# PROGEST: Un sistema con interfaccia amichevole per il recupero e la gestione delle informazioni relative alle attività di ricerca del CNR. Alcuni aspetti metodologici.

*R.Amaranti* (*)- *B.Pocobelli* (*)- *R.Sprugnoli* (1)- *P.Venerosi.* (2)

(1) Dipartimento di Informatica - Università degli Studi di Firenze
(2) Istituto di Elaborazione dell'Informazione - CNR Pisa
(*) Contratti consulenza professionale CNR



## Sommario

Nella realizzazione dell'interfaccia, particolare attenzione è stata rivolta allo sviluppo di metodologie di recupero differenziate a seconda del tipo di informazione contenuta nei dati. Dal punto di vista operativo il sistema si basa, oltre che sulle funzioni proprie di un DBMS, su due tecniche complementari per il recupero semantico delle informazioni: una di tipo 'full-text' ed una che permette di scorrere un linguaggio documentario appositamente strutturato (browsing). Sono previste ricerche per l'esplorazione della base di dati lungo la struttura gerarchica del vocabolario e la rete delle associazioni semantiche. La presentazione 'a viste' del vocabolario e la visualizzazione del 'contesto corrente' (cioè dei concetti coinvolti nell'interrogazione) aiutano l'utente durante la navigazione. L'interfaccia è stata realizzata in ambiente DOS 3.30 e utilizza il dBASEIII plus arricchito da una estesa indicizzazione necessaria a realizzare l'aspetto non tradizionale dell'applicazione.

## Introduzione

PROGEST è un prototipo di interfaccia amichevole per il recupero delle informazioni relative alle attività di ricerca dei progetti finalizzati del Consiglio Nazionale delle Ricerche (CNR), realizzato in ambiente DOS nell'ambito del Progetto Strategico 'Trasferimento delle Tecnologie dei Progetti Finalizzati'.

Attualmente, la raccolta dei dati relativi all'attività di ricerca dell'Ente viene effettuata in modo centralizzato dal CNR. I dati raccolti attraverso schede di rilevazione vengono inviati al Servizio Informatica Area Milanese (SIAM) del CNR che li elabora in modo da ottenere sia una uscita on-line accessibile sulle reti nazionali ed internazionali, sia una produzione di cataloghi cartacei per la diffusione. Il sistema di gestione della banca dati dei progetti di ricerca del CNR che consente l'interscambio di informazioni a livello internazionale con i corrispondenti Enti di Ricerca di Canada, Francia, Germania, Giappone, UK, USA e Svezia è noto con il nome di EXIRPTS [NAL90].

In questo contesto, la realizzazione dell' interfaccia in ambiente DOS è un obiettivo che si pone nel senso di un'ulteriore espansione della disseminazione delle informazioni, consentendo agli utenti finali un accesso ai dati semplificato e trasparente ed un loro uso anche personalizzato.

L'ipotesi su cui si è basato il lavoro di progettazione è quella di realizzare una interfaccia centrata sulle esigenze dell'utente. Dall'esame delle relazioni esistenti tra contenuto dell'informazione presente nei dati, tipi di utente e tecniche di IR è dipesa la scelta delle tecniche e delle particolari metodologie di recupero delle informazioni sviluppate nell'interfaccia.

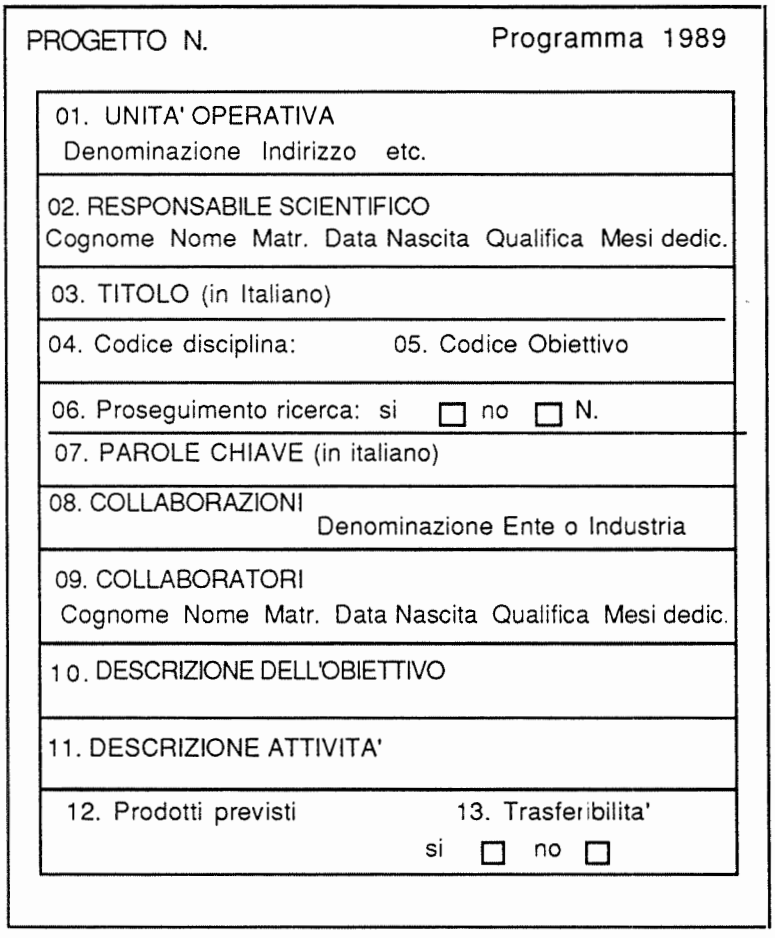


**Fig 1**: *Scheda di rilevazione progetti di ricerca*

La logica che sottostà all'applicazione è quella di dotare l'utente di uno strumento di controllo perchè la ricerca intrapresa risulti ad ogni passo significativa e di favorirlo con una gamma di funzioni che con l'alternarsi del loro impiego renda possibile una esplorazione graduale e non ripetitiva della base di dati [BAL88].

