

**InformaticaUmanistica**

# Esercitazione Funzionalità avanzate di Greenstone

*Pasquale Savino*

*ISTI - CNR*



UNIVERSITÀ DI PISA

# Greenstone

## Esercitazioni Parti VI - IX

# Sommario

- ◆ **Scopo dell'esercitazione**
- ◆ **Il classificatore PHIND**
- ◆ **Indici per full-text di frasi e fielded search**
- ◆ **Uso del classificatore AZCompactList**
- ◆ **Modifica dell'interfaccia di presentazione**

# Scopo dell'esercitazione

- ◆ **Familiarizzarsi con l'uso di diversi classificatori usati in Greenstone**
- ◆ **Differenze tra la ricerca di frasi basata su PHIND e la ricerca di frasi basata su fielded search**
- ◆ **Personalizzazione dell'interfaccia di presentazione della collezione**

# Browsing su frasi

- ◆ Associare metadati strutturati per la classificazione può essere molto oneroso. Se questa informazione non è disponibile, in Greenstone può essere utilizzato il browsing su frasi presenti nel documento
- ◆ Frase: una sequenza di parole che appaiono più di una volta nella collezione
- ◆ L'estrazione avviene automaticamente
- ◆ Key phrases
- ◆ Browser di frasi
  - Le frasi sono organizzate gerarchicamente
  - Ordinate per documento e per frequenza all'interno della collezione
  - Le foglie di questa gerarchia sono i documenti
- ◆ Esempi: [FAO Collection](#), [The Complete Works of Shakespeare](#)
  - ➔ <http://www.sadl.uleth.ca/nz/cgi-bin/library?a=p&p=about&c=fi1998>
  - ➔ <http://www.sadl.uleth.ca/nz/cgi-bin/library?a=p&p=about&c=allshake>

Search frasi che contengono la parola "locust"

Vengono visualizzate le frasi con informazioni sul numero di documenti

Seleziono una delle frasi ("desert locust") e trovo tutte le frasi che la contengono

phind classifier demo

HOME HELP PREFERENCES

topics

search titles a-z language topics

Search for  Previous Next

locust (first 10 of 102 phrases)

	docs	freq
<b>Desert locust</b>	719	1787
locust numbers	127	285
migratory locust	89	192
locust operations	84	112
locust infestations	64	110
locust situation	66	94
locust damage	50	85
locust control	45	84
locust infestation	46	76
forecast locust	25	59
<a href="#">Get more phrases</a>		

Desert locust (first 10 of 102 phrases)

	docs	freq
Desert locust situation	140	188
Desert locust adults	99	153
solitary Desert locust	60	112
Desert locust activity	75	103
Desert locust infestations	60	86
Desert locust control	41	59
Desert locust breeding	39	49
FAO Desert locust	44	45
isolated Desert locust	22	40
Desert locust Bulletin	32	36
<a href="#">Get more phrases</a>		

phind classifier demo

HOME HELP PREFERENCES

topics

search titles a-z language topics

Search for  Previous Next

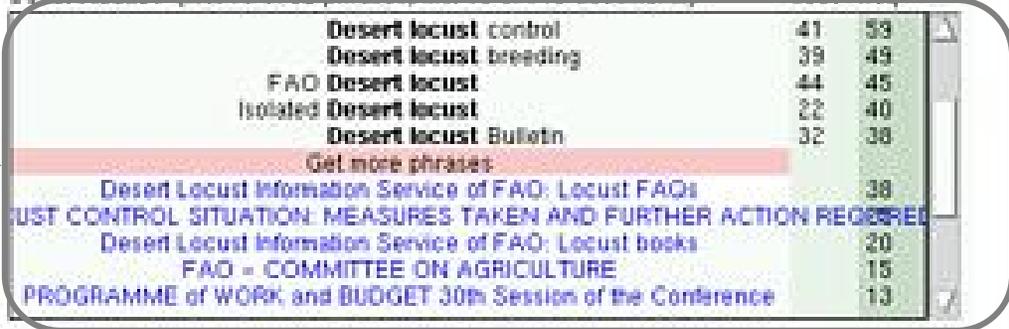
locust (first 10 of 102 phrases)

	docs	freq
<b>Desert locust</b>	719	1767
locust numbers	127	285
migratory locust	89	192
locust operations	84	112
locust infestations	64	110
locust situation	66	94
locust damage	50	65
locust control	45	84
locust infestation	46	76
forecast locust	25	59
<a href="#">Get more phrases</a>		

