálása

nyomtatás

könyvjelző

teljes eléséri útvonal kijelzése
Arcanum szövegtár > lexikonok, szótárak > Finály Henrik: A latin nyelv szótára > M > Menosca, ae, nn.
a csomópontok linkként szerepelnek!

anacleto

találati lista

keresés szűkítése, finomítása

előző/következő találati lista

lista elemei számosságának módosítása

logaritmikus lapozás
1- · 31- · 41- · 51- · 61- · 71- · 81- · 91- · 161- · 261- · 361- · 381-

találatok szókörnyezetének megjelenítése

a keresett kifejezés kivilágítása

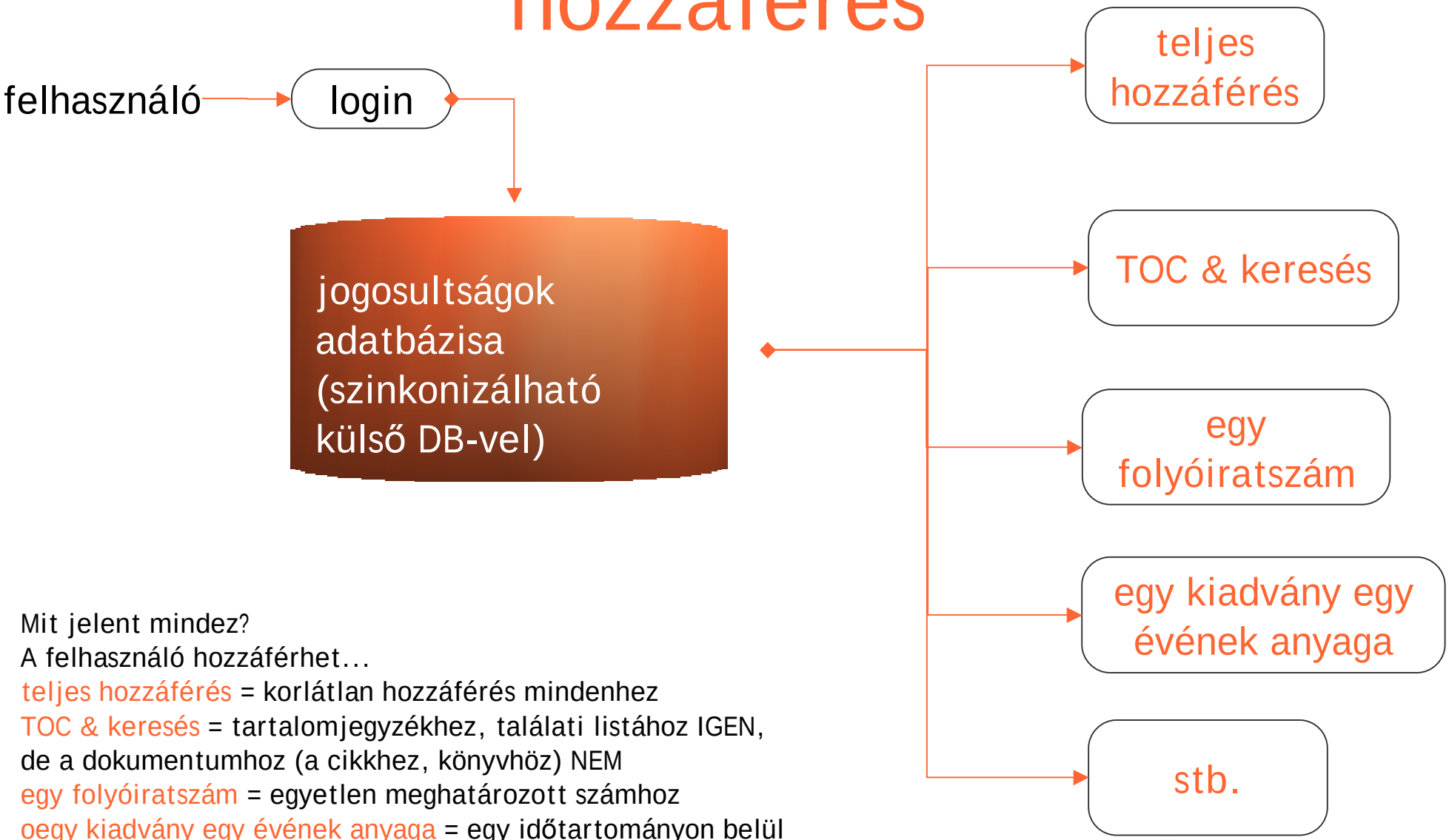
találatokra szűkített tartalomjegyzék behívása

adminisztrációs felület

- alapbeállítások (könyvtárak, fájlok, rendezés, naplózási szintek stb.)
- indexelés, index karbantartás, index megtekintése
- naplók megtekintése
- felhasználók és jogok beállítása
  IP-filter
- ékezet-konverziók
  Kosice eredetileg Košice, most mind a két alak kereshető
- stopszavak

anacleto

hozzáférés



Mit jelent mindez?
A felhasználó hozzáférhet…
teljes hozzáférés = korlátlan hozzáférés mindenhez
TOC & keresés = tartalomjegyzékhez, találati listához IGEN, de a dokumentumhoz (a cikkhez, könyvhöz) NEM
egy folyóiratszám = egyetlen meghatározott számhoz
oegy kiadvány egy évének anyaga = egy időtartományon belül (pl. 2004.V.-2005. IV.) az adott kiadvány minden cikkéhez

Szoftver-követelmények

Böngésző:

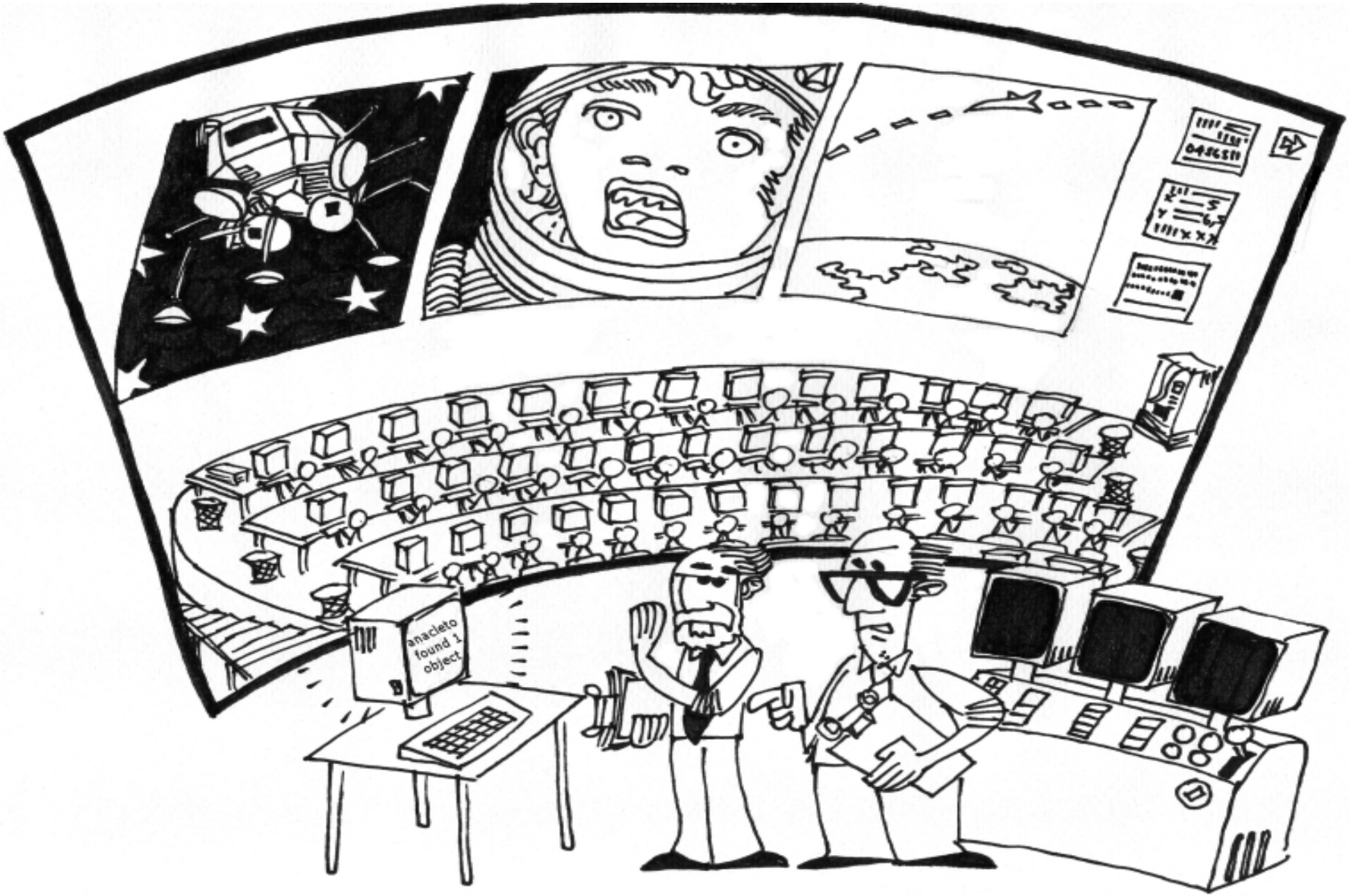
IE5.0+, Firefox, Opera 7.5+, Safari

Java alkalmazásszerver:

bármelyik (JBoss-szal es Tomcat 5.0, 5.5-el teszteltük)

Operációs rendszer:

MS Windows, NT, Linux, Solaris (minden olyan op. rendszer, amelyhez létezik Java támogatás)



Anacleto found 1 object