# Introduzione all'indicizzazione / Serafina Spinelli

Bologna, 10.11.2005

## Introduzione e Definizioni

### Un po' di contesto

Fra gli obbiettivi di una biblioteca (e non solo...):

1. Organizzare/rappresentare oggetti/informazioni
2. Rendere possibile la ricerca/il recupero di oggetti/informazioni

Perchè?

- motivi "di scala"...
- massimizzare l'efficacia
- minimizzare le barriere

### Indicizzazione

Di cosa parliamo quando parliamo di "indicizzazione".

**1. Serrai (1974)**
Indicizzare vuol dire assegnare uno o più caratteri di riconoscimento o di recupero ad un documento. Questi caratteri possono esprimersi come simboli di una classe o sottoclasse in un sistema classificatorio, o come dei soggetti, ossia singole parole o combinazioni di parole, in un catalogo, detto appunto per soggetti.

**2. Maltese (1982)**
Indicizzare un documento significa indicarne il contenuto dal punto di vista del soggetto, di ciò di cui si parla, dare del documento una descrizione da indice, cioè una descrizione molto breve del suo soggetto.

**3. Caffo (1988)**
L'indicizzazione è la tecnica per costruire accessi attraverso il contenuto semantico dei documenti, distinti da altre forme di accesso, e comprende sia il processo di analisi concettuale del documento, sia la traduzione del contenuto informazionale del documento in un linguaggio d'indicizzazione.

**4. Petrucciani (1984)**
L'indicizzazione consiste nell'attribuire ai documenti una rappresentazione contratta, più o meno strutturata, che permette una esplorazione selettiva non praticabile sui documenti stessi.

**5. Petrucciani (1991)**
L'indicizzazione consiste nell'attribuire ai documenti, per il recupero, delle brevi "rappresentazioni" indicative del loro contenuto (voci d'indice, intestazioni di soggetto, simboli di classificazione, parole chiave, ecc.). In genere, contemporaneamente, viene ad essi attribuita una descrizione bibliografica, cioè una "carta d'identità" contenente i propri connotati.

**6. Cheti (MIAC, 1996)**
E' l'operazione mediante la quale si creano gli accessi al contenuto semantico del documento. Consta delle fasi di analisi concettuale e di traduzione dei concetti individuati e delle relazioni logiche individuate nei termini e nelle forme proprie del linguaggio di indicizzazione prescelto.

**7. Bogliolo (1998)**
Indicizzare significa creare indici, cioè un'organizzazione sistematica di oggetti simbolici (parole, frasi, codici alfa-numerici) finalizzati a consentire a un utente di trovare l'informazione relativa a un documento ospitato in un determinato archivio.

**8. Gnoli (2000)**
Rappresentazione sintetica [dei contenuti] dei documenti mediante l'attribuzione di indici ricercabili.

**Gli standard:**

**1. ISO 5127/5 (Vocabolario, 1981)**
L'indicizzazione è l'azione mirante a rappresentare i risultati dell'analisi di un documento con gli elementi di un linguaggio naturale o di un linguaggio documentario, generalmente per facilitarne il reperimento.

**2. ISO 5963 (Metodi per l'analisi dei documenti..., 1985)**

L'indicizzazione è l'azione di descrivere o identificare un documento nei termini del suo contenuto concettuale.

## Sistema d'Indicizzazione

Insieme delle procedure per l'organizzazione e la rappresentazione del contenuto dei documenti.

## Linguaggio d'Indicizzazione o Linguaggio Documentario

Codice attraverso il quale si rappresenta e trasmette il contenuto informazionale del documento, allo scopo di renderne possibile il recupero.

Strumento per la rappresentazione coerente, formalizzata e sintetica del contenuto concettuale dei documenti, funzionale alla segnalazione e al reperimento dei documenti stessi [GRIS].

Insieme di termini (**semantica**) ammessi per descrivere il contenuto dei documenti o per organizzare una ricerca su di essi, e insieme di regole (**sintassi**) che stabiliscono l'ordine e le modalità di citazione dei termini.

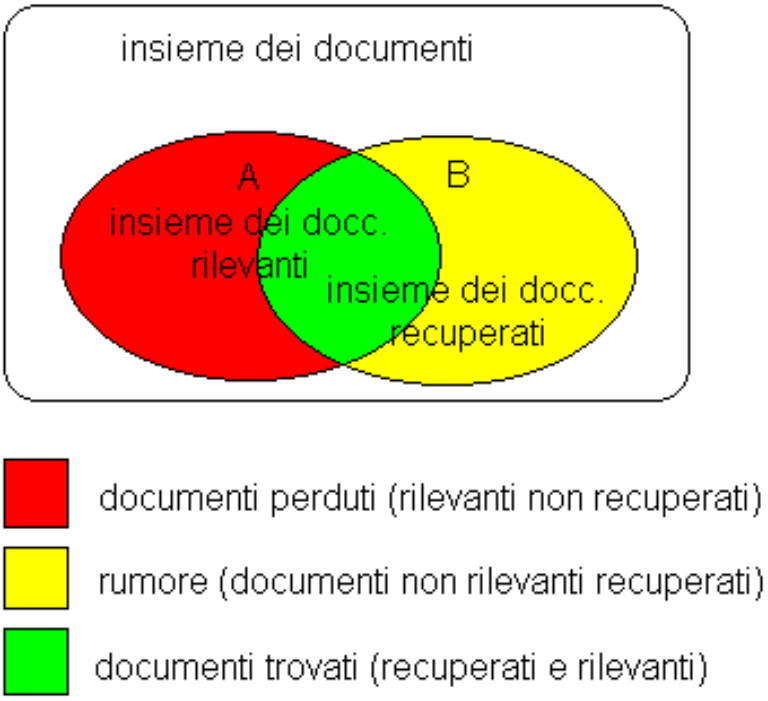
## Sistema Documentario

Contesto in cui si effettuano le procedure di trattamento della documentazione (Biblioteca, Centro di documentazione...) e di ricerca e reperimento dell'informazione.

## Information retrieval

Insieme di tecniche per il reperimento dei documenti rilevanti rispetto ad una determinata esigenza informativa dell'utente.

Criteri di valutazione dei risultati della ricerca: rilevanza, richiamo e precisione.




## Grado di Richiamo

Numero di documenti rilevanti recuperati rispetto al numero totale di documenti rilevanti della biblioteca.
Formula di calcolo:
*GR = docc. rilevanti recuperati : totale docc. rilevanti*

## Grado di Precisione

Numero di documenti rilevanti rispetto al numero di documenti recuperati.
Formula di calcolo:
*GP = docc. rilevanti recuperati : totale docc. recuperati*

## Rumore

Documenti recuperati non rilevanti.

## Futility point

Quantità massima di documenti fra i quali un utente è disposto a cercare quelli che effettivamente rispondono alle sue esigenze informative. Di solito stimato attorno a 30.

# Il Linguaggio d'Indicizzazione

Benchè possa essere utilizzato come linguaggio d'indicizzazione anche il linguaggio naturale, di norma per linguaggio d'indicizzazione s'intende un **linguaggio controllato** o **formalizzato** o **artificiale**, con il quale si faccia **indicizzazione per concetti** e non indicizzazione per termini.

**Linguaggio controllato** è un linguaggio all'interno del quale sia esercitato il controllo degli equivalenti semantici e sintattici (della sinonimia semantica e sintattica), e cioè:

- che ogni concetto o combinazione di concetti sia espresso sempre dallo stesso termine o combinazione di termini
- che ogni termine o combinazione di termini rappresenti sempre lo stesso concetto o combinazione di concetti, e quindi
- che vi sia un rapporto biunivoco fra termini e concetti

Fare **indicizzazione per concetti** (o indicizzazione **assegnata**) significa:

- stabilire una descrizione standardizzata per ogni concetto
- usare questa descrizione ogni volta che è appropriata, indipendentemente da fatto che coincida con quella usata dall'autore
- usare questa descrizione in fase di ricerca in modo da realizzare l'"incontro" con i documenti

L'indicizzazione per concetti è perciò impegnativa in fase di **input** (cioè di indicizzazione del documento), ma economica in fase di **output** (cioè di ricerca).

Fare **indicizzazione per termini** (o indicizzazione **derivata**) significa:

- utilizzare per indicizzare il documento gli stessi termini usati dall'autore (ad es. derivandoli dal titolo, sottotitolo, abstract ecc.), fondandosi sulla presunzione che ne rappresentino bene il soggetto
- in fase di ricerca, cercare di individuare e combinare correttamente tutti i termini che gli autori potrebbero aver utilizzato per esprimere quel soggetto

L'indicizzazione per termini è perciò economica in fase di **input** (cioè di indicizzazione del documento), ma impegnativa in fase di **output** (cioè di ricerca).

Come il linguaggio naturale, il linguaggio d'indicizzazione è costituito da **vocabolario** (insieme di termini) e **sintassi** (insieme di regole di combinazione dei termini) attraverso i quali **si dà un nome ai concetti**, e cioè:

- vengono formulate le rappresentazioni del contenuto semantico dei documenti
- vengono effettuate le richieste informative sul contenuto concettuale dei documenti da parte degli utenti

Si può distinguere fra **vocabolario d'indicizzazione** (costituito dai termini **preferiti**, cioè effettivamente attribuibili al documento per la sua indicizzazione) e **vocabolario d'accesso** (costituito sia dai termini preferiti che dai **termini non preferiti**, cioè dai termini che non possono essere utilizzati per l'indicizzazione e che rimandano a termini preferiti), che può essere usato dall'utente in fase di ricerca e che lo indirizza verso i termini del vocabolario d'indicizzazione.

# Linguaggi d'Indicizzazione

## Enumerativi

Sono quelli che enumerano tutti e soli i termini o le combinazioni di termini (quindi sia soggetti semplici che soggetti composti) che l'indicizzatore può usare. In generale sono:

- più rigidi
- più difficili da aggiornare
- meno espressivi (possono esprimere solo ciò che elencano)

Esempi: CDD, [CDU], Soggettario, LCSH

## Analitico-Sintetici

Sono quelli che elencano solo soggetti semplici, che vanno poi combinati secondo appropriate regole sintattiche. In generale sono:

- più flessibili
- più facili da aggiornare
- più espressivi

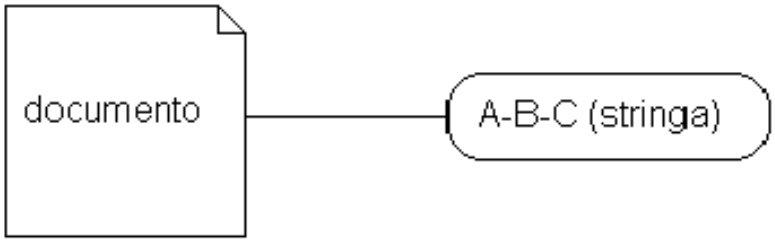
Esempi: Classificazioni a faccette, Bliss, Thesauri (propriamente "vocabolari")

## Precoordinati

Sono quelli nei quali i termini che esprimono i concetti (A, B, C) vengono coordinati (cioè combinati secondo regole sintattiche che ne determinano l'ordine di citazione) prima, cioè al momento dell'indicizzazione.

La stringa che ne risulta è collegata nel suo insieme al documento e dà un'immagine complessiva del contenuto del documento.

Esempi: CDD, CDU, Soggettario

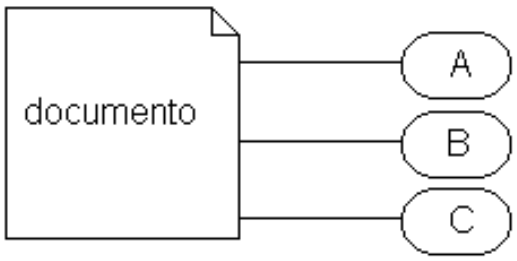



## Postcoordinati

Sono quelli in cui all'atto dell'indicizzazione i termini vengono collegati direttamente al documento e non fra di loro. La coordinazione viene fatta solo al momento della ricerca da parte dell'utente, attraverso strumenti come gli operatori booleani.

Ogni termine dà accesso al documento ma non ne descrive il contenuto complessivo.

Esempi: Thesauri non integrati con norme sintattiche, Parole chiave.




Approfondimenti: Risorse on-line per l'indicizzazione; Introduzione ai thesauri; Thesaurus regionale toscano; Catalogo della Biblioteca della Giunta regionale toscana; introd. alla CDD su Alice; la pagina della Classificazione Decimale Dewey della LIUC; ricerca per CDD nell'opac UniBo; introd. alla CDU su Alice; pagina CDU Online. Per fare una pausa: la finalissima di "Lotta di classe" di Kurzweil. Ma il 616.85270086947095335 esiste davvero?.

# Operazioni fondamentali dell'indicizzazione

## 1. Analisi concettuale

Identificazione del contenuto concettuale del documento per poterne poi organizzare la rappresentazione attraverso i codici del linguaggio documentario.

Fonte normativa: **ISO 5963 del 1981**: *Methods for examining documents, determining their subjects and selecting indexing terms*, in 9 sezioni:

1. scopi e campi di applicazione della norma;
2. altre norme correlate;
3. definizione dei termini usati;
4. analisi dei tre stadi dell'indicizzazione (1. esame del documento e determinazione del soggetto; 2. identificazione dei concetti principali; 3. traduzione nei termini di un linguaggio di indicizzazione);
5. analisi del primo stadio e raccomandazioni sulle parti del documento da considerare più attentamente;
6. identificazione dei concetti secondo un procedimento di scomposizione del soggetto simile all'analisi per faccette; scelta dei concetti da indicizzare;
7. problema della selezione dei termini di indicizzazione, con riferimento alla ISO 2788 (thesauri);
8. controllo di qualità e coerenza dell'indicizzazione;
9. conclusioni, raccomandazioni di standardizzazione dei metodi, principali problemi dell'analisi e direttive generali.

In dettaglio:

Le principali **fonti** per l'esame del documento e l'accertamento del contenuto sono:

titolo, sottotitolo, indice, sommario, introduzione, conclusioni, riferimenti bibliografici, illustrazioni, fonti esterne.

L'individuazione dei concetti secondo il procedimento di 'scomposizione del soggetto' viene guidata da una serie di domande che costituiscono la cosiddetta **lista di controllo**:

- il documento tratta di un prodotto, di un fenomeno, di una condizione particolari?
- è presente un concetto di attività (azione, operazione, processo?)
- c'è un oggetto verso cui l'attività è diretta, che ne è il traguardo, il destinatario, il paziente?
- si parla anche di agenti o fattori dell'attività?
- ci si riferisce a maniere particolari (strumenti, tecniche, metodi) per compiere una certa azione?
- questi elementi sono considerati nel contesto di un particolare ambiente, località, area geografica, epoca?
- sono indicate delle variabili dipendenti o indipendenti?
- l'argomento è trattato da un punto di vista o con un metodo particolare, normalmente non associati ad esso?

Approfondimenti: il Manuale ipertestuale di analisi concettuale.

## 2. Traduzione nel linguaggio d'indicizzazione

Traduzione dei termini usati nell'analisi concettuale nei termini o segni del linguaggio documentario.

### 2.1 Scelta dei termini

Scelta dei termini che, all'interno del vocabolario del linguaggio d'indicizzazione adottato, sono deputati a rappresentare univocamente i concetti identificati in fase di analisi concettuale.

### 2.2 Costruzione della stringa

Ordinamento dei termini secondo l'ordine di citazione previsto dal linguaggio adottato.

# Principi fondamentali dell'indicizzazione per soggetto

Focalizziamo la nostra attenzione sull'indicizzazione per soggetto, intendendo per questa:

> la rappresentazione coerente, formalizzata e sintetica del contenuto concettuale dei documenti, funzionale alla segnalazione e al reperimento dei documenti stessi, per mezzo di un linguaggio di **indicizzazione verbale di tipo precoordinato**, costituito da:
> - un insieme controllato di termini scelti dalla lingua naturale per esprimere univocamente i singoli concetti (**vocabolario**)
> - un insieme di norme che regolano la combinazione dei termini in sequenze sintatticamente unitarie (**sintassi**).

I principi fondamentali dell'indicizzazione per soggetto sono principi che sottostanno all'attività d'indicizzazione:

- costituiscono le direttive generali alle quali deve uniformarsi l'attività d'indicizzazione
- determinano i requisiti o qualità fondamentali di un sistema d'indicizzazione

### Uniformità e unità

**Uniformità**: all'interno di un certo linguaggio, un concetto o combinazione di concetti è sempre rappresentato da un solo termine o sequenza di termini (controllo della sinonimia).

**Unità**: all'interno di un certo linguaggio, un termine o sequenza di termini rappresenta sempre un solo concetto o combinazione di concetti (controllo della polisemia).

Il **controllo della sinonimia** si realizza in campo **semantico** attraverso la scelta di un termine preferito e l'approntamento di rinvii dai sinonimi e quasi-sinonimi, in campo **sintattico** attraverso l'applicazione di una ordine di citazione uniforme dei termini nella costruzione delle stringhe. In caso di più ordini di citazione ammessi, va effettuata la scelta di un ordine di citazione preferito e l'approntamento di rinvii dagli altri.

Il **controllo della polisemia** avviene precisando, nel caso di termini omonimi o polisemici, il preciso significato col quale il termine è utilizzato all'interno del linguaggio d'indicizzazione.

## Esaustività

Numero di concetti identificati come essenziali per la descrizione del soggetto ed effettivamente tradotti nei termini del linguaggio d'indicizzazione.

Sommarizzazione (identificazione del soggetto complessivo) vs indicizzazione approfondita (estrazione di tutti i concetti ritenuti importanti).

In un linguaggio che esprima in forma di descrizione sintetica il contenuto del documento, l'indicizzatore dovrebbe identificare tutti i concetti essenziali per esprimere il soggetto, ossia il tema complessivo o centrale o **tema di base** del documento.

## Coestensione

La descrizione del soggetto del documento dovrebbe essere tradotto in **una sola stringa** che contenga tutti gli elementi indispensabili ad individuare il soggetto, e non in più stringhe, solo dalla cui unione si evince il soggetto complessivo.
Es.: soggetti di: *Guida alla catalogazione in SBN : libro antico* in Opac LC e in Opac Unibo.

## Specificità

**Precisione** con cui un particolare concetto identificato nel soggetto del documento è rappresentato dal linguaggio d'indicizzazione.

I concetti dovrebbero essere espressi nel modo più specifico possibile, specie se il vocabolario è strutturato in modo tale da garantire l'accesso anche partendo da termini più generali.

Il problema è particolarmente sentito nel caso delle classificazioni, che raramente arrivano ad un livello di dettaglio sufficiente da garantire un buon grado di specificità per le biblioteche specializzate.

## Coerenza

Omogeneità di trattamento da parte dello stesso indicizzatore in tempi diversi, o da parte di diversi indicizzatori.

E' un requisito dipendente dall'indicizzatore o dal gruppo di indicizzatori, e dalla qualità degli strumenti di indicizzazione.

## Accessibilità

E' un requisito dipendente dal software del sistema d'indicizzazione, che dovrebbe prevedere una molteplicità di accessi alle rappresentazioni del contenuto semantico dei documenti.

Ad es.:

- deve consentire l'accesso alla stringa non solo nel suo complesso ma anche da ognuno dei termini che la costituiscono;
- deve consentire l'accesso non solo dai termini preferiti ma anche, con meccanismi di rinvio, a partire da quelli non preferiti;
- deve consentire di accedere a ciascun termine a partire da tutti quelli ad esso correlati, esplicitandone i rapporti semantici.

## Criteri e metodi della teoria moderna dell'indicizzazione

La teoria moderna dell'indicizzazione ha elaborato una serie di criteri e di metodi volti a garantire il rispetto dei principi fondamentali dell'indicizzazione.

I più importanti:

### definizione delle relazioni semantiche e delle relazioni sintattiche

Le **relazioni semantiche** o **a priori** sono quelle che un termine intrattiene con altri termini in virtù del proprio significato, sono universalmente valide ed indipendenti dal documento cui il termine è assegnato.

Sono le relazioni esplicitate dai thesauri, dagli schemi di classificazione (parzialmente), dalla struttura sindetica (richiami - v.a., ** - e rinvii - v. * -) dei soggettari.

Le **relazioni sintattiche** o **a posteriori** sono quelle che un termine intrattiene con gli altri termini della stessa stringa in virtù dei rapporti determinati dal soggetto del documento cui è assegnata la stringa, e sono valide solo nel contesto di quella stringa.

Sono le relazioni esplicitate dalle stringhe di soggetto.

### uso dell'analisi categoriale per l'analisi delle relazioni sintattiche

Consiste nell'analisi dei termini della stringa sulla base di parametri come:

- **categorie generali**: es.: entità astratte, concrete, individuali
- **faccette**, cioè principi di suddivisione che, applicati ad una certa categoria, generano gli elementi relativi:
  es.: <persone>
  - secondo la professione
  - secondo la razza
  - secondo l'età
- **ruoli**: es.: oggetto, azione transitiva, agente, parte, proprietà, localizzazione

L'analisi categoriale aiuta l'indicizzatore a:

- ordinare i termini nella stringa
- scegliere la forma singolare o plurale

- scegliere la forma composta o scomposta
- stabilire le relazioni a priori fra termini

### ordine di citazione e scelta della costruzione passiva

L'ordine di citazione si ispira in genere a due criteri:

- rispettare i nessi logici fra i termini, in particolare le relazioni uno-a-uno, interpretabili in termini di ruoli
- se nel soggetto è presente un concetto di attività, ordinare i termini ponendo in prima posizione quello che svolge il ruolo di oggetto che subisce l'azione, secondo la sequenza base OGGETTO-AZIONE-AGENTE

Alcuni fra i più noti ordini di citazione:

- **Ranganathan**: Personalità-Materia-Energia-Spazio-Tempo
- **Ordine di Citazione Standard**: Cose-Specie-Parti-Materiali-Proprietà-Processi-Operazioni-Agenti-Spazio-Tempo
- **Coates**: Cosa-Parte-Materiale-Proprietà-Azione-Agente

Approfondimenti: Guida GRIS.

## Il panorama internazionale

### GSARE

Guidelines for subject authority and reference entries / IFLA. - München : Saur, 1993.

Si tratta di linee guida per la creazione e la manutenzione di liste di voci di autorità di soggetto (analogamente a quanto fa GARE per l'authority control in generale), emanate dall'IFLA, che tendono alla realizzazione di coerenza e uniformità nella scelta della forma dei descrittori di soggetto.

### FRBR

Functional Requirements for Bibliographic Records : Final Report / IFLA Study Group on the Functional Requirements for Bibliographic Records ; approved by the Standing Committee of the IFLA Section on Cataloguing. München : SAUR, 1998.

Nel 1991 lo Standing Committee dell'IFLA Section of Cataloguing ha commissionato uno studio che definisse i **requisiti funzionali delle registrazioni bibliografiche**.
Lo scopo di questo studio è delineare in termini definiti con chiarezza le funzioni svolte dalla registrazione bibliografica relativamente ai diversi media, le diverse applicazioni e le diverse necessità dell'utente. Lo studio intende coprire l'arco completo delle funzioni della registrazione bibliografica nel senso più vasto del termine, vale a dire non solo elementi descrittivi ma anche punti di accesso (nome, titolo, soggetto, ecc.), altri elementi di organizzazione (classificazione ecc.) e annotazioni.

Lo studio intende appoggiarsi su basi teoriche solide e condivise (non a caso fa uso di tecniche di costruzione di **modelli del tipo entità-relazione (E-R)**), ma anche evitare ogni pregiudizio verso qualunque tipo di codice di catalogazione esistente.

FRBR definisce **tre categorie di entità**, delle quali i gruppi due e tre esistono attraverso le loro

relazioni con le entità del Gruppo 1.

Le entità del **Gruppo 1** sono l'**opera**, l'**espressione**, la **manifestazione** e l'**item** (la copia, l'esemplare). Esse costituiscono il fondamento del modello e della sua metodologia.

Le entità del **Gruppo 2** comprendono **persone** (individui) o **enti** (organizzazioni o gruppi di individui e/o organizzazioni). Queste entità rappresentano i responsabili della custodia intellettuale o artistica delle entità del Gruppo 1 e della produzione fisica e/o la distribuzione delle manifestazioni del Gruppo 1. Possono fungere anche da soggetti.

Le entità del **Gruppo 3** costituiscono un insieme aggiuntivo di entità che fungono da soggetto di lavori intellettuali: comprendono **concetto** (nozione o idea astratta), **oggetto** (una cosa materiale), **evento** (un'azione o un fatto), o **luogo** ( una localizzazione). Queste entità fungono di solito da soggetti delle opere. Possono essere correlate con una sola opera o una moltitudine di opere e ciascuna opera può essere correlata a varie entità del Gruppo 3.

Gli obiettivi di FRBR vanno oltre la tradizionale attenzione alla descrizione bibliografica, e includono nella propria sfera anche un esame, per quanto meno sviluppato, dei punti di accesso o "elementi di organizzazione".

Gli attributi logici delle entità persona, ente, concetto, oggetto, evento e luogo (quindi le entità dei gruppi 2 e 3) vengono presi in considerazione solo nella misura in cui sono rispecchiati in maniera tipica nel record bibliografico. FRBR in sostanza **non** analizza i dati aggiuntivi che di norma vengono riportati in un record di autorità né le relazioni fra quelle entità che sono in genere rispecchiate dall'apparato sindetico del catalogo. Per le entità del gruppo 3, praticamente individua solo l'attributo "**termine**", che definisce come la parola, frase, o gruppo di caratteri usati per dare un nome o designare un concetto, un oggetto, un evento o un luogo. FRBR riconosce che concetti, oggetti, eventi e luoghi possono essere designati da più termini, o da più varianti formali di un certo termine. Di norma le agenzie bibliografiche selezionano uno di questi termini come intestazione uniforme, mentre gli altri possono essere trattati come termini varianti.

Comunque FRBR riconosce la necessità di estendere in futuro il modello in modo da trattare i dati relativi alle voci di autorità, tanto che l'appendice A cerca di stabilire una corrispondenza fra gli attributi logici delle entità e gli elementi dei dati previsti da ISBD, GARE e GSARE (Guidelines for subject authority and reference entries dell'IFLA).

## Principles Underlying Subject Heading Languages

Nel 1990 la Sezione Classificazione e indicizzazione della Divisione Controllo bibliografico dell'IFLA ha istituito un Gruppo di lavoro sui "Principi sottostanti ai linguaggi d'indicizzazione per soggetto", che ha condotto uno studio articolato in più fasi, avente i seguenti scopi:

1. facilitare l'accesso per soggetto all'informazione a livello internazionale
2. fornire uno strumento che faciliti lo sviluppo di linguaggi d'indicizzazione stabilendo cosa s'intende per un buon linguaggio e quali sono i suoi auspicabili principi di costruzione e applicazione
3. promuovere la comunicazione fra diversi linguaggi, identificando le comunanze e fornendo una struttura per lo studio comparativo
4. fornire un'impalcatura logico-teorica per particolari standard e linee guida per la costruzione o applicazione di linguaggi d'indicizzazione per soggetto.

Per "principi" il gruppo ha inteso i "postulati fondamentali" che devono guidare la costruzione e l'applicazione di qualsiasi linguaggio d'indicizzazione per soggetto i cui obbiettivi siano il miglioramento di precisione e richiamo nella ricerca. Non quindi regole specifiche per la selezione del vocabolario o la costruzione di una semntica o di una sintassi.

Risultato di questo studio è un documento, pubblicato a stampa nel 1999 ma in bozza già nel 1995, articolato come segue:

- 1. parte: contesto, definizioni e principi
- 2. parte: censimento e rassegna dei principi all'interno di un campione di linguaggi effettivamente esistenti.

L'ambito entro cui dovrebbero agire i "principi" è il più ampio possibile, e comprende sia linguaggi precoordinati sia postcoordinati. Vengono individuati **11 principi**, di cui **9 definiti Principi di costruzione**:

- principio dell'intestazione uniforme (*Uniform Heading Principle*): ogni concetto (o individuo) indicizzato all'interno di un linguaggio deve essere rappresentato da un'unica intestazione;
- principio di sinonimia (*Synonymy Principle*): i sinonimi, i termini che esprimono lo stesso concetto, devono essere "controllati", cioè ricondotti all'intestazione prescelta per rappresentare il concetto; questo principio aumenta il potere di richiamo di un linguaggio;
- principio di omonimia (*Homonymy Principle*): termini che possono avere significati diversi devono essere "disambiguati", cioè se ne devono ricavare intestazioni diverse, una per ognuno dei significati; questo principio diminuisce il rumore e aumenta la precisione della ricerca;
- principio di uniformità dei nomi propri (*Naming Principle*): nomi di persone, luoghi, enti, opere, ecc. devono essere espressi nella stessa forma adottata dai cataloghi per autori e titoli;
- principio semantico (*Semantic Principle*): per esprimere la struttura semantica, paradigmatica, del linguaggio, i termini devono essere collegati da relazioni di equivalenza, gerarchiche e di coordinazione;
- principio sintattico (*Syntax Principle*): per esprimere soggetti complessi e composti, la sintassi di un linguaggio d'indicizzazione deve collegare i termini di una stringa di soggetto con relazioni sintagmatiche, e non paradigmatiche (cioè non semantiche);
- principio di coerenza (*Consistency Principle*): ogni nuovo soggetto che entra nel linguaggio deve essere omogeneo per forma e struttura ai soggetti analoghi già presenti;
- principio della garanzia bibliografica (*Literary Warrant Principle*): il vocabolario di un LI dev'essere sviluppato in maniera dinamica, basato sulla garanzia bibliografica (sulla letteratura effettivamente esistente), costruito "a posteriori";
- principio del lessico orientato all'utente (*User Principle*): il vocabolario del LI deve riflettere l'uso linguistico corrente della comunità di utenti cui si rivolge.

e **2 Principi di applicazione**:

- principio dell'intestazione specifica (*Specific Heading Principle*): un'intestazione di soggetto (o un insieme di intestazioni) dev'essere coestesa rispetto al contenuto concettuale del documento cui si riferisce. Il livello di coestensione può essere soggetto ad "aggiustamenti" per evitare problemi di richiamo;
- principio della politica d'indicizzazione (*Subject Indexing Policy Principle*): bisogna sviluppare ed esplicitare le politiche di indicizzazione che guidano l'analisi e la rappresentazione del contenuto concettuale dei documenti, sia per venire incontro alle esigenze degli utenti sia per fare un trattamento coerente della documentazione.

I linguaggi analizzati nella seconda parte dello studio sono di ambito generale (cioè non specialistici), e dotati di una qualche forma di precoordinazione; per la maggior parte sono gestiti dalle biblioteche nazionali e usati nella bibliografia del paese, ma anche adoperati da altri tipi di biblioteche. Molti di loro forniscono manuali o raccolte di regole e liste di intestazioni.
Lo studio ha verificato che tutti i sistemi rispondono in generale ai principi di costruzione (soprattutto a quelli semantici, e soprattutto tramite regole e procedure, più che in maniera dichiarata), mentre i principi di applicazione sono meno espliciti o evidenti.

Le conclusioni del gruppo di lavoro sono state le seguenti:

- lo sviluppo di una dichiarazione dei principi sottostanti ai LIS internazionalmente condivisibile

è possibile, e supportata nella teoria e nella pratica dagli esistenti sistemi di recupero per soggetto;
- gli attuali sistemi d'indicizzazione convergono molto di più sul controllo terminologico e la struttura paradigmatica, molto meno sugli aspetti sintagmatici della rappresentazione della conoscenza;
- il fatto che la pratica di definire politiche d'indicizzazione sistematiche e complete sia meno sviluppata, mostra la necessità di enfatizzare gli aspetti gestionali dei sistemi di recupero per soggetto, e supportarli condividendo esperienze e producendo raccomandazioni;

## IME ICC (International Meeting of Experts on an International Cataloguing Code)

Serie di incontri di esperti di codici catalografici promossi sempre dalla Sezione di Catalogazione dell'IFLA, con l'obbiettivo di definitire nuovi principi di catalogazione che sostituiscano i "principi di Parigi" (1961) e di promuovere lo sviluppo di un codice di catalogazione internazionale per la descrizione bibliografica e l'accesso. Il contesto di riferimento è sempre quello di FRBR. Frutto del primo meeting sono i "Principi di Francoforte" (2003), attualmente in versione draft, disponibili online anche in traduzione italiana.

Approfondimenti: Testo di FRBR (in formato PDF); Seminario FRBR (Firenze, 27-28 gennaio 2000); Pino Buizza, Indicizzazione per soggetto e FRBR, "Bibliotime", VI, 1 (marzo 2003); Principles Underlying Subject Heading Languages: An International Approach / Maria Inês Lopes; un ampio commento dei "Principi di Francoforte" anche in relazione alla catalogazione semantica è nella relazione di Pino Buizza, Verso nuovi principi e nuovi codici a Bibliocom 2004; Alberto Cheti, Il punto di vista del GRIS sulla "relazione di soggetto" in FRBR a Bibliocom 2004.

## Il progetto di rinnovamento del Soggettario

Nel 2000 la Biblioteca nazionale centrale di Firenze, produttrice della BNI, ha affidato ad un gruppo di lavoro composto in larga parte da membri del GRIS uno *Studio di fattibilità relativo al rinnovamento del Soggettario*. Il Gruppo ha prodotto un documento articolato in sette proposte, e presentato a Firenze, nei giorni 5-6 Aprile 2001, nel corso di un seminario ad inviti finalizzato ad un confronto e una valutazione collettiva delle proposte.

Le proposte avanzate dallo studio hanno riguardato i seguenti ambiti:

1. **la precisione del linguaggio: specificità dei termini e coestensione delle stringhe.**

   La proposta prende le mosse dal riconoscimento della necessità di un linguaggio il più possibile selettivo, cioè in grado di escludere i documenti non pertinenti e selezionare quelli pertinenti. Questo risultato (aumento congiunto delle capacità di richiamo e di precisione) è ottenibile solo assegnando questi due obiettivi ai due differenti piani dell'indicizzazione, quello semantico e quello sintattico. La precisione nella rappresentazione dei singoli concetti dipende dalla specificità dei termini impiegati per esprimerli (specificità terminologica). La precisione nella rappresentazione delle relazioni sintattiche tra i concetti necessari alla definizione di un tema comporta invece la precoordinazione, la scelta di una forma di espressione chiara e non ambigua delle relazioni fra i termini (ordine di citazione+connettivi), la traduzione di un tema unitario in un'unica sequenza sintatticamente strutturata di termini (coestensione).
   Va quindi garantita nel nuovo Soggettario la massima ospitalità lessicale (la possibilità di comprendere nel vocabolario anche termini molto specialistici); il vocabolario va dotato di una coerente struttura semantica che funga da guida per la scelta del termine appropriato e per la gradazione del richiamo da parte dell'utente; deve essere infine abbandonata l'attuale prassi di rappresentare alcuni soggetti composti con più voci non coestese.

2. **le unità di base del linguaggio e le loro combinazioni: termini e stringhe vs. voce principale e suddivisioni.**

   Le unità di base del linguaggio non devono essere quelle attualmente identificate come "voci principali" e "suddivisioni"; esse devono piuttosto essere i termini che esprimono concetti singoli o unitari, la cui combinazione, secondo le regole della sintassi, dà luogo alle stringhe di soggetto, che rappresentano i soggetti composti.
   La sequenza *Voce principale - Suddivisioni generiche - Suddivisione geografica - Suddivisione cronologica - Suddivisione formale*, attualmente usata dal Soggettario, non può assicurare un ordine di citazione standard, che è invece assicurato dalla combinazione di singoli termini secondo regole derivanti dalla loro funzione logico-sintattica.

3. **i metodi di controllo delle relazioni sintattiche.**

   In un moderno sistema di indicizzazione, la forma preferita di precoordinazione è quella sintetica. A differenza di un linguaggio enumerativo, le cui espressioni (stringhe di soggetto) sono solamente quelle specificate in una lista di autorità, un linguaggio sintetico è un linguaggio che consente di generare stringhe di soggetto mediante la combinazione dei termini del vocabolario secondo regole di sintassi.
   Il Soggettario attuale, pur essendo teoricamente enumerativo (elenca anche le voci costruite, nelle quali le suddivisioni compaiono al seguito della voce principale e non costituiscono una voce a sé), è nell'uso ampiamente sintetico, e riconosce esplicitamente alle suddivisioni generiche elencate un carattere esemplificativo. Tuttavia, anche quando la costruzione di nuove voci è affidata all'indicizzatore, la loro correttezza è basata prevalentemente sull'autorità della lista, mediante istruzioni specifiche o l'analogia con voci in essa già presenti, piuttosto che sull'aderenza ad una regola generale.
   Nel Soggettario attuale, il problema della costruzione delle stringhe di soggetto assume fondamentalmente la forma della scelta della voce principale, intesa come "prima voce", "parola d'ordine", "soggetto", e la successiva aggiunta delle eventuali suddivisioni. La scelta della voce pricipale si basa, più che su un ordine di citazione prestabilito, su una valutazione del grado di importanza, di significatività dei concetti che costituiscono il soggetto: il concetto più importante, più significativo è espresso nella voce principale; gli altri, con funzione completiva, nelle suddivisioni. Il Soggettario non enuncia criteri di scelta riconducibili ad un principio unico, ma anzi adotta comportamenti diversi a seconda dei casi, per cui convivono criteri di tipo lessicale (p.e., la precedenza assegnata ai nomi di persona, che dà luogo al soggetto biografico), criteri di tipo semantico (p.e., la precedenza assegnata alle entità rispetto alle attività, ai processi e alle discipline), criteri di tipo sintattico (p.e., la priorità assegnata all'intero rispetto alle sue parti, o ad un individuo o una classe rispetto alle proprietà, agli aspetti, ai punti di vista, alle azioni che li riguardano).
   La scelta delle suddivisioni si basa sulla loro ricerca nella lista, dove si possono trovare due tipi di suddivisioni: quelle proprie di una voce principale e quelle libere, cioè applicabili a una o più categorie di voci principali. Negli ultimi due aggiornamenti del Soggettario, compaiono liste di suddivisioni generiche separate e corredate da note, rinvii e dall'indicazione del tipo di voce principale con cui sono usate.
   La proposta prevede di sostituire il modello sintattico attuale, che consiste nella distinzione "voce principale/suddivisione", col modello sintattico basato sullo "schema di ruoli", molto più efficace e produttivo, in base al quale è possibile costruire le regole sintattiche, secondo un'organizzazione gerarchica del tipo:

   - Principi (p.e., la costruzione passiva)
   - Regole (p.e., un'azione transitiva segue l'oggetto verso cui è diretta)
   - Istruzioni specifiche (p.e., un'istruzione relativa ad un certo termine o a un'intera categoria di termini)

4. **le norme per l'ordine di citazione.**

   In un linguaggio di tipo sintetico (quale dovrebbe diventare il nuovo Soggettario), la

disposizione degli elementi significativi nella stringa (ordine di citazione) viene regolata da norme, la cui efficacia dipende in massima parte dal tipo di analisi categoriale e dai principi generali sui quali sono fondate.
L'analisi categoriale è un'analisi degli elementi linguistici che definiscono il tema da indicizzare, volta all'identificazione della loro categoria di appartenenza: nel vecchio Soggettario le categorizzazioni più diffuse erano di tipo lessicale (nomi propri/nomi comuni) e/o semantico (entità/attività, concreto/astratto ecc.), che consentivano perciò di stabilire l'ordinamento sintattico solo di sequenze poco articolate, costituite da un numero di elementi non superiore al numero delle categorie individuate, ed appartenenti ognuno ad una categoria differente (un nome proprio ed uno comune, un concetto di entità ed uno di attività, un concetto concreto ed uno astratto).
Per regolare la costruzione di strutture sintattiche di maggiore ampiezza ed articolazione è indispensabile quindi riferirsi non al significato, ma alla funzione logica degli elementi della stringa (scopo/strumento, azione/oggetto, azione/agente, intero/parte ecc.). L'analisi categoriale di tipo logico-funzionale è modulare ed esaustiva, e permette di applicare i medesimi criteri nella rappresentazione di temi di diversa complessità, garantendo, con la formulazione di stringhe sempre coestese, la massima espressività del linguaggio, e livelli elevati di predittività e coerenza.
Ai due diversi criteri di analisi categoriale corrispondono principi sintattici differenti. In relazione alle categorizzazioni lessicali e semantiche sono stati infatti stabiliti principi di ordinamento che attengono al significato degli elementi da disporre nella stringa, come il principio della concretezza decrescente, formulato da Ranganathan, e quello del concetto più significativo, formulato da Coates. Alle categorizzazioni logico-funzionali sono invece associati due principi sintattici basati esclusivamente sulle relazioni che uniscono i concetti nella definizione di un tema: il principio della relazione uno a uno (i concetti devono essere citati nella stringa in modo che sia preservata ed evidenziata, per quanto possibile in un ordinamento lineare, ogni relazione sintattica diretta) ed il principio della dipendenza logica (tra due concetti deve essere citato per primo quello che costituisce il presupposto logico della funzione svolta dall'altro). Da quest'ultimo principio discende quello della costruzione passiva, e quindi la norma di citare sempre come concetto chiave (o voce principale) il concetto verso il quale è diretta un'azione o che è obiettivo di una funzione agentiva o strumentale.

5. **le norme per il controllo della morfologia dei termini.**

Gli aspetti del controllo terminologico presi in considerazione in questa proposta sono il numero, la scomposizione e la disambiguazione. La proposta è costruita tenendo conto di quanto previsto dalla norma ISO 2788-1986 (sulla costruzione dei thesauri) in maniera abbastanza fedele, anche se con qualche scostamento. Si tenga conto che il Soggettario attuale manca di principi e criteri omogenei per il controllo della morfologia dei termini, anche perché essi non erano ancora stati formulati in maniera esplicita e completa al momento della sua pubblicazione (1956).

6. **i metodi di controllo delle relazioni semantiche.**

La formalizzazione del vocabolario e delle relazioni semantiche è un principio irrinunciabile dei linguaggi di indicizzazione. Essa è riconducibile ad alcune regole fondamentali:

- un concetto è rappresentato da uno e un solo termine; a ogni termine corrisponde uno e un solo concetto;
- le relazioni che sono esplicitate nel vocabolario controllato sono sempre relazioni a priori;
- le relazioni tra termini sono a loro volta formalizzate, vale a dire che ogni relazione individua una struttura classificatoria data, il cui significato è omogeneo per tutte le relazioni individuate dalla struttura del vocabolario.

Il controllo terminologico implementato nell'attuale Soggettario riduce a due soli casi tutte le possibili relazioni semantiche tra termini: la relazione tra sinonimi (o relazione di equivalenza, vedi e il reciproco *) e una seconda relazione (vedi anche e il reciproco **), deputata ad esprimere sia le relazioni verso il basso e verso l'alto (la relazione gerarchica dei thesauri) sia

quelle trasversali (la relazione associativa dei thesauri).
La proposta propugna per il nuovo Soggettario l'adozione del modello a tre relazioni (quello tipico dei thesauri), in quanto dotato di una maggiore ricchezza e flessibilità nella rappresentazione delle relazioni semantiche, e sostanzialmente compatibile con il modello a due relazioni (mentre non è vero il contrario).

7. **la struttura di una voce del vocabolario.**

   La proposta è più quella di un record di controllo semantico che di una vera e propria voce di vocabolario, contiene perciò anche elementi per il controllo gestionale e della visualizzazione. Il record risulta così articolato in tre parti:

   - struttura di controllo terminologico
   - struttura di visualizzazione, selezione e compatibilità
   - parte gestionale

Dopo un fase di ulteriori approfondimenti, seminari, e confronti con studiosi ed esperti italiani e stranieri, la fase di studio si è conclusa nel 2002 con la stesura di un *Progetto preliminare*, la messa a punto dei principali aspetti organizzativi (ad es. lo smantellamento del vecchio sistema, la definizione dei costi, le modalità per realizzare il lavoro secondo piani a breve e medio termine, ecc.), e la pubblicazione in volume di tutti i testi documentari e progettuali sotto il titolo di *Per un nuovo Soggettario* (Bibliografica 2002).

Nel 2004 il Progetto e i suoi avanzamenti sono stati sintetizzati in un documento intitolato *Il nuovo Soggettario italiano: dallo studio al progetto*, a cura di Anna Lucarelli, Leda Bultrini e Alberto Cheti, disponibile online sul sito della BNCF.

Il documento illustra le principali **scelte metodologiche** operate dal progetto:

- il modello analitico sintetico piuttosto che quello enumerativo
- l'articolazione in termini e stringhe piuttosto che in voci principali e suddivisioni
- la specificità e la coestensione
- l'analisi categoriale per il controllo delle relazioni semantiche e l'adozione del modello a tre relazioni
- l'analisi dei ruoli per il controllo delle relazioni sintattiche

le componenti **architetturali** del nuovo Soggettario:

- le norme
- il vocabolario (un thesaurus multidisciplinare di circa 60.000 termini strutturati)
- il corredo sintattico-applicativo
- l'archivio delle stringhe di soggetto

alcuni aspetti qualificanti delle future "**Norme**":

- **Principi**
  - Unità (nei due aspetti di uniformità e univocità)
    - Uniformità
    - Univocità
  - Predittività (prevedibilità, nel senso di coerenza e omogeneità)
  - Specificità (nei tre aspetti che seguono)
    - Esaustività (dell'analisi)
    - Coestensione (della stringa)
    - Specificità (dei termini)
- **Analisi concettuale dei documenti**
  - Il soggetto

    - La procedura
        - Esame delle fonti
        - Selezione dei concetti
    - L'enunciato di soggetto
- **Controllo del vocabolario**
    - Morfologia
        - Singolare e plurale (principio della numerabilità e dello scostamento categoriale)
        - Termini composti (metodo di analisi e i criteri di scomposizione della norma ISO 2788)
        - Nomi propri
        - Omografia (tecniche di disambiguazione: preferita la qualificazione generica)
    - Struttura del vocabolario
        - Relazioni di equivalenza
        - Relazioni gerarchiche (generica: sempre, tutto-parte: solo se la parte implica l'intero, esemplificativa: individuo-classe)
        - Relazione associativa (limitata alla definitoria)
    - Categorie e categorizzazione (schema delle categorie individuate, in via sperimentale: Attività, Discipline, Materiali, Oggetti, Organismi, Organizzazioni, Persone, Processi, Spazio, Strumenti, Strutture)
- **Costruzione delle stringhe di soggetto**
    - Analisi dell'enunciato di soggetto
    - Tipi di relazioni (due tipi principali, che attivano due diverse strutture logiche: relazione transitiva e relazione di appartenenza)
    - I ruoli
    - L'ordine di citazione
    - Punteggiatura e connettivi
    - Il nucleo
        - Principio della relazione uno a uno
        - Principio della dipendenza logica
    - Complementi

le principali caratteristiche del **Vocabolario** e la **struttura del record** (termine e suo corredo semantico, di applicazione, gestionale):

- esempio di record strutturato:

**Monumenti**

TT Strutture
BT Costruzioni
*<secondo la forma>*
NT Archi di trionfo
*<secondo la funzione>*
NT Monumenti commemorativi
NT2 Monumenti ai caduti
NT Monumenti nazionali
NT Monumenti sepolcrali
NT2 Mausolei
NT2 Piramidi
RT Opere d'arte
RT Edifici
SN Usare sia nel significato specifico di costruzione consacrata alla memoria di fatti e personaggi importanti; sia nel senso, collettivo, di insieme di costruzioni che testimoniano la storia e l'arte di un luogo, di una nazione, etc., per es., i monumenti di Firenze. È essenziale che il significato di "costruzione" sia sotteso al significato individuato; non si deve cioè considerare monumento ciò che non è costruito (p.e., le tradizioni, gli oggetti, etc., in quanto tali)
HN Usato nel *Soggettario 1956*; le *Liste di aggiornamento 1986-1998* hanno: Monumenti ai caduti dei campi di concentramento
Categoria: Strutture
Classificazione: 725.94 (DDC21)
Fonte1: *Soggettario 1956*
Varianti: Monumento
Identificativo: M000001
Status del record: Esempio
Fonte2: Gruppo

- esempio di record con nota sintattica:

**Appalti**

TT Strumenti
BT Contratti
NT Appalti comunali
NT Appalti edilizi
NT Subappalti
RT Appaltatori
RT Capitolati di appalto
RT Contratti di appalto
Nota sintattica: ***Elem. trans.*** *Segue il termine che rappresenta l'oggetto/meta e precede il termine che rappresenta il beneficiario, lo strumento o l'agente, p.e,* Biblioteche comunali - Servizi - Appalti alle cooperative di lavoro; Servizi pubblici - Appalti degli enti locali [*precedentemente* 1. Lavori pubblici - Appalti, 2. Enti locali - Funzioni in materia di lavori pubblici]; Lavori pubblici - Appalti - Ruolo delle associazioni temporanee di imprese - Legislazione [*precedentemente* 1. Associazioni temporanee di imprese - Legislazione, 2. Lavori pubblici - Appalti - Legislazione]; Lavori pubblici - Appalti - Impiego dell'autocertificazione - Legislazione [*precedentemente* 1. Lavori pubblici - Appalti; 2. Autocertificazione - Legislazione]. ***Parte/Propr.*** *Quando non sia citato un oggetto/meta, segue il termine che rappresenta il possessore (appaltatore), p.e.,* Ferrovie dello Stato - Appalti.
Categoria: Strumenti
Fonte1: *Soggettario 1956*
Identificativo: A000002
Status del record: Esempio
Fonte2: Gruppo

- altro esempio di record con nota sintattica:

**Malattie**
TT Processi
BT Processi patologici
*<secondo l'agente>*
NT Malattie parassitarie
*<secondo il modo di trasmissione>*
NT Malattie infettive
NT Malattie ereditarie
*<secondo gli organi e le parti>*
NT Malattie cerebrovascolari
NT Malattie gastrointestinali
NT Distrofia muscolare
*<secondo il paziente>*
NT Malattie infantili
RT Malati
RT Patologia
Nota sintattica: ***Parte/propr.*** *Segue il termine che rappresenta il possessore (singoli individui, gruppi di persone, organismi e loro parti), p.e.,* Leopardi, Giacomo - Malattie; Adolescenti - Malattie; Gatti - Malattie; Apparato digerente - Malattie; Bambini - Sistema nervoso - Malattie [*precedentemente* Sistema nervoso - Malattie - Infanzia]; Laringe - Vasi sanguigni – Malattie.
Categoria: Processi
Classificazione: 616 (DDC21)
Fonte1: *Soggettario 1956*
Identificativo: M000003
Status del record: Esempio
Fonte2: Gruppo

- esempio di record con nota sintattica generale (si troverà sotto termini che indicano una categoria o una classe molto generale. La sua peculiarità consisterà nel riferirsi non tanto al temine sotto il quale compare, ma a tutti i termini appartenenti a quella categoria o classe, che condividono le valenze sintattiche indicate nella nota):

**Attività**

NT Attività artistiche
NT Attività economiche
NT Attività religiose
NT Attività ricreative
NT Attività tecniche
NT Giochi
Nota sintattica generale: ***Elem. trans.*** *I termini della faccetta* Attività *seguono il termine che rappresenta l'oggetto/meta e precedono il termine che rappresenta il beneficiario, lo strumento o l'agente, p.e.* Catechismo - Insegnamento agli adolescenti; Alimenti - Conservazione - Impiego dell'acido para-ossi-benzoico; Paesi in via di sviluppo - Assistenza economica della Comunità europea.
***Parte/Propr.*** *Quando non sia citato un oggetto/meta, i termini della faccetta* Attività *seguono il termine che rappresenta il possessore (agente), p.e.* Anziani - Attività ricreative; Popoli primitivi – Agricoltura.
Categoria: Attività
Fonte1: Struttura
Identificativo: A000003
Status del record: Esempio
Fonte2: Gruppo

Approfondimenti: Alberto Cheti, Il punto sul GRIS e gli sviluppi attuali, "Bibliotime", VI, 1 (marzo 2003); Anna Lucarelli, La revisione del Soggettario, "Bibliotime", VI, 1 (marzo 2003); Anna Lucarelli, Fra principi internazionali e tradizione europea: sviluppi italiani nell'indicizzazione per soggetto a Bibliocom 2004.

## Il progetto "Opac Semantici"

Svolta fra 2003 e 2004, la ricerca "OPAC semantici" ha analizzato più di 150 cataloghi italiani dal punto di vista degli accessi semantici, allo scopo di rispondere a domande come:

- quali tecniche d'indicizzazione sono più comunemente applicate nei cataloghi in linea italiani?
- quali funzionalità proprie degli OPAC si sono aggiunte ai tradizionali percorsi della ricerca bibliografica?
- quali altre potrebbero essere arricchite da un migliore impiego delle informazioni semantiche?

I dati raccolti hanno permesso di definire un "indice di semanticità" dei cataloghi, che dovrebbe esprimere l'efficacia degli attuali cataloghi nel supportare ricerche intorno al contenuto concettuale dei documenti.