Desert locust (first 10 of 87 phrases, first 10 of 719 documents)

	docs	freq
Desert locust control	41	59
Desert locust breeding	39	49
FAO Desert locust	44	45
Isolated Desert locust	22	40
Desert locust Bulletin	32	38
<a href="#">Get more phrases</a>		
<a href="#">Desert Locust Information Service of FAO: Locust FAQs</a>		38
<a href="#">LOCUST CONTROL SITUATION: MEASURES TAKEN AND FURTHER ACTION REQUIRED</a>		20
<a href="#">Desert Locust Information Service of FAO: Locust books</a>		20
<a href="#">FAO - COMMITTEE ON AGRICULTURE</a>		15
<a href="#">PROGRAMME of WORK, and BUDGET 30th Session of the Conference</a>		13

Scendendo nella gerarchia di frasi posso arrivare a documenti che contengono la frase



Quindi posso  
visualizzare il  
documento



\_testTopicpage\_ - Netscape  
File Edit View Go Window Help

HOME HELP PREFERENCES

search titles a-z language topics

 DESERT LOCUST INFORMATION SERVICE  
of the MIGRATORY PESTS GROUP

[Locust](#) [Mapper](#) [News](#) [Archives](#) [Plagues](#) [Activities](#) [Books](#) [FAQ](#)

### Frequently Asked Questions (FAQs) about Desert Locusts

- [What is the difference between locusts and grasshoppers?](#)
- [What is a Desert Locust?](#)
- [What countries are affected by the Desert Locust?](#)
- [Do Desert Locust plagues occur with any regularity?](#)
- [How long does a Desert Locust live?](#)
- [How many eggs does a Desert Locust female produce?](#)
- [How far and how fast can Desert Locusts migrate?](#)
- [How big are swarms and how many locusts are there in a swarm?](#)
- [What percentage of the Desert Locust's exoskeleton is chitin?](#)
- [How much food can a Desert Locust eat?](#)
- [What is the relationship between locusts and ecology?](#)
- [Why do locusts change their behaviour?](#)
- [Are there other important species of locusts?](#)
- [Can locusts hurt humans?](#)
- [How can locusts be controlled?](#)
- [Who carries out locust control operations?](#)
- [Are there any non-chemical ways to kill locusts?](#)
- [Can locusts be detected by satellites?](#)
- [Why are Desert Locust so difficult to control?](#)
- [Do people really eat locusts?](#)
- [What is a Desert Locust composed of?](#)

Document Done

# Form-based searching

- ◆ **Possibilità di combinare ricerche su campi di metadati diversi**
  - Ad es. Creator = “Salton” AND Title = “Information Retrieval”
- ◆ **Ricerca full text su parole singole e su frasi**
- ◆ **Nella modalità “advanced search” è possibile ordinare i risultati in base alla rilevanza con l’interrogazione e specificare se va usato lo stemming delle parole**



# bibliography collection

[HOME](#) [HELP](#) [PREFERENCES](#)

## search

[search](#) [titles a-z](#) [authors a-z](#) [dates](#) [phrases](#)

Search for  of

**Word or phrase**

Clear Form

**... in field**

- full records
- full records
- Creator
- BookConfOnly
- Source
- Number
- Month
- Abstract**
- Keywords
- Date
- JournalsOnly
- EntryType
- Keyword
- Volume
- Title
- Year
- PublisherAddress
- Booktitle
- Edition
- Editor
- Pages
- Chapter
- Publisher
- Author
- Note
- Journal