Per l'indagine dell'universo delle attività di ricerca del CNR sono stati utilizzati opuscoli illustativi dell'attività dell'Ente, le relazioni annuali degli Istituti e lavori presentati a convegni che focalizzavano aspetti interessanti per l'analisi di questa realtà [NEG90, NAL90]. La selezione dei dati utili per la descrizione delle attività di ricerca si è basata invece sui supporti che fornivano una rappresentazione strutturata del progetto di ricerca (schede di rilevazione dei progetti,vedi fig.1 e registrazioni su file).

Ai fini della sperimantazione è stata essenziale la base documentaria fornita dal file dei progetti di ricerca del 1989 messo a disposizione dal SIAM sulla quale si è svolto, in una precedente sperimentazione, buona parte del lavoro preparatorio di analisi. Questa ha consentito di individuare l'insieme di record riferibili all'area disciplinare della Computer Science, prescelta come base di conoscenze nei sistemi di interfaccia da sviluppare. Inoltre ha reso possibile sperimentare il funzionamento dei linguaggi documentari proposti dal CNR e usati dai ricercatori ai fini di un esame del sistema di interrogazione disponibile e quindi della qualità della risposta. L'esito di questa indagine ha rafforzato la convinzione dell'utilità, come verrà detto più avanti, di inserire nel sistema di indicizzazione delle ricerche i livelli più generali di un linguaggio specialistico autorevole che consentisse modalità di interrogazione più articolate. Questo ampliamento dell'area descrittiva con l'aggiunta dei due campi necessari ad ospitare i nuovi livelli di codifica ha comportato l'indicizzazione manuale dei record dell'area informatica.

Su questa base, in mancanza di formati di comunicazione standard per questo tipo di dati, si è ricorsi ad un programma ad hoc, in ambiente VM, che operasse sul file del SIAM per la selezione dei record e dei campi utili alla creazione della base documentaria dell'interfaccia, provvedendo anche all'operazione di conversione del file in uno importabile nel sistema in sviluppo [BIA90].

La possibilità di poter usufruire immediatamente della gamma di dati ritenuti necessari e sufficienti dal CNR ha consentito di puntare direttamente alla ricognizione del tipo di informazione in essi contenuta e a fissarne la tipologia per distinguere le possibili modalità di accesso.

## 1. Modello di dati e modello dell'utenza

Nel corso dell'indagine sono stati individuati tre tipi di dati utili per la definizione del modello [AGO87]: i dati strutturati (ad es. i nomi dei ricercatori che collaborano al progetto) per i quali vale la regola di recupero eatto o 'exact matching', cioè recupero deterministico tipicamente utilizzato dai DBMS; i dati tipo testo utilizzati per la descrizione della ricerca (ad es. il titolo della ricerca) ed i dati cosiddetti ausiliari che hanno un significato convenzionale diretto al recupero semantico dell'informazione (ad es. i termini di un vocabolario controllato). Un'ulteriore distinzione si è resa necessaria per i dati strutturati, tra quelli direttamente accessibili e quelli che potevano essere recuperati come testo in quanto intrinsecamente non significativi per una interrogazione. Ottenuta la tipologia, le modalità di accesso sono dipese dall'esame della potenziale utenza, dalle sue esigenze e dalla efficacia della sua rappresentazione.

Nell'analisi dell'utenza è stata individuata una precisa figura professionale, quella dell'amministratore della base di dati con funzioni di gestione e manutenzione. Queste funzioni, per la loro specificità, comportano un alto grado di competenza e responsabilità operativa ed è per questo motivo che si è deciso di permetterne l'uso solo ad un utente specializzato nel settore.

Nel corso di una ulteriore indagine sono stati individuati i potenziali utenti nelle figure dei tecnici, ricercatori, imprenditori, amministrativi. In mancanza di una vera e propria sperimentazione di laboratorio che consentisse di analizzare il comportamento di questo tipo di utenza abbiamo ipotizzato, sulla base dell'esperienza, che, in linea generale, esistessero per ognuna di queste figure degli interessi preferenziali verso un certo tipo di informazione presente nella base documentaria, ma che nessuna di queste figure potesse essere veramente distinta sulla base di univoci interessi. Si è preferito quindi parlare di 'approccio flessibile' per esprimere la realtà delle diverse esigenze informative che, nel tempo, uno stesso utente può avere e le diverse preferenze che utenti dello stesso tipo hanno nei confronti di differenti modalità di recupero dell'informazione.

La precisazione delle esigenze professionali segna infatti una prima approssimazione al modello di utenza; ma è l'interpretazione delle flessibilità del comportamento e delle

esigenze dell'utente e la capacità di farle aderire alle funