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**

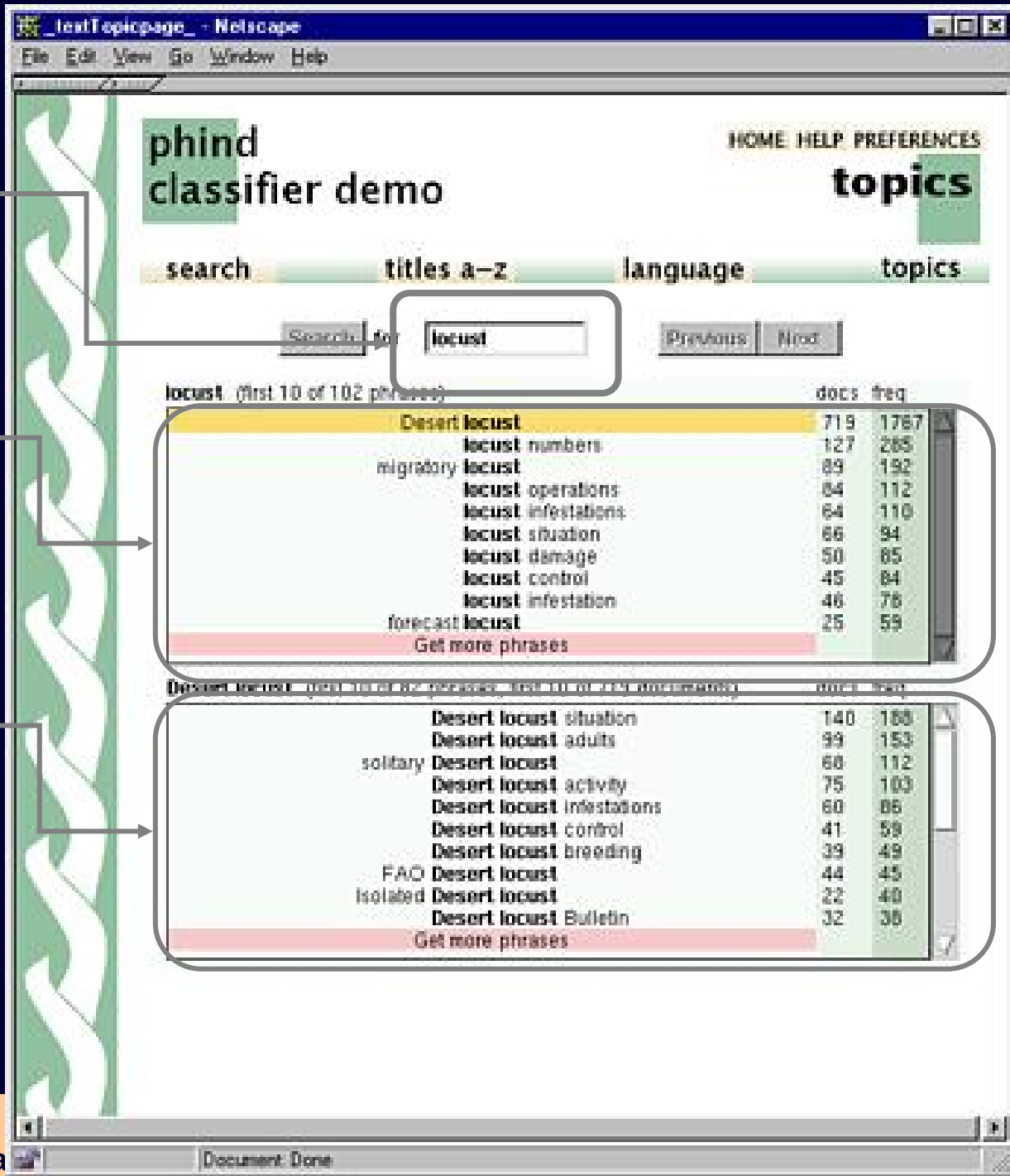
Browsing su frasi

- **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 gerachicamente
  - 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**

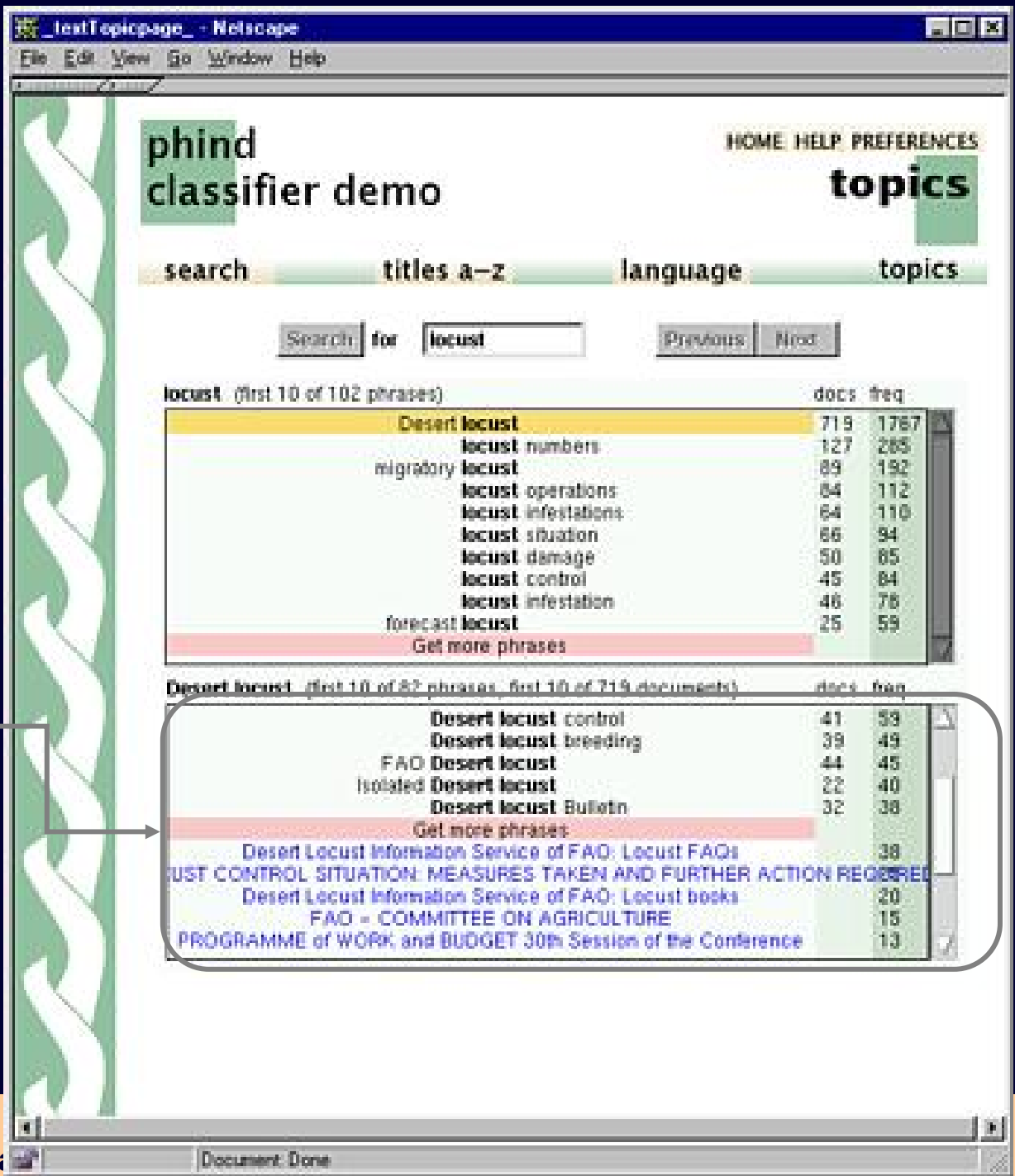
Search frasi che contengono la parola “locust”

Vengono visualizzate le frasi con informazioni sul numero di documenti

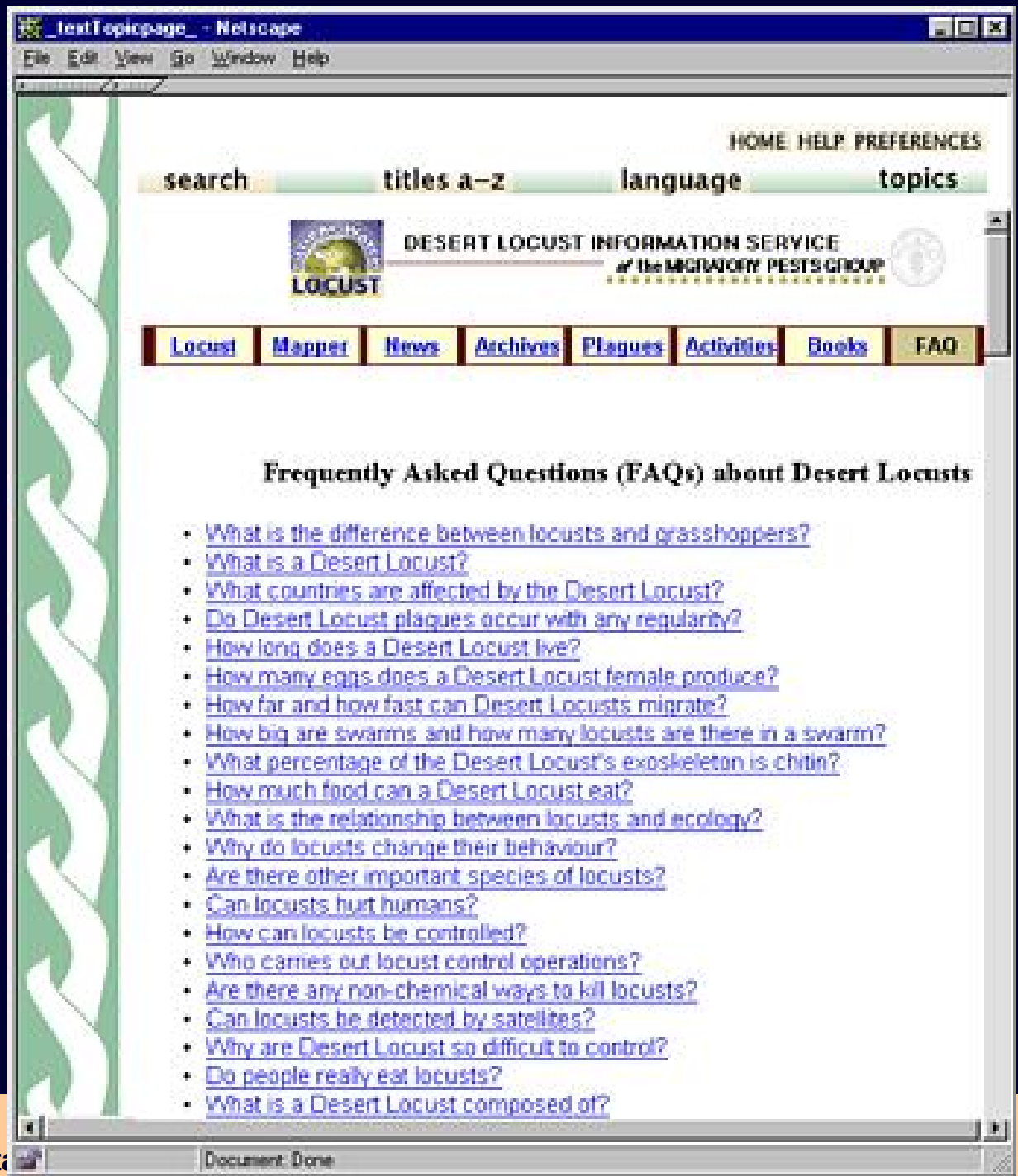
Seleziono una delle frasi (“desert locust”) e trovo tutte le frasi che la contengono



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

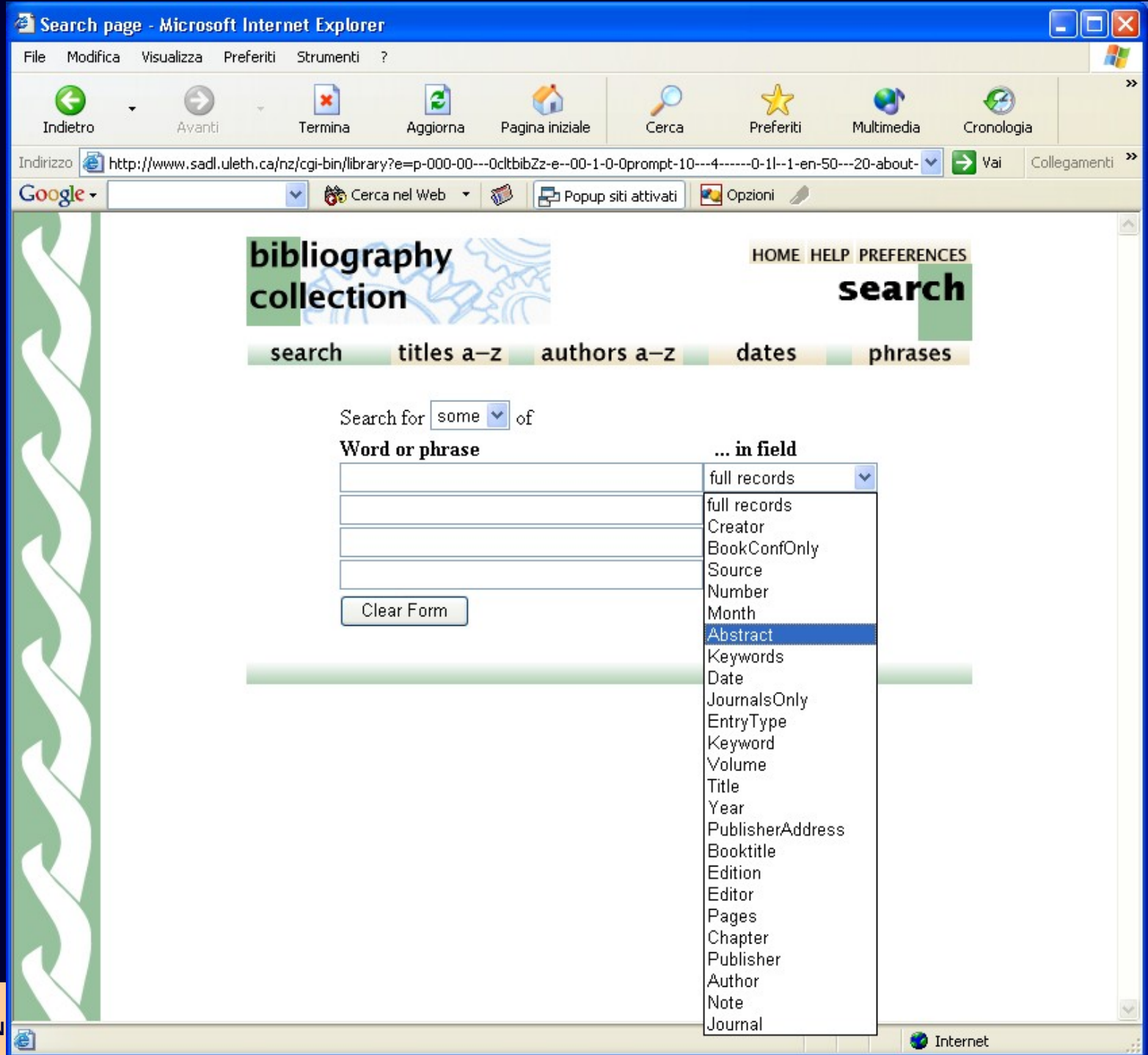


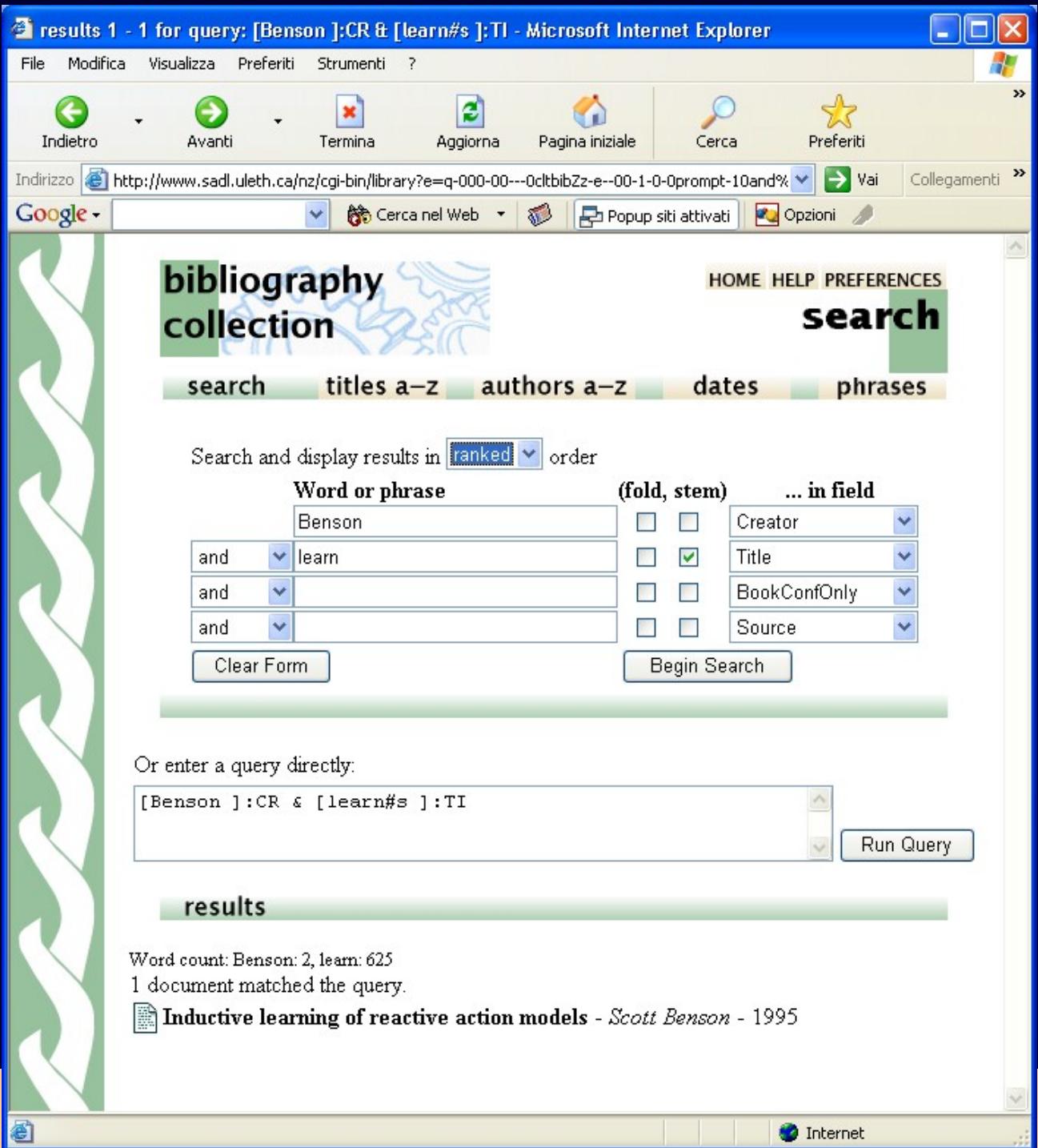
Quindi posso visualizzare il documento



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**


Search page - Microsoft Internet Explorer
File Modifica Visualizza Preferiti Strumenti ?
Indietro Avanti Termina Aggiorna Pagina iniziale Cerca Preferiti Multimedia Cronologia
Indirizzo http://www.sadl.uleth.ca/nz/cgi-bin/library?e=p-000-00---0cltbibZz-e--00-1-0-0prompt-10---4------0-1l--1-en-50---20-about-
Vai Collegamenti
Google Cerca nel Web Popup siti attivati Opzioni
bibliography collection
HOME HELP PREFERENCES
search
search titles a–z authors a–z dates phrases
Search for some of
Word or phrase
... 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
Clear Form
Internet


results 1 - 1 for query: [Benson]:CR & [learn#s]:TI - Microsoft Internet Explorer
File Modifica Visualizza Preferiti Strumenti ?
Indietro Avanti Termina Aggiorna Pagina iniziale Cerca Preferiti
Indirizzo http://www.sadl.uleth.ca/nz/cgi-bin/library?e=q-000-00---0cltbibZz-e--00-1-0-0prompt-10and%
Vai Collegamenti
Google Cerca nel Web Popup siti attivati Opzioni
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 Creator
and learn Title
and BookConfOnly
and Source
Clear Form
Begin Search
Or enter a query directly:
[Benson]:CR & [learn#s]:TI
Run Query
results
Word count: Benson: 2, learn: 625
1 document matched the query.
Inductive learning of reactive action models - Scott Benson - 1995
Internet

Form-based searching

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

```
Searchtype form
Indexes nome-indici
```

Esercitazioni

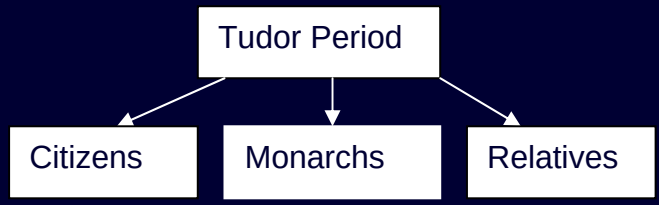
- **Parte VI – Utilizzo dei classificatori**
  - Creazione del classificatore PHIND per la ricerca di frasi
  - Creazione di un indice per la ricerca di frasi
    - ➔ **l'indice che creeremo ora permette di specificare esplicitamente le frasi da cercare**
  - Creazione di partizioni (sottocollezioni) 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

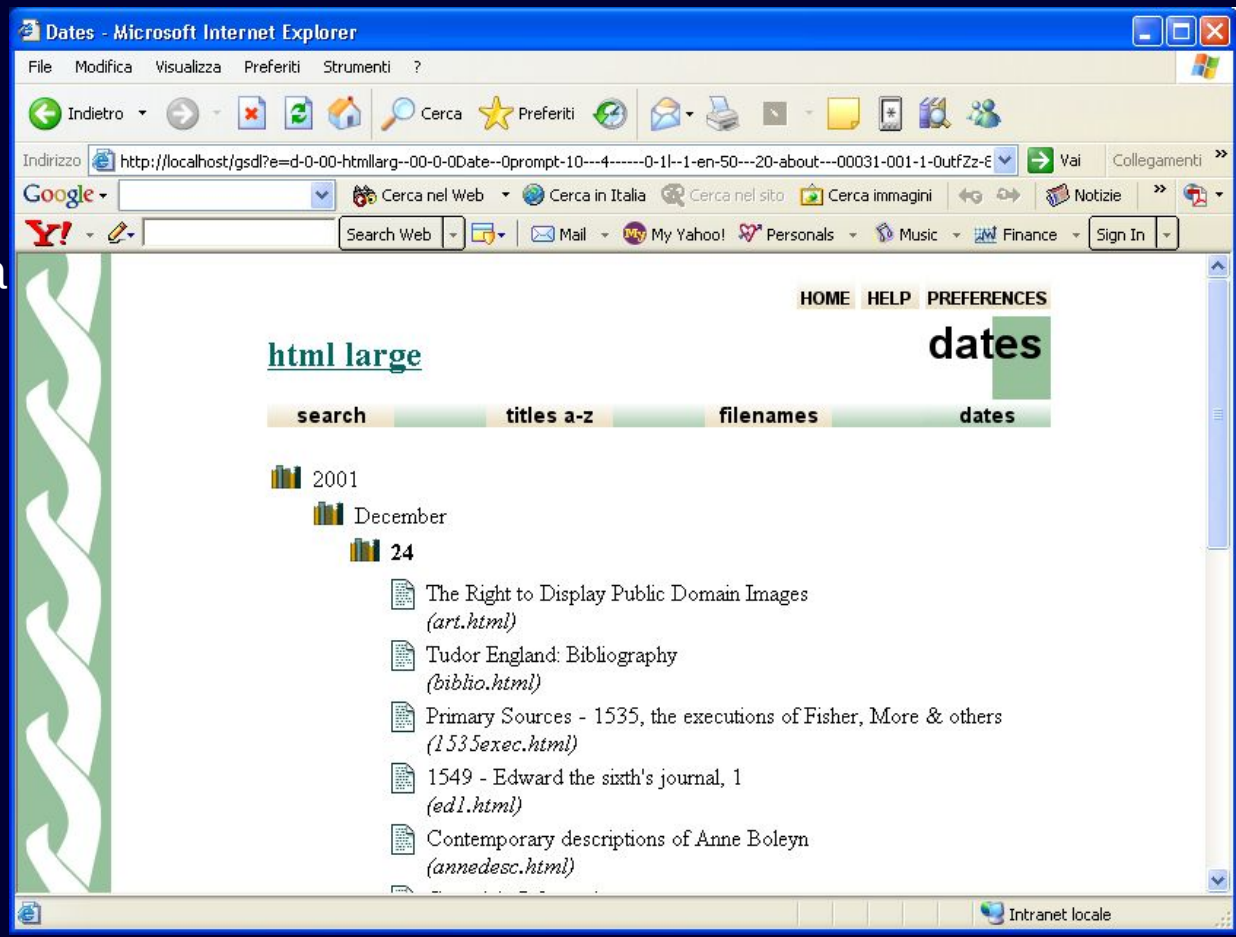
- **Parte VI (cont.)**
  - **Creare una gerarchia di soggetti (dc.Subject)**