# bibliography collection

HOME HELP PREFERENCES

## search

search titles a-z authors a-z dates phrases

Search and display results in **ranked** order

Word or phrase	(fold, stem)	... in field
Benson	<input type="checkbox"/> <input type="checkbox"/>	Creator <input type="text"/>
and <input type="text"/> learn	<input type="checkbox"/> <input checked="" type="checkbox"/>	Title <input type="text"/>
and <input type="text"/>	<input type="checkbox"/> <input type="checkbox"/>	BookConfOnly <input type="text"/>
and <input type="text"/>	<input type="checkbox"/> <input type="checkbox"/>	Source <input type="text"/>

Or enter a query directly:

### results

Word count: Benson: 2, learn: 625

1 document matched the query.

 **Inductive learning of reactive action models** - *Scott Benson* - 1995

# Form-based searching

- ◆ **Inserire nel file di configurazione (collect.cfg)**

```
Searchtype form
```

```
Indexes nome-indici
```

# Esercitazioni

## ◆ Parte VI – Utilizzo dei classificatori

- Creazione del classificatore PHIND per la ricerca di frasi
- Creazione di un indice per la ricerca di frasi
  - ➔ **l'indice che creeremo ora permette di specificare esplicitamente le frasi da cercare**
- Creazione di partizioni (sottocollezioni) della collezione in base al nome dei file
  - ➔ **Si definiscono dei filtri che permettono di selezionare tutti i file contenuti in determinate cartelle**
  - ➔ **Si creano le partizioni utilizzando i filtri creati**
  - ➔ **Si creano degli indici di ricerca per ogni partizione**

# Esercitazioni

## ◆ Parte VI (cont.)

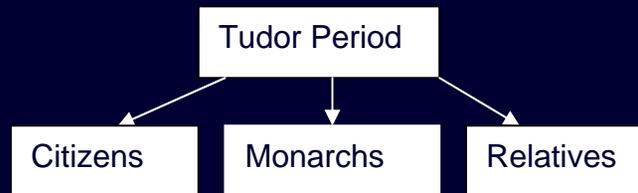
### ■ Esercizio

- Si associ ad alcuni file della collezione un certo valore per dc.Creator, mentre ad altri si associ un valore diverso
- Si provi a creare delle partizioni per i file con i due autori
  - [Si usi il panel Enrich per associare a gruppi diversi di file un valore per dc.Creator
  - Si usi il panel Design, sezione Partition Index per creare i filtri e le partizioni]

# Esercitazioni

## ◆ Parte VI (cont.)

- Creare una gerarchia di soggetti (dc.Subject)



- Si usa la possibilità di specificare valori gerarchici per i metadati usando il carattere “\”
  - ➔ Ad es. “Tudor Period\Citizens”
- Si crea un classificatore che permette di visualizzare la gerarchia di soggetti ed i documenti associati ad ogni soggetto (Hierarchy classifier)

# Esercitazioni

## ◆ Parte VI (cont.)

### ■ Esercizio

- Si provi ad associare una data a gruppi diversi di documenti
- Si crei uno Hierachy classifier per le date
- Come vengono visualizzate le date? Si ottiene la seguente visualizzazione?

Microsoft Internet Explorer window showing a web application interface for "Dates". The browser title is "Dates - Microsoft Internet Explorer". The address bar shows the URL: `http://localhost/gsdll?e=d-0-00-htmlarg--00-0-0Date--0prompt-10---4-----0-11--1-en-50---20-about---00031-001-1-OutfZz-E`. The page content includes a navigation menu with "HOME", "HELP", and "PREFERENCES". Below this, there are tabs for "search", "titles a-z", "filenames", and "dates". The "dates" tab is active, showing a hierarchical view of documents. The hierarchy starts with "2001", followed by "December", and then "24". Under "24", there are several document titles with their corresponding HTML filenames: "The Right to Display Public Domain Images (art.html)", "Tudor England: Bibliography (biblio.html)", "Primary Sources - 1535, the executions of Fisher, More & others (1535exec.html)", "1549 - Edward the sixth's journal, 1 (ed1.html)", and "Contemporary descriptions of Anne Boleyn (annedesc.html)". The browser's status bar at the bottom indicates "Intranet locale".

# Esercitazioni

- ◆ **Parte VII – Creazione di una BD di record MARC**
  - Uso del classificatore AZCompactList che crea dei gruppi per i documenti che hanno uno stesso valore per il metadato
  - Semplice modifica della visualizzazione del classificatore AZCompactList

# Esercitazioni

- ◆ **Parte VIII – Partizionamento dell'indice full-text sulla base di valori dei metadati**
  - In questa esercitazione useremo nuovamente la collezione Tudor e partizioneremo l'indice per la ricerca full-text in quattro parti separate.
  - Per fare questo definiremo 4 sottocollezioni ottenute “filtrando” i documenti sulla base dei valori dei metadati nel campo **dc.Subject and Keywords**.
  - Quindi assegneremo un indice ad ogni sotto-collezione. Questo ci permetterà di limitare la ricerca ad un sottoinsieme dei documenti.
  - **Attenzione:** c'è un errore nelle note dell'esercitazione VIII
    - ➔ Nella collezione tudor non è stato mai assegnato il valore “Tudor Period|Others” al metadato dc.Subject

Tudor - Microsoft Internet Explorer

File Modifica Visualizza Preferiti Strumenti ?

Indirizzo http://146.48.85.78:1025/gsd/?a=p&p=about&c=tudor&l=en&uq=1164126409031

Google Cerca

Segnalibri 658 bloccati Controllo Impostazioni

Collegamenti HotMail gratuita Personalizzazione collegamenti Windows WindowsMedia Channel Guide Hotmail Il meglio del Web

Indietro Cerca Preferiti Adobe

HOME HELP PREFERENCES

# Tudor

about

search titles a-z filenames

Search for  of  that contain  of the words

citizens  
monarchs  
others  
relatives

Begin Search

## About this collection

Tudor

### How to find information in the Tudor collection

There are 3 ways to find information in this collection:

- search for particular words
- access publications by title
- access publications by filename

You can *search for particular words* that appear in the text from the "search" page. This is the first page that comes up when you begin, and can be reached from other pages by pressing the *search* button.

You can *access publications by title* by pressing the *titles a-z* button. This brings up a list of books in alphabetic order.

You can *access publications by filename* by pressing the *filenames* button. This brings up a list of entries, sorted by original filename.

Internet

# Esercitazioni

## ◆ Parte IX – Una collezione multimediale

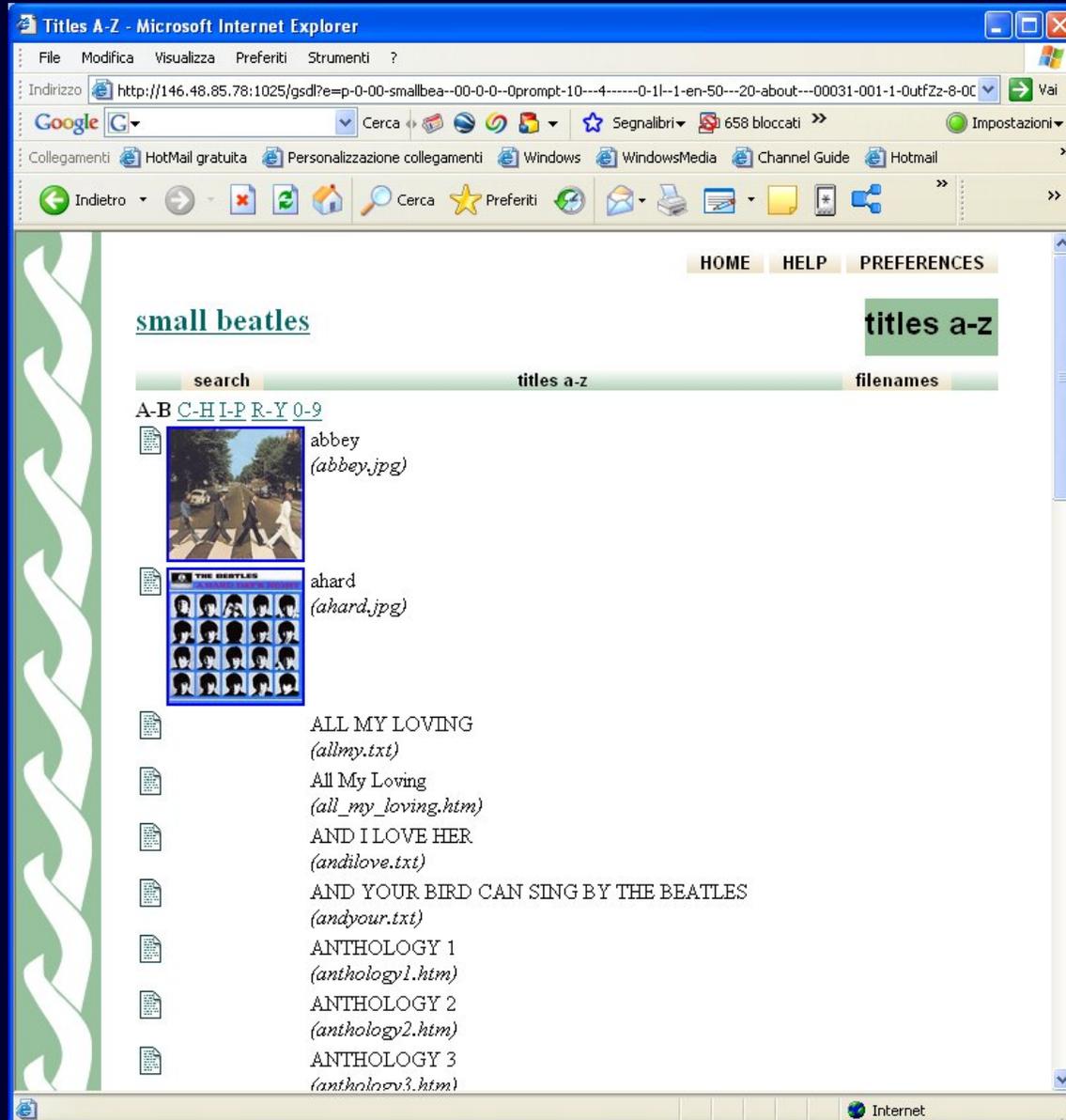
- Copieremo una collezione multimediale con informazioni sui dischi dei Beatles, brani musicali, ecc.
- La collezione è già stata creata, quindi è sufficiente copiare la cartella con i dati della collezione in “collect”
- Si provi ad esplorare la collezione appena creata.
- Nella prossima esercitazione, proveremo a creare una collezione identica a questa



# Esercitazioni

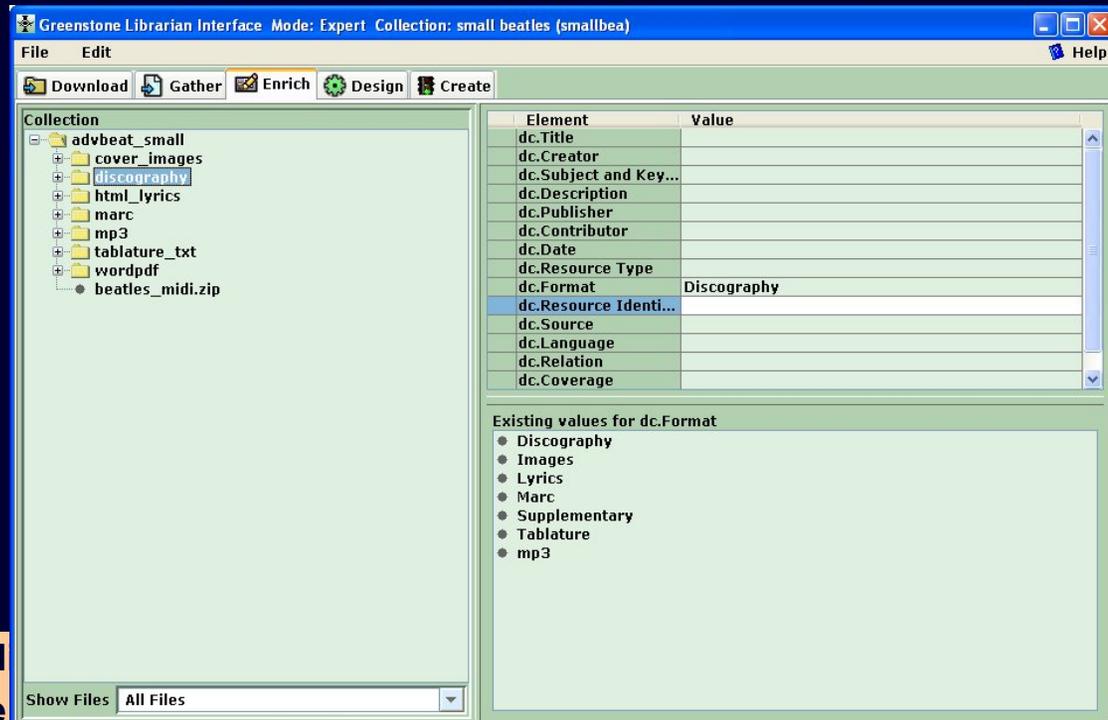
## ◆ Parte X – Creazione di una collezione multimediale

- Ora ricostruiremo la collezione sui Betales.
- Useremo un sottoinsieme dei file, per ridurre i tempi di costruzione della collezione (cartella “advbeat\_small”)
- Si prova a creare la collezione con i dati iniziali.



## ◆ Parte X (cont) – Creazione di una collezione multimediale

- Correggeremo manualmente alcuni metadati (ad es. dc.Title per i file magicalmysterytour.htm)
  - ➔ Il nuovo valore viene inserito nel campo dc.Title
  - ➔ Questo implica che dovremo creare un classificatore unico per dc.Title ed ex.Title
- Vogliamo ora permettere il browsing per ogni diverso tipo di media (per la discografia, per gli audio, ecc.)
  - ➔ Assegniamo un valore al metadato dc.Format per ogni tipo di oggetto (tipi di oggetti simili si trovano nella stessa cartella)



# ◆ Parte X (cont) – Creazione di una collezione multimediale

- ➔ Rimuoviamo il Browsing classifier per ex.Source
- ➔ Creiamo un AZCompactList classifier per dc.Format

The image shows two overlapping windows. The foreground window is titled 'Configuring Arguments' and contains the following configuration for 'AZCompactList':

- metadata: dc.Format
- firstvalueonly
- allvalues
- sort
- removeprefix
- removesuffix
- mingroup
- minnesting
- mincompact
- maxcompact
- doclevel: top - Whole document.
- freqsort
- recopt

The foreground window also shows configuration for 'BasClas':

- buttonname: browse
- no\_metadata\_formatting
- builddir
- outhandle: STDERR

The background window is a Microsoft Internet Explorer browser displaying a web page titled 'small beatles'. The page has a search bar and a list of categories: Discography, Images, Lyrics, Marc, mp3, Supplementary, and Tablature. The browser's address bar shows a URL starting with 'http://146.48.85.78:1025/gsd/?e=d-0-00-smallbea--00-0-0--0prompt-10---4-----0-11--1-en-50---20-about---00031-001-1-OutfZz'.

## ◆ Parte X (cont) – Creazione di una collezione multimediale

- Eliminiamo la visualizzazione di testo non necessario (per tutti i file multimediali sono stati creati dei documenti html vuoti).
  - ➔ La visualizzazione viene modificata in format features della VList

### Format originale

```
<td valign="top">[link][icon][/link]</td>
<td
valign="top">[ex.srclink]{Or}{[ex.thumbicon],[ex.srcicon]}[ex./srclink]</td>
<td valign="top">[highlight]
{Or}{[dls.Title],[dc.Title],[ex.Title],Untitled}
[/highlight]{If}{[ex.Source],<br><i>([ex.Source])</i>}</td>
```

### Format modificato

```
<td valign=top>
{If}{[dc.Format] eq 'Audio',
    [srclink][srcicon][/srclink],
    {If}{[dc.Format] eq 'Images',
        [srclink][thumbicon][/srclink],
        [link][icon][/link]}}</td>
<td valign=top>[highlight]
{Or}{[dls.Title],[dc.Title],[Title],Untitled}
[/highlight]{If}{[ex.Source],<br><i>([ex.Source])</i>}</td>
```

## ◆ Parte X (cont) – Creazione di una collezione multimediale

- Eliminiamo anche la visualizzazione del nome del file sorgente modificando ancora la riga della format feature

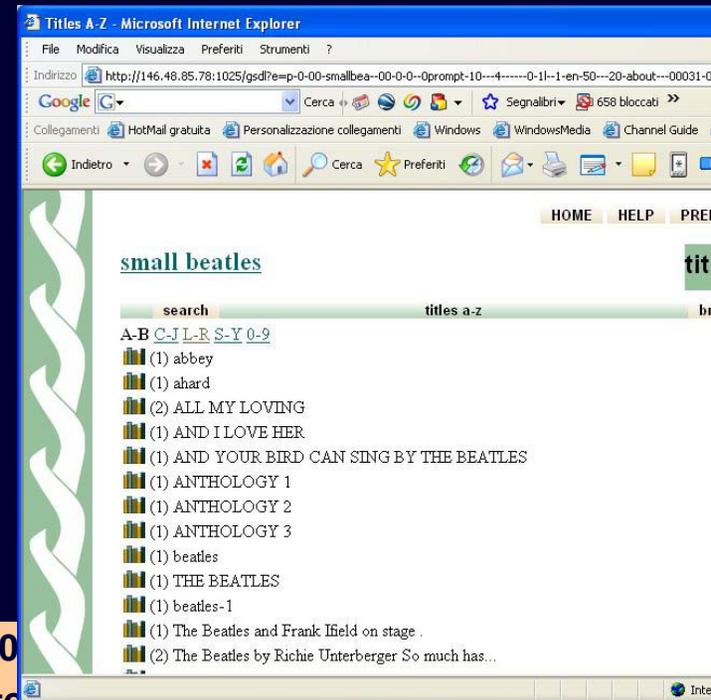
```
<td valign=top>
{If}{[dc.Format] eq 'Audio',
  [srclink][srcicon][/srclink],
  {If}{[dc.Format] eq 'Images',
    [srclink][thumbicon][/srclink],
    [link][icon][/link]}}</td>
<td valign=top>[highlight]
{Or}{[dls.Title],[dc.Title],[Title],Untitled}
[/highlight]{If}{[ex.Source],<br><i>([ex.Source])</i>}</td>
```

```
<td valign=top>
{If}{[dc.Format] eq 'Audio',
  [srclink][srcicon][/srclink],
  {If}{[dc.Format] eq 'Images',
    [srclink][thumbicon][/srclink],
    [link][icon][/link]}}</td>
<td valign=top>[highlight]
{Or}{[dls.Title],[dc.Title],[Title],Untitled}
[/highlight]{If}</td>
```

## ◆ Parte X (cont) – Creazione di una collezione multimediale

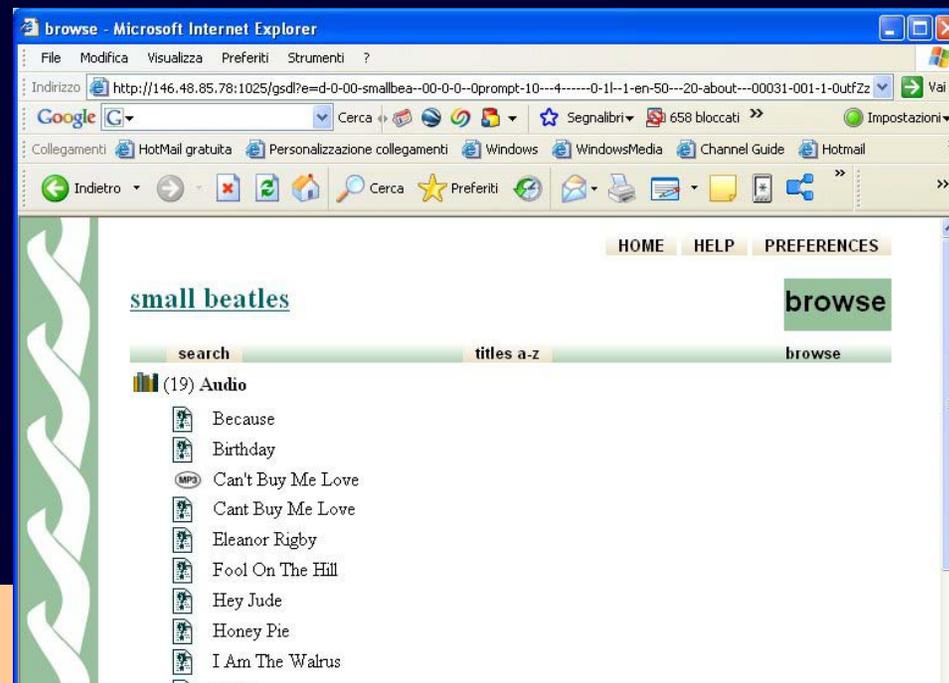
- Uso di AZCompactList al posto di AZList
  - ➔ Spesso più documenti hanno lo stesso titolo per cui vogliamo raggrupparli insieme
  - ➔ Sostituiamo il classificatore AZList per dc.Title, ex.Title con AZCompactList
  - ➔ Per ogni documento (anche se ripetuto) viene visualizzata una icona dello scaffale
  - ➔ Possiamo anche visualizzare il numero di oggetti presenti in ogni gruppo, modificando il format features della VList come segue

```
<td valign=top>
{If}{[dc.Format] eq 'Audio',
[srclink][srcicon][/srclink],
{If}{[dc.Format] eq 'Images',
[srclink][thumbicon][/srclink],
[link][icon][/link]}}</td>
<td>{If}{[numleafdocs],([numleafdocs])}</td>
<td valign=top>[highlight]
{Or}{[dls.Title],[dc.Title],[Title],Untitled}
[/highlight]</td>
```



# Parte X (cont) – Creazione di una collezione multimediale

- Aggiunta di un browser di frasi (Phind)
- Modifica dell'icona della collezione
- Gestione dei file MIDI
  - ➔ Verrà usato il plugin UnknownPlug, un plugin generico che può operare su qualunque tipo di file
  - ➔ Configurare il plugin in modo da gestire i file con estensione .mid e da assegnare a questi file un formato MIDI ed un mime type audio/midi
  - ➔ Per visualizzare questi file, bisogna anche assegnare il metadato dc.Format=Audio al file beatles\_midi.zip



## Parte X (cont) – Creazione di una collezione multimediale

- Ora “ripuliremo” la lista dei titoli
  - Ad es. se abbiamo “Anthology 1”, “Anthology 2” vogliamo che questi titoli siano raggruppati
  - A questo scopo useremo una espressione regolare per specificare quali sono i suffissi del valore del metadato ex.Title che devono essere rimossi
  - `(?i)(\\s+\\d+)(\\s+[[:punct:]].*)`
  - Questa stringa si specifica nel campo removesuffix del classificatore per ex.Title
  - Modificheremo poi le modalità di presentazione (sfondo, immagini, ecc.) cambiando alcuni file di macro
  - Infine modificheremo ancora la format feature della VList per visualizzare icone diverse per i diversi tipi di media