

Tipologia dei dati e organizzazione delle informazioni Sistemi di indicizzazione e recupero



6

Indicizzazione manuale e automatica Sistemi di ricerca

Ricerca

- **Per valore esatto** —————→ **DBMS**
si cercano i "record" i cui "campi" soddisfano un certo valore
- **Per contenuto semantico** —————→ **IRS**
si cercano documenti che contengano parole o frasi di interesse per l'utente

In tutti i casi l'informazione viene reperita per mezzo di **INDICI**

che descrivono

- le entità (caso db)
- i documenti (caso information retrieval)
- le pagine Web (caso ricerca in rete)

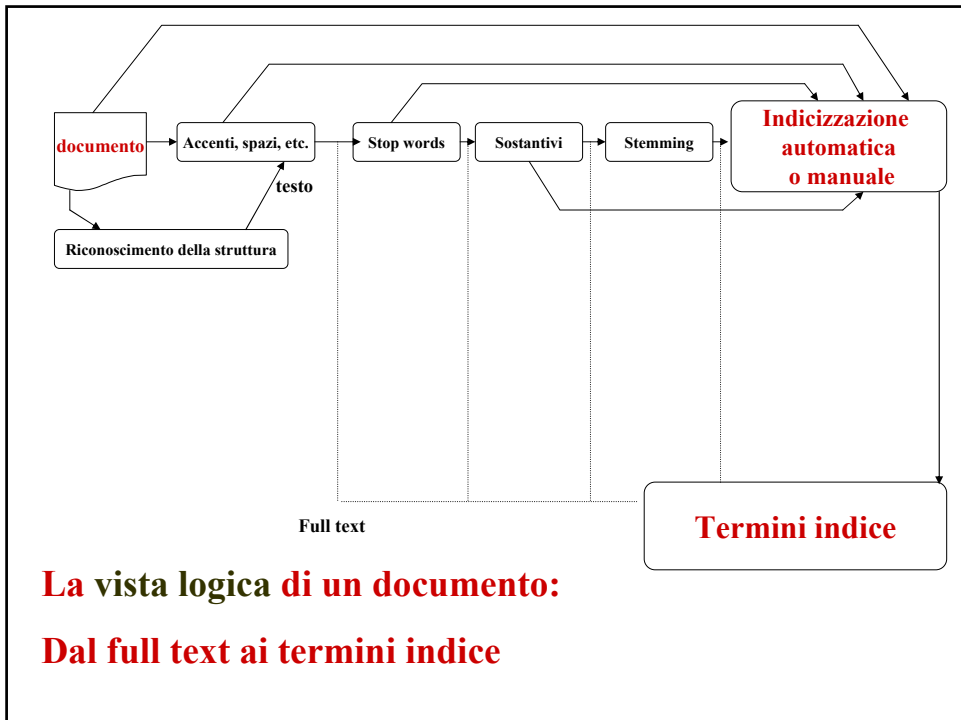
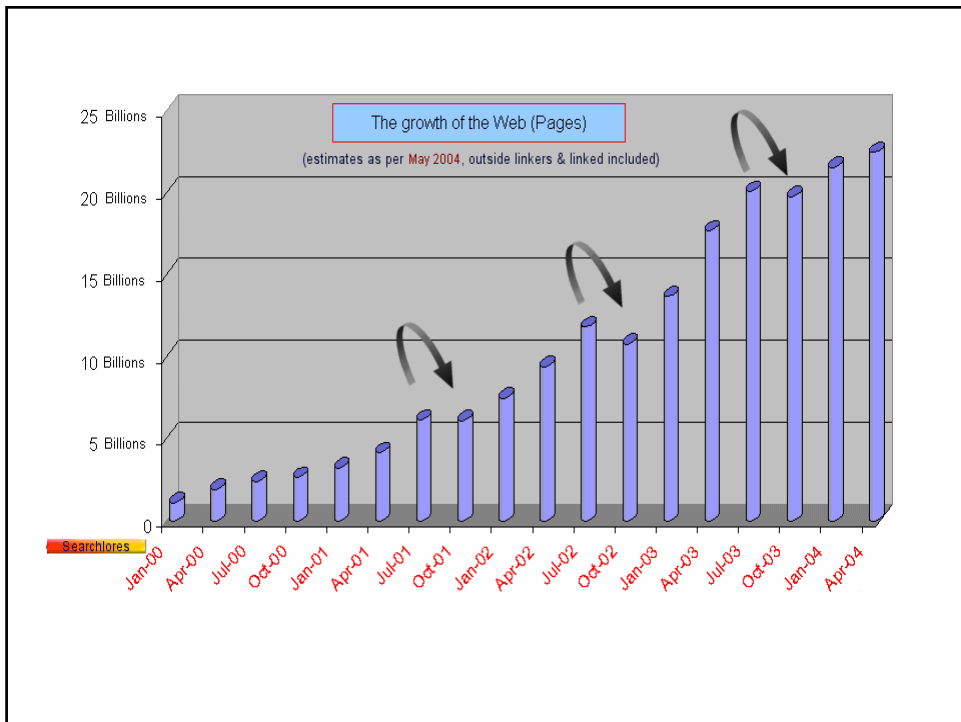
Gli **INDICI** sono **metadati** cioè dati sui dati

I “dati sui dati”

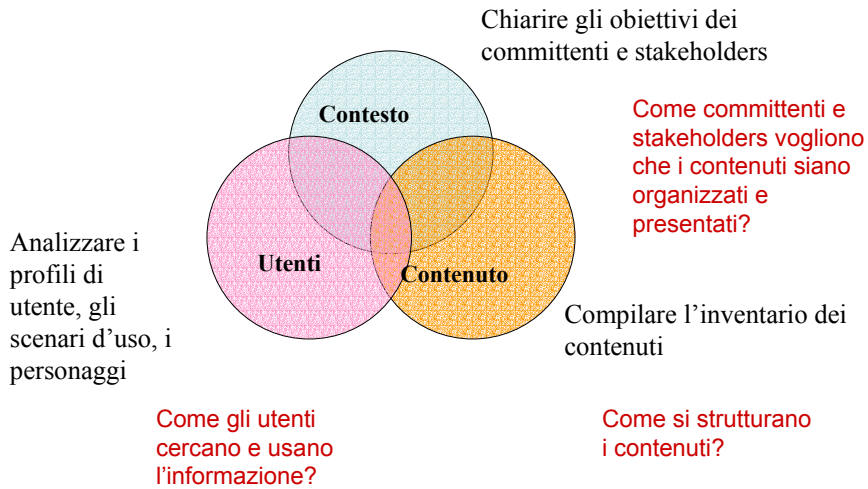
Vengono creati sia manualmente che automaticamente.

Sono indispensabili per far trovare documenti Web e quindi farli leggere.

Per grandi volumi sono creati automaticamente




Metadati e Information Architecture



Tipi e funzioni di metadati

TIPO	DEFINIZIONE	ESEMPI
Amministrativi	Metadati usati per l'amministrazione e gestione	Fonte Proprietà e diritti Documenti per l'accesso legale Localizzazione delle fonti Controllo delle versioni o release
Descrittivi	Metadati usati per descrivere e identificare le informazioni	Record di catalogazione Indici Relazioni di Iperlink Annotazioni
Conservativi	Metadati relativi alle pratiche per la conservazione	Documentazione sulle misure adottate per preservare le versioni fisiche o digitali dei dati (es la data dell'ultima copia, eventuale migrazione su altri sistemi)
Tecnici	Relativi alle modalità tecniche di creazione dei dati e alle procedure adottate per la sicurezza	Digitalizzazione (e.g., formati, rapporti di compressione, routines di scalatura etc) Data di autenticazione e sicurezza (Es., encryptions, passwords)
Uso	Livello e tipo di uso della sorgente di informazioni	Permettono di memorizzare il tipo di uso e di utente (Es. un codice che rende visibili o no i dati per una certa categoria di utenti)



Tessuto n. 86

Descrizione: **Frammento. Seta; damasco;**
37.5 x 31.5cm
Italia(?), XVII secolo
damasco

Den. tecnica: 9.4 x 13.6cm

Rapp. disegno: organzino di seta tinto in verde
trama seta tinta in giallo binata a 2 capi

Ordito:

Tessimento:

Cimosse:

Armature: fondo legato in raso da 5 faccia catena
disegno (opera) legato in raso da 5 faccia
trama

Il frammento è molto sfrangiato ai bordi; il tessuto si presenta molto morbido, duttile e piatto. Per la fortuna e gli sviluppi della tecnica si rimanda al saggio di Jolanda Silvestri (...). In questo frammento il contrasto tra sfondo e disegno è sottolineato anche dalla differenza del colore; giallo per la trama, verde per l'ordito.

Lo schema di base a maglie ovali a doppia punta, disposte in righe sfalsate, è quasi annullato dal fitto intrico di rami fioriti e foglie abitate da varie specie di uccelli. Il disegno è riferibile al repertorio decorativo che illustra specie diverse del mondo animale e vegetale usato anche per i ricami nel corso del secolo XVI che si mantiene fino al secolo XVIII soprattutto nei damaschi di lino (...).

Il rapporto piuttosto ridotto del disegno e l'effetto minuto delle decorazioni fanno pensare al suo uso per un abito. Dalla Podreider (...) era stato datato alla seconda metà del secolo XIV e classificato come diaspro lucchese. Due tessuti simili della raccolta di Isabella Errera sono invece datati tra il XVI e il XVII secolo e più correttamente indicati come damaschi. Damaschi a disegni piccoli di questo tipo si trovano, con attribuzione all'Italia,

Metadati

INDICI

TABELLA: SCHEDA
NSCHEDA: 86
MODULO: cm. 9.4 x 13.6

DATA: Secolo 17|inizio, secolo 17|meta', secolo 17|fine

LUOGO: Italia

FASCEORIZ: 0
FASCEVERT: 0
FIGURATIVO: 0
FILEVERT: 0
FILEDIAG: 0
FILEORIZ: 0
M_OVALI: 2
M_ESAGONAL: 0
M_ROMBOIDA: 0
M_MISTILINEA: 0
SCACCHIERA: 0
SERPENTINA: 0

BICROMIA: BICROMIA|Verde|fondo
BICROMIA|giallo|fondo
BICROMIA|Giallo|pelo
BICROMIA|verde|pelo

VEGETALI: VEGETALI|RAMO|Ramo naturalistico|fiorito
VEGETALI|FOGLIA
ANIMALI: ANIMALI|Animali reali|uccello

TIPOLOGIA: DAMASCO

MATERIE: SETA|seta

BLOCCO1: frammento ù cm. 37,5 x 31,5
BLOCCO2: damasco Italia ù XVII sec.
BLOCCO3: inv. n. 2667
BLOCCO4: damasco (tessuto "senza rovescio") seta

DIDA: 86
frammento
37,5 x 31,5 cm.
damasco

I metadati.

Loro ruolo nell'architettura dell'informazione

Per migliorare la navigazione e il recupero dei dati da parte dell'utente, gli autori di pagine web hanno la possibilità di aggiungere parole o frasi che ne descrivono il contenuto attraverso i cosiddetti metadati.

Esempio di metadati aggiunti ad un portale sul cavallo e l'equitazione sotto forma di parole chiave nel linguaggio HTML:

```
<title>EQUINET - Il portale italiano del cavallo e dell'equitazione</title>
.....
<meta name="keywords" content="equitazione, cavallo, horse, cheval,
equitation, endurance, salto ostacoli, monta western, turismo equestre,
vacanze a cavallo, dressage, ippica, centro ippico, reining, purosangue arabo,
mascalcia, monta maremmana, veterinaria">
```

I metadati non compaiono nell'interfaccia utente, ma sono disponibili ai motori di ricerca.

Vocabolari controllati

Nella sua forma più semplice un vocabolario controllato è un sottoinsieme di un linguaggio che rappresenta un sapere specialistico, per esempio un elenco (indice) dei termini specifici di una disciplina (arte, medicina, economia, ecc.)

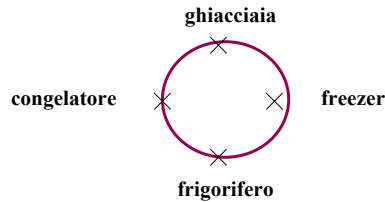
Un vocabolario controllato di questo tipo può essere:

- deciso da uno o più esperti, o
- costruito automaticamente scartando dai testi del settore le parole cosiddette “non-stop” (articoli, preposizioni, pronomi, ecc.)

Vocabolari controllati: anelli di sinonimi

Un primo arricchimento del vocabolario controllato è costituito dalla introduzione dei sinonimi, o meglio di termini considerati equivalenti secondo certi criteri, nella stessa lingua o in lingue diverse, comprendendo anche errori ortografici comuni.

Poiché nessuno dei termini equivalenti è considerato preferito, si parla di *anelli di sinonimi*.



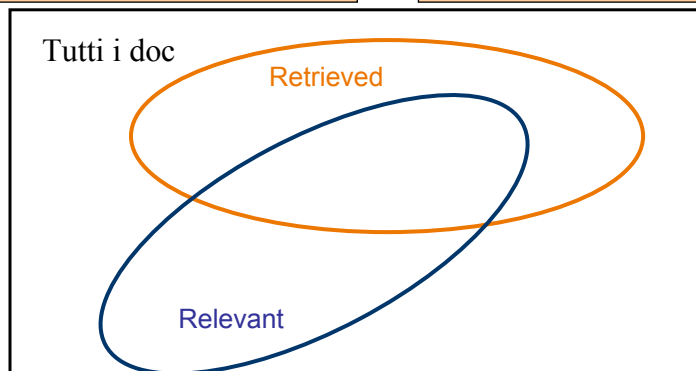
Pro e contro: maggiore quantità di risultati (**richiamo o recall**), minore rilevanza (**precisione o precision**).

Precision e Recall

Sono due criteri standard di valutazione dei sistemi di Information Retrieval

$$\text{Precision} = \frac{|\text{RelRetrieved}|}{|\text{Retrieved}|}$$

$$\text{Recall} = \frac{|\text{RelRetrieved}|}{|\text{Rel in Collection}|}$$



Algoritmi di ricerca

- Algoritmi a schema comparativo, confrontano la query con un indice o (più raro) con l'intero testo cercando la stessa stringa (*pattern matching*)
- Algoritmi che recuperano i documenti indicizzati con metadati simili (*similar pages*)

Query builder

- correttori ortografici
- analisi fonetica
- stemmer
- elaborazione linguaggio naturale
- vocabolari controllati e thesauri

Vocabolari controllati: termini preferiti

Viene chiamato “**authority file**” un elenco di termini preferiti scelti da qualche fonte autorevole per un certo settore.

I **termini preferiti** possono svolgere più di una funzione:

- per gli autori e indicizzatori, da guida terminologica
- per la gestione di dizionari controllati, da identificatori unici per gli “anelli di sinonimi”
- per l'utente, da suggerimento per l'uso di termini corretti e standard nella ricerca, e da “sfolgimento” terminologico nella navigazione.

Vocabolari controllati: schemi di classificazione

Un vocabolario controllato diventa uno schema di classificazione, (schema organizzativo) o **tassonomia**, quando i termini vengono organizzati in una gerarchia.

Uno schema di classificazione svolge un triplice ruolo:

- per l'architetto dell'informazione, come **strumento di organizzazione** e etichettatura dei documenti
- per l'utente, come **ausilio alla navigazione** (se, come in Yahoo!, è resa visibile come parte integrante dell'interfaccia)

`home>science>computer science>artificial-intelligence`

- per l'utente, **nella ricerca**, quando gli vengono mostrate le categorie in cui è stato trovato il termine dell'interrogazione

`shopping>animali>cani`

familiarizzandolo con lo schema di classificazione del sistema

Tassonomia o Schema di classificazione

Gli elementi di un gruppo sono separati in sottogruppi mutuamente esclusivi, non ambigui, che presi nel loro insieme, coprono tutte le possibilità

Una tassonomia dovrebbe essere semplice, facile da ricordare e facile da usare.

Uno dei più noti esempi di tassonomia è quella di Linneo per la biologia.

Ontologia

tassonomia concettuale, non linguistica

cose, eventi, relazioni in un particolare dominio

La Classificazione Decimale Dewey (DDC)

data di pubblicazione: 1876

- E' un elenco gerarchico con 10 categorie di livello superiore e una profondità variabile a seconda delle categorie.
- E' usata nelle biblioteche di più di 130 Paesi.
- E' presente in molte interfacce di visualizzazione.

http://www.lib.duke.edu/libguide/fi_books_dd.htm (Class. Dewey)



Dewey numbers divide humanity's knowledge, ideas, and artistic creations into **ten major categories** spanning a range from **000 to 999**:

000 Generalities

100 Philosophy & psychology

200 Religion

300 Social sciences

400 Language

500 Natural sciences & math



600 Technology (Applied sciences)

700 The arts

800 Literature & rhetoric

900 Geography & history

2

Each major category divides into **nine sub-categories** spanning a range of 10 to 90.
For example:

500 Natural science & mathematics

510 Mathematics

 520 Astronomy & allied sciences

530 Physics

540 Chemistry & allied sciences

550 Earth sciences

560 Paleontology & paleozoology

570 Life sciences

580 Botanical sciences

590 Zoological sciences

3

Each sub-category is further divided into **nine specialized topics** ranging from 1 to 9:

520 Astronomy

521 Celestial mechanics

522 Techniques, equipment, etc.

523 Specific celestial bodies 

524 [Unassigned]

525 Earth (Astronomical geography)

526 Mathematical geography

527 Celestial navigation

528 Ephemerides

529 Chronology

4

By adding **decimals**, the specialized topics are broken down even further:

523.3 Moon

523.4 Planets

523.5 Meteors, solar wind, zodiacal light

523.6 Comets

523.7 Sun

523.71 Constants and dimensions

523.72 Physics of

523.73 Motions

523.74 Photosphere

523.75 Chromosphere and corona

523.76 Solar interior

523.78 Eclipses

Vocabolari controllati: thesauri

Un Thesaurus è un vocabolario controllato in cui vengono esplicitate relazioni semantiche fra termini. Precisamente:

- **relazioni di equivalenza** fra i termini (anelli di sinonimi)
- **relazioni gerarchiche** fra i termini preferiti (schemi di classificazione)

e inoltre

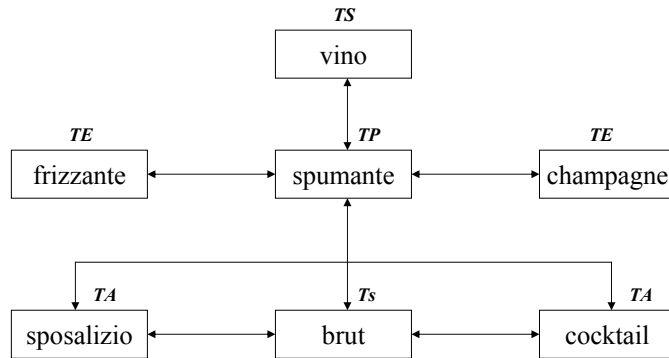
- **relazioni associative** fra i termini



Esempio

Associativa è per esempio **la relazione di contestualità** fra termini, come “forchetta” e “coltello”, “autostrada” e “casello”, “Waterloo” e “Napoleone”.

Relazioni semantiche in un thesaurus di vini



**TP=termine preferito
(variante, sinonimo)**

TE=termine equivalente

TS=termine sopraordinato (più generale)

TA=termine associato

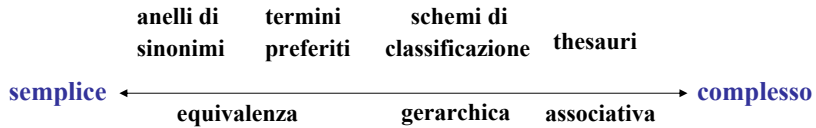
Ts=termine sottordinato (più specifico)

Gli strumenti per l'indicizzazione

- Vocabolari controllati
- Anelli di sinonimi
- Termini preferiti (Authority file)
- Tassonomie e schemi organizzativi (gerarchie tra termini di un vocabolario)
- Thesauri: Vocabolari controllati con relazioni tra termini

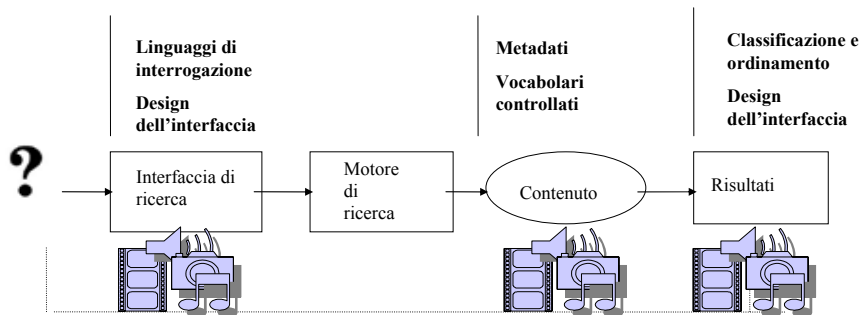
Vocabolari controllati e thesauri. Riepilogo

(vocabolario)



(relazioni)

Bisogna facilitare un certo tipo di **utente** a trovare i **contenuti** che cerca in un certo **contesto**



I media possono avere il ruolo di

- **Elementi del Linguaggio di interrogazione**
- **Contenuto informativo fatto di testi, immagini etc**
- **Risultati**

Linee guida per la costruzione/gestione di Thesauri

Ci sono vari standard nazionali e internazionali che offrono linee guida per la costruzione di thesauri, fra cui il più diffuso è lo

standard ANSI/NISO Z39.19 (USA,1994)

Buoni motivi per attenersi alle linee guida dello standard USA :

- i problemi generali della classificazione sono affrontati sistematicamente
- gran parte del software per la gestione dei thesauri è progettato per aderire allo standard
- compatibilità e integrazione tecnologica ne risultano avvantaggiate

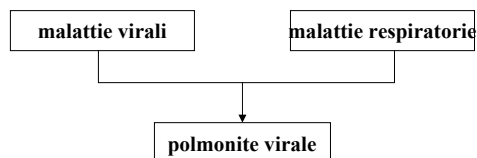
Problemi generali della classificazione

• Arbitrarietà nella scelta delle relazioni semantiche, in particolare la relazione associativa. Lo standard ANSI/NISO ne propone molte, fra cui:

causa-effetto; processo-agenti; concetto-proprietà; azione-prodotto, ecc.

• Arbitrarietà nella scelta dei termini preferiti. ANSI/NISO fornisce indicazioni lessicali (per es. **frequenza del termine, specificità**) e formali (**categoria grammaticale; ortografia; plurale/singolare; abbreviazioni, ecc.**)

• Trattamento e rappresentazione (nell'interfaccia) delle poligerarchie, cioè di gerarchie dove un termine è sottordinato di più termini. Es.



Poligerarchie: un esempio nel web (*trad. italiana*)

Alcuni sistemi fanno ampio uso di poligerarchie, che vengono indicate nell'interfaccia con il segno @ accanto ai termini che fanno riferimento ad altri rami della gerarchia.

Yahoo! Indice
Intrattenimento

[Domestico](#) > [Intrattenimento](#)

Yahoo Interno!

- [Yahoo! Intrattenimento](#) - movies, musica, TV, astrologia e più.