  - **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 (Hierachy 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?

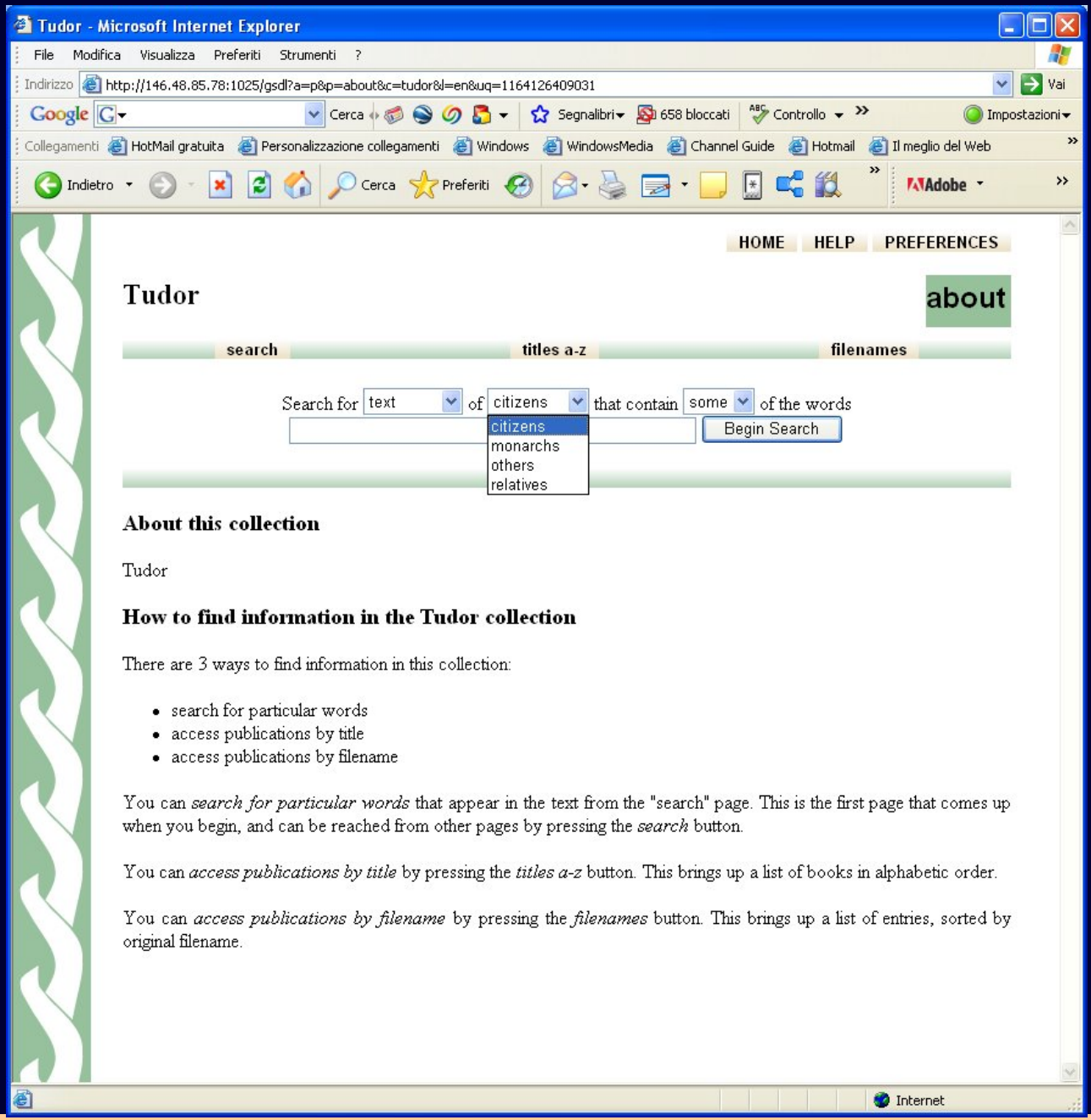


Esercitazioni

- **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/gsdl?a=p&p=about&c=tudor&l=en&uq=1164126409031
Vai
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 text of citizens that contain some 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