La "scoperta", non troppo inattesa, è che gli opac sfruttano solo la mimima parte delle potenzialità messe a disposizione dall'indicizzazione. Tipicamente, ad esempio, non implementano la "ricerca in due fasi" (per termine e per stringa di soggetto) auspicata da Gris e dallo studio per il nuovo Soggettario, né si può effettuare una ricerca o un browsing sugli equivalenti verbali delle notazioni di classificazione. Nell'opac del polo bolognese:

- le ricerche per soggetti e classi non sono nella maschera di default ma solo in quella avanzata

- la ricerca sulle liste è solo sull'inizio stringa e non sui singoli termini né sulle singole parole
- la ricerca nel campo "classificazioni" si può effettuare sia per notazione sia per equivalente verbale (anche se ciò non è affatto intuitivo!), ma porta sempre direttamente ai record
- una ricerca su lista del campo classificazioni può essere fatta solo sulle notazioni

Esempi a confronto su BO e FI:

- cerca "autori fiorentini" sull'opac bolognese: Lista (non trova niente) e Ricerca (trova tutto ma va direttamente ai record senza consentire ulteriori selezioni)
- cerca "autori fiorentini" sull'opac fiorentino: Browse su soggetto e Ricerca
- cerca "851.91" sull'opac bolognese: Lista (trova la lista delle notazioni che cominciano con 851.91) e Ricerca (trova tutto ma va direttamente ai record senza consentire ulteriori selezioni)
- cerca "851.91" sull'opac fiorentino: Scorri Lista Classificazioni (trova la lista delle notazioni che cominciano con 851.91) e Ricerca (trova tutto ma va direttamente ai record senza consentire ulteriori selezioni); ma ci sono anche le Ricerche sui descrittori di classificazione, con cui è possibile accedere alla Ricerca per parole contenute nella descrizione, all'Elenco delle descrizioni che iniziano con il termine specificato, al Browse sulle descrizioni.

Approfondimenti: sito del progetto Opac semantici; Claudio Gnoli, Riccardo Ridi, Giulia Visintin, Di che parla questo catalogo? Un'indagine sugli accessi semantici negli opac italiani, "Biblioteche oggi", 22 (2004), 8, p 23-29; Riccardo Ridi, Claudio Gnoli, Giulia Visintin, Come vogliamo chiamarli? Operatori booleani e tecniche di information retrieval negli opac italiani, "Bibliotime", VII, 3 (novembre 2004).

## L'indicizzazione nel contesto digitale

Il continuo aumento della informazione a testo pieno disponibile online rende sempre più pressante la necessità di disporre di procedure attraverso le quali ottenere **per via algoritmica** (cioè attraverso programmi) un risultato paragonabile a quello che otterrebbe un indicizzatore esperto attraverso la comprensione dei documenti.

Definiamo due macrocontesti di azione:

- **collezioni omogenee e conoscibili a priori**: grandi corpora testuali dotati di un livello sia pur minimo di omogeneità, ad esempio grandi collezioni di testi digitali di un certo ambito disciplinare;
- **collezioni eterogenee e a crescita imprevedibile**: il corpus complessivo di tutto ciò che è disponibile in Internet (o quantomeno tendenzialmente tale).

All'interno del primo contesto si stanno sperimentando tecniche che derivano più o meno esplicitamente dalle vecchie tecniche di indicizzazione automatizzata, integrate però dal riconoscimento della necessità di costruire una "impalcatura semantica" che superi i problemi associati alle tecniche di riconoscimento lessicale "esatto". Ne sono esempi la LSI (latent semantic indexing), una famiglia di metodi di indicizzazione e retrieval che organizzano l'informazione in uno "spazio semantico" (che utilizza i modelli di spazio vettoriale) all'interno del quale termini e documenti strettamente associati da un punto di vista concettuale sono spazialmente vicini. Al momento dell'interrogazione, attraverso l'elaborazione dei termini usati per la query si identifica un punto nello spazio semantico e si selezionano i documenti collocati nella zona circostante a quel punto. E' possibile anche implementare meccanismi di ranking basandosi sulla maggiore o minore distanza nello spazio fra documenti e query, nonché implementare tecniche di retroazione sulla rilevanza, utilizzando i termini dei documenti pertinenti selezionati per rendere più efficace la query.

L'indicizzazione (a diversi gradi di automatizzazione) e il retrieval che agiscono all'interno dell'intero

WWW pongono naturalmente problemi ancora più complessi.

Storicamente, i primi strumenti di orientamento sono state le raccolte di risorse per materia (indici manuali per classi, subject tree, per es. Yanoff, il primo Yahoo, che ha comunque tuttora come struttura principale quella classificata).

Si trattava di schemi di classificazione su base disciplinare, mantenuti da un singolo o un gruppo di persone, sulla base di un'attività "umana e manuale" di ricerca, valutazione, aggiornamento e strutturazione delle risorse in rete. Sono stati gli unici strumenti di accesso per soggetto alle risorse durante tutta la prima "pioneristica" fase di WWW, durante la quale l'entità delle risorse era tale da rendere possibile padroneggiarle con metodi manuali.

In alcuni casi non avevano un vero e proprio andamento classificatorio, ma erano una semplice "flat list" in cui erano compresenti allo stesso unico livello gerarchico sia discipline che soggetti. Es.: The WWW Virtual Library.

In una seconda fase si sono poi affermati:

- i motori di ricerca generici (cioè gli indici automatici per parole, es. Altavista, Webcrawler, Lycos, ecc.)
- i punti di partenza, i portali generici
- i metadati (dati sui dati, sistemi per la descrizione delle risorse elettroniche, derivanti da un'attività di "catalogazione" manuale, tipicamente da parte dell'autore).

Quel che sta accadendo oggi, e uno sguardo al futuro:

- affermazione di motori che fanno uso di tecniche sofisticate per la valutazione dei documenti e il ranking delle risposte: l'es. più noto è quello di Google, che "misura" l'importanza citazionale di ogni documento utilizzando un algoritmo che considera il numero di link al documento (distinguendo fra pagine repertoriali e fonti primarie); Google attribuisce un alto valore, per valutare la pertinenza di un documento rispetto ad una ricerca, anche al testo da cui parte il link al documento (anchor text) e all'ordine e alla prossimità fra i termini di ricerca
- motori basati sull'esplorazione della rete da parte di "agenti intelligenti" (cfr. vortali, vectories e harvesting)
- portali specifici, "Vortali" (portali verticali= portali prodotti e mantenuti da motori specializzati, che vengono inizializzati con una lista di siti rilevanti per un certa area disciplinare e a partire da quelli sono in grado di "catturare" gli altri rilevanti), "Vectories" (directories verticali, cioè disciplinarmente specializzate= raccolte di risorse classificate automaticamente da software specializzati che vengono inizializzati con alcuni "training documents", dai quali mutuano un primo schema di classificazione, che poi popolano attraverso il reperimento di altri documenti dello stesso ambito tematico)
- introduzione di elementi di automazione nella produzione dei metadati; Dublin Core e altri schemi di metadati;
- metadati esterni e tecniche di "harvesting" (= mietitura di siti): uso di robot specializzati nella raccolta di "indicizzazioni" (fondamentalmente i cosiddetti metadati esposti) di documenti, le cui rappresentazioni vengono depositate in maniera distribuita in luoghi e secondo formati convenzionali, in modo da essere riconosciute e catturate dai motori;
- topic maps: elaborate per descrivere strutture di conoscenza e associarle con risorse informative pertinenti, di recente diventate uno standard ISO (ISO13250 del 2000); le TMs nascono con l'obiettivo di creare un meccanismo di indicizzazione che si adatti ad ogni supporto e ad ogni oggetto o insieme di oggetti informativi (l'oggetto di riferimento non è più il singolo documento come nell'ambiente cartaceo, ma un insieme di oggetti i cui confini reciproci possono anche essere indefiniti). Esse riprendono una serie di concetti, di formalizzazioni e di funzionalità derivanti dagli strumenti "classici" di mappatura delle conoscenze, cioè indici analitici, glossari e thesauri, e li rielaborano e combinano con quelli derivanti dalle reti semantiche;
- web semantico: dall'indicizzazione alla rappresentazione della conoscenza, formalizzata e

"incapsulata" a priori nei documenti web; dal web "machine-representable" al web "machine-understandable"; dal web "deposito di informazioni" al web "fornitore di servizi".

**Conclusioni**

Approfondimenti: Metadati e indicizzazione semantica, in 3 parti, di Ingo Bogliolo; Guide per la ricerca in Internet, di Mariateresa Pesenti; DoIS.

---

A cura di S. Spinelli, ultimo aggiornamento 8.11.2005.