Categorie

▪ Attori ed attricesse (15155) NEW!	▪ Magia (330) NEW!
▪ Parchi di tema e di divertimento (444)	▪ Movies e pellicola (29291) NEW!
▪ Premi (15)	▪ Musica (87785) NEW!
▪ Libri e letteratura @	▪ Notizie e mezzi (381)
▪ Chiacchierate e tribune (77)	▪ Organizzazioni (12)
▪ Comedy (1121)	▪ Effettuando Le Arti @
▪ Comics ed animazione (5441) NEW!	▪ Radio @
▪ Elettronica Di Consumatore (1338)	▪ Cose Ripartite con scelta casuale (60)
▪ Concorsi, indagini e scrutinio (407) NEW!	▪ Rassegne (38)

Le poligerarchie creano complessità di navigazione e disorientamento.

Yahoo! Directory
Entertainment

[Home](#) > [Entertainment](#)

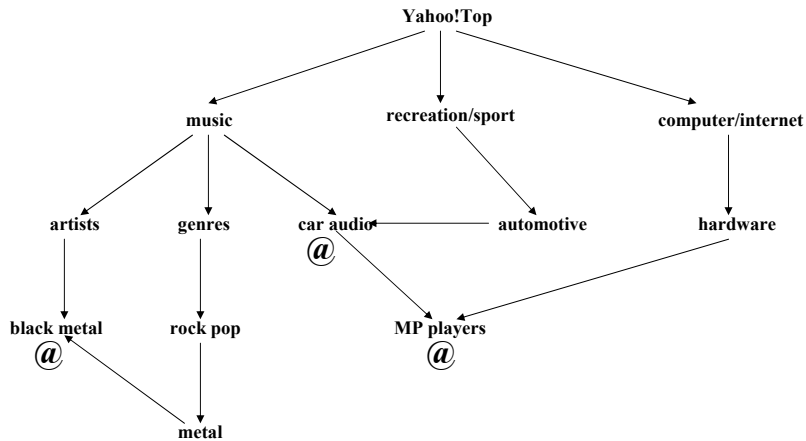
Inside Yahoo!

- [Yahoo! Entertainment](#) - movies, music, TV, astrology, and more.

Categories

▪ Actors and Actresses (15155) NEW!	▪ Magic (330) NEW!
▪ Amusement and Theme Parks (444)	▪ Movies and Film (29291) NEW!
▪ Awards (15)	▪ Music (87785) NEW!
▪ Books and Literature@	▪ News and Media (381)
▪ Chats and Forums (77)	▪ Organizations (12)
▪ Comedy (1121)	▪ Performing Arts@
▪ Comics and Animation (5441) NEW!	▪ Radio@
▪ Consumer Electronics (1338)	▪ Randomized Things (60)
▪ Contests, Surveys, and Polls (407) NEW!	▪ Reviews (38)
▪ Employment (561) NEW!	▪ Shopping and Services@
▪ Entertainment and Media Production@	▪ Sports Entertainment (1234) NEW!
▪ Events (251)	▪ Television Shows (12972) NEW!
▪ Food and Drink@	▪ Trivia (97)
▪ Gambling@	▪ Virtual Cards (527)
▪ Games@	▪ Web Directories (34)

Un esempio di poligerarchia da Yahoo



Classificazione a “faccette”

Altri sistemi hanno adottato la classificazione a “faccette”, o punti di vista.

Non ci si preoccupa di **collocare** un oggetto in una gerarchia, ma di **descriverlo** in termini di sue proprietà o caratteristiche mutuamente esclusive.

Non una singola grande tassonomia, ma tante piccole tassonomie che rispecchiano altrettanti diversi punti di vista.

La classificazione a faccette è stata proposta negli anni '30 dal bibliotecario indiano Ranganathan.

Oggi è adottata da molti siti, soprattutto di ambiente commerciale.

L'esempio dei vini

Faccette

Tipo

Regione di origine

Produttore o cantina

Annata

Prezzo

Valori di esempio

Rosso (Merlot, Pinot noir), Bianco (Chablis, Chardonnay), Frizzante, Rosé, Dessert,...

Australia, California, Francia, Italia, ...

Ferrari, Berlucci, Santa Maria della Versa, Blackstone, Clos, du Bois, ...

1968, 1990, 1999, 2002,...

.....

Classificazione a "faccette": i vini per regione, tipo e ...

The screenshot shows the wine.com website interface. At the top, there is a navigation bar with links for 'home', 'shop', 'gift center', 'corporate gifting', 'wine clubs', 'fine wine', 'top sellers', 'top rated', 'accessories', and 'wine tasting'. Below this is a search bar with 'Shipping to: CA' and 'California' selected. A 'Browse by type' sidebar on the left lists categories like 'Valentine's Day Gift Ideas', 'Wine Packs', 'Region' (France, California, Italy, Washington, South America, Spain, Australia, Oregon, Other Regions), and 'Type' (Red Wines, White Wines, Sparkling, Dessert). The main content area features a 'Welcome to wine.com!' message, a '10% Savings on Case Purchases' banner, and several product listings. One listing is for 'Dom Perignon 1995 with Gift Box' with a special price of \$99.99. Another is for 'wines.com Taste of Italy Valentine's Six Pack' with a special price of \$89.99. A 'limited time offer' for a wine club is also visible.



... produttore



The screenshot shows the wine.com website interface. On the left, there is a 'browse by winery' section with a list of wineries including B, C, D, E, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, V, W, X, Z, and 'Other wineries'. The main content area displays several wine products with their names, prices, and promotional offers. For example, 'L'arche v. 1999 zombi' is priced at \$14.99 (15% off), 'L'arche v. 1999 Estate' at \$110.99 (20% off), and 'Insta-pull Lever Corkscrew' at \$39.99 (20% off). A 'what's hot' section on the right highlights 'Gustaf Z800' and 'Chateau St. Michel'. At the bottom, there is a 'VeriPro by VISA' logo and a phone number: 'To order wine by phone, call 1-877-289-6886'.

Navigazione per regioni

The screenshot shows the wine.com website with the 'Wine Shop' section. The navigation bar includes 'home', 'shop', 'gift center', 'corporate gifting', 'wine clubs', 'fine wine', 'top sellers', 'top rated', 'accessories', and 'wine tasting'. The 'shop' menu is expanded to show 'by type', 'by region', 'by winery', 'top ten', and 'wine packs'. The 'Shipping to' dropdown is set to 'CA' (California). The main heading is 'Wine Shop BROWSE BY REGION'. Below this, there is a paragraph: 'Perhaps you're in the market for a pungent Italian wine. Maybe something exhilarating from new Zealand? Or are you strictly into Californian wines. In this portion of the site, you can shop all over the world!'. A list of regions is provided: Australian, Austrian, Californian, Canadian, French, German, Italian, New Zealand, Oregon, Portuguese, South African, and South American. A globe image is shown next to the list. On the right, there is a 'wine.com Favorites' section with a list of products: 'NASC 95 Cabernet Sauvignon', 'Coccha v. 1999 2000 Marques de Casa Coccha Cabernet Sauvignon' (priced at \$14.99), and 'Penfold 1928 Bin 407 Cabernet Sauvignon' (priced at \$19.99).

Ricerca a “faccette” combinabili



Pregi e futuro della classificazione sfaccettata

- Dalla parte dei **progettisti**:

grande potenza e flessibilità nella presentazione di scelte di navigazione e ricerca.

- Dalla parte degli **utenti**:

possibilità di formulare interrogazioni simulando il linguaggio naturale

E' un approccio destinato ad imporsi sulle soluzioni a tassonomia unica.

Metadati, vocabolari controllati e thesauri saranno i mattoni fondamentali dei futuri siti web sempre più flessibili e sfaccettati.

ESERCIZIO N. 5 e N.6

Dovete dimostrare di conoscere gli strumenti per l'indicizzazione

- Vocabolari controllati
- Anelli di sinonimi
- Termini preferiti (Authority file)
- Tassonomie e schemi organizzativi (Gerarchie tra termini di un vocabolario)

Trovate in rete un esempio per ciascuno degli strumenti elencati