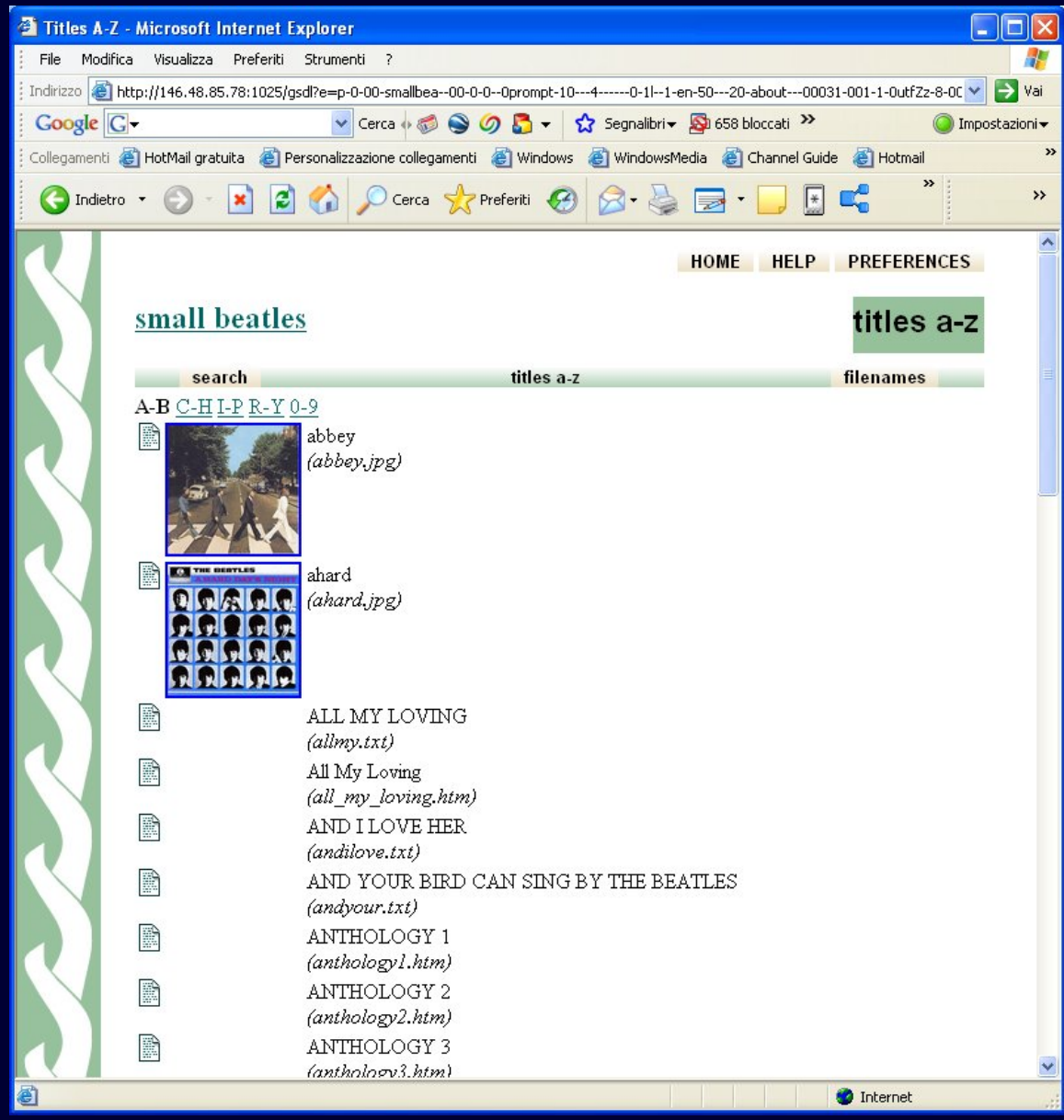
Esercitazioni

- **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


Browse - Microsoft Internet Explorer
File Modifica Visualizza Preferiti Strumenti ?
Indirizzo http://146.48.85.78:1025/gsdl?e=d-0-00-advbeat%5flarge--00-0-0--0prompt-10---4------0-1l--1-en-50---20-about---00031-001-1-0u
Vai
Google Cerca Segnalibri 658 bloccati Impostazioni
Collegamenti HotMail gratuita Personalizzazione collegamenti Windows WindowsMedia Channel Guide Hotmail
Indietro Cerca Preferiti
THE BEATLES
HOME HELP PREFERENCES
browse
search titles a-z browse phrases
(21) Audio
(50) Discography
(14) Images
(264) Lyrics
(50) MARC
(7) Supplementary
(31) Tablature
Internet

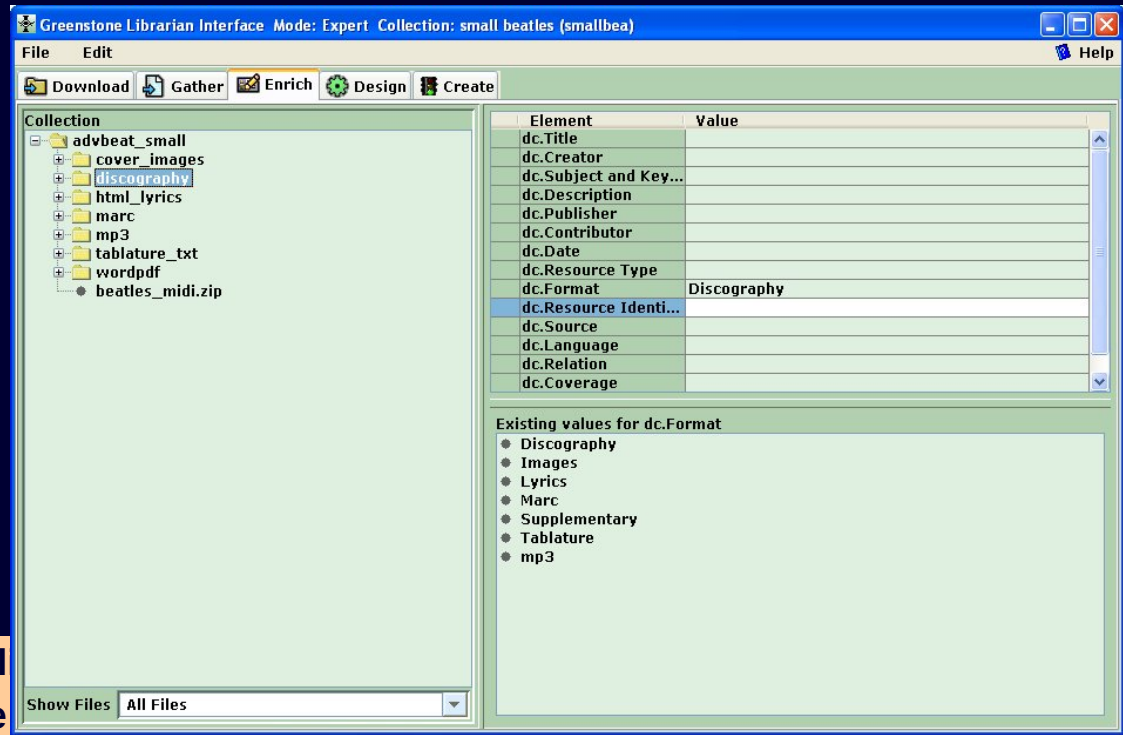
Esercitazioni

- **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.



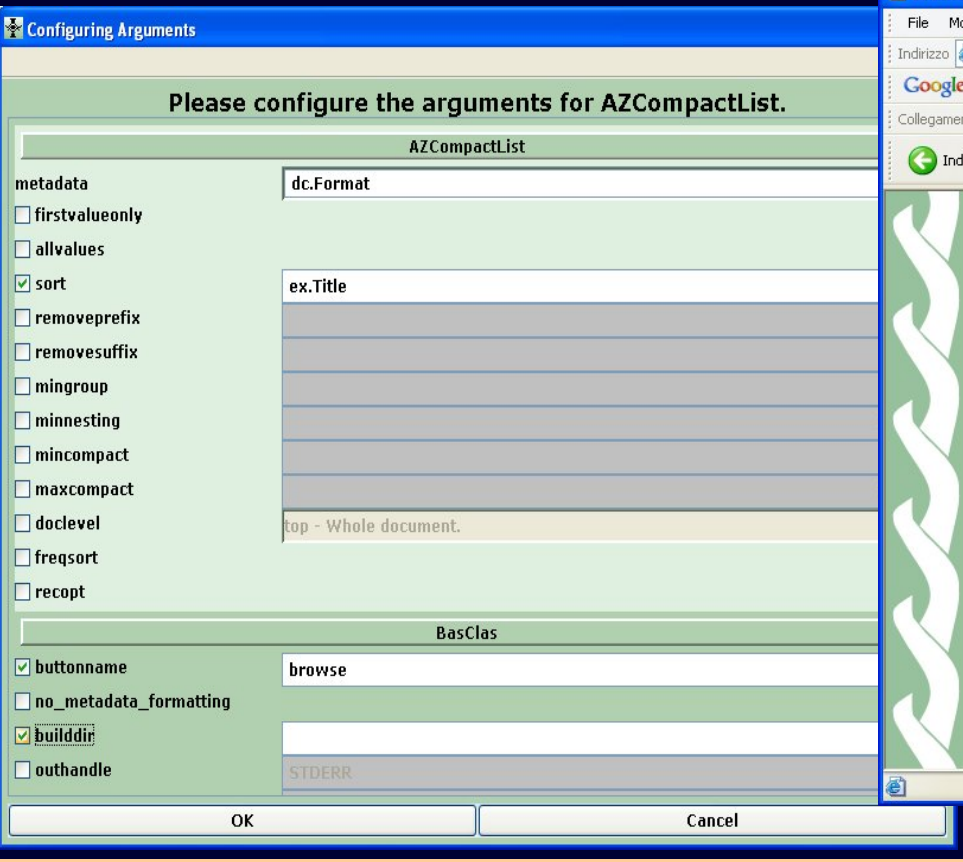
Parte X (cont) – Creazione di una collezione multimediale

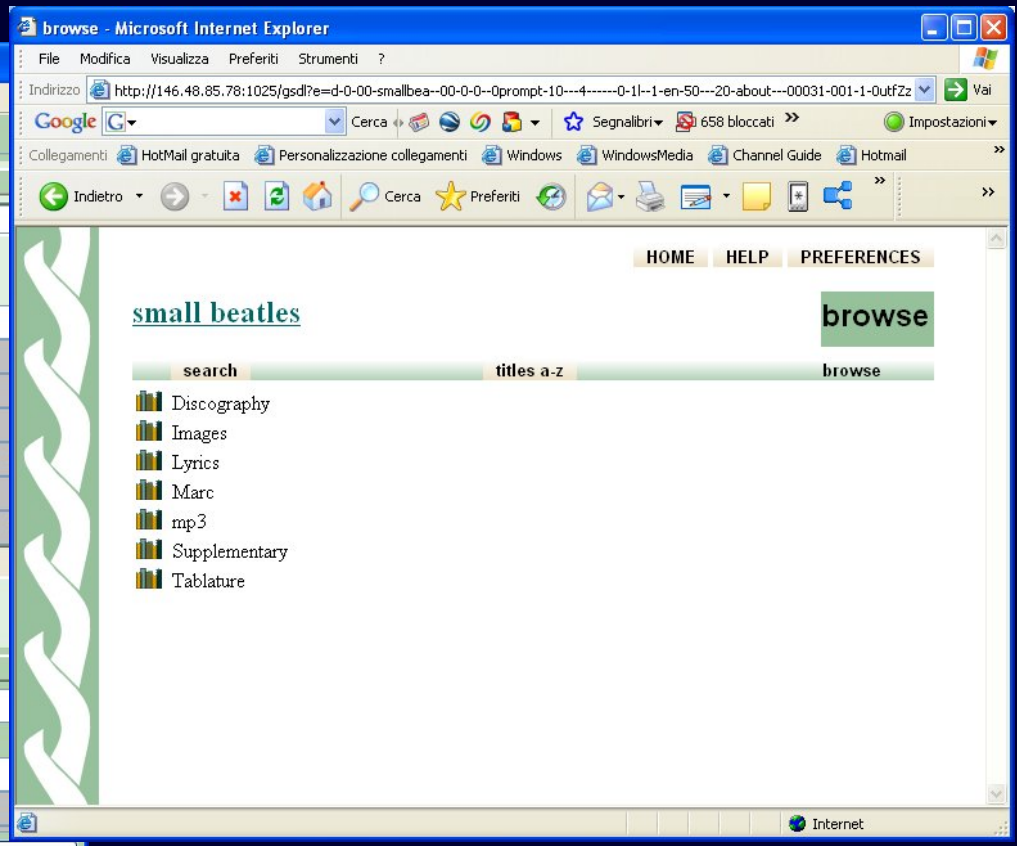
- Correggeremo manualmente alcuni metadati (ad es. dc.Title per i file magicalmisterytour.htm)
  - ➔ **Il nuovo valore viene inserito nel campo dc.Title**
  - ➔ **Questo implica che dovremo creare un classificatore unico per dc.Title ed ex.Title**
- Vogliamo ora permettere il browsing per ogni diverso 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





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./srcl
ink]</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

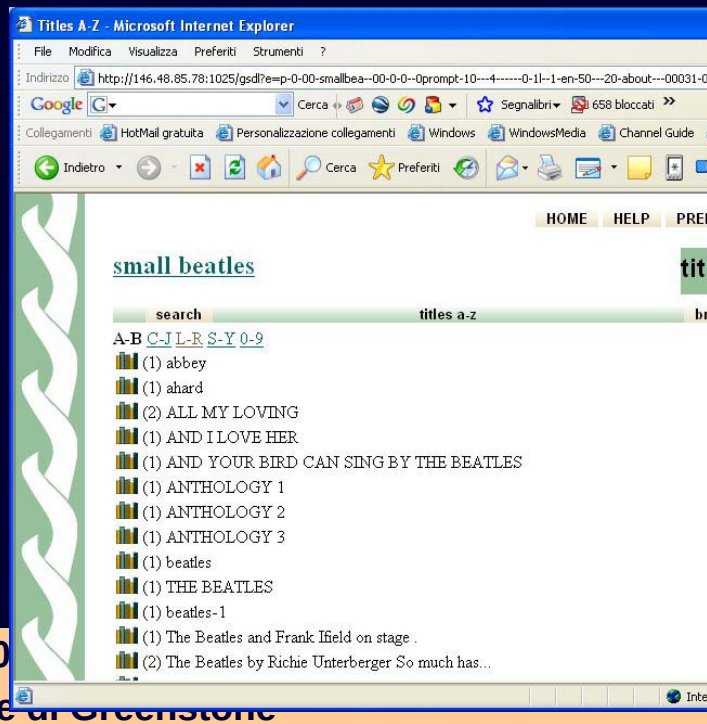
```
<td valign=top>
{If}{[dc.Format] eq 'Audio',
     [srclink][srcicon][/srclink],
     {If}{[dc.Format] eq 'Images',
          [srclink][thumbicon][/srclink],
          [link][icon][/link]}}</td>
<td valign=top>[highlight]
{Or}{[dls.Title],[dc.Title],[Title],Untitled}
[/highlight]{If}{[ex.Source],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

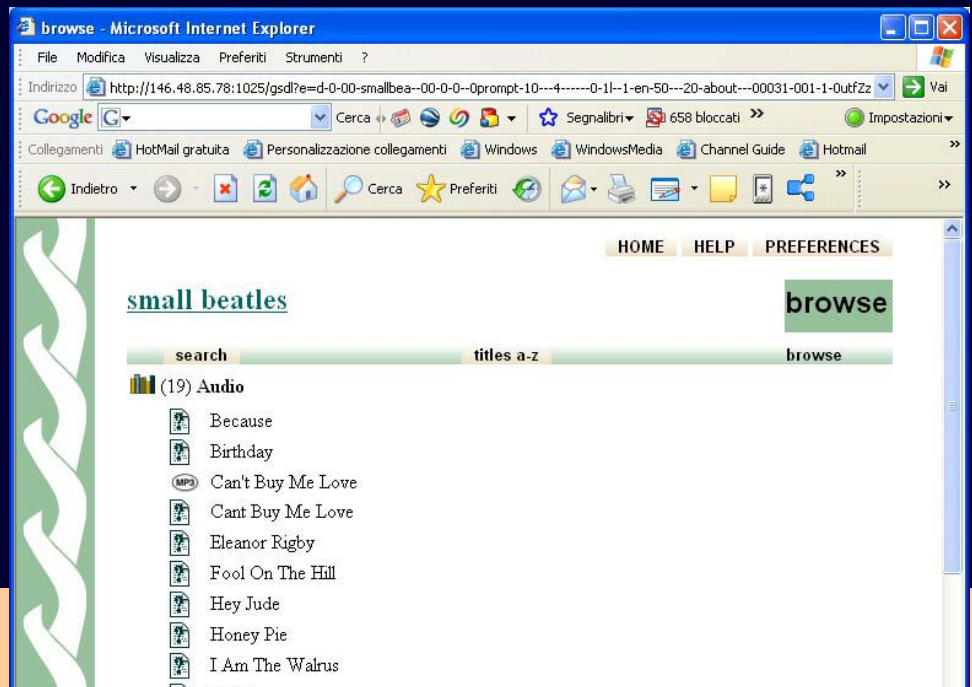
- Uso di AZCompactList al posto di AZList
  - ➔ **Spesso più documenti hanno lo stesso titolo per cui vogliamo raggrupparli insieme**
  - ➔ **Sostituiamo il classificatore AZList per dc.Title, ex.Title con AZCompactList**
  - ➔ **Per ogni documento (anche se ripetuto) viene visualizzata una icona dello 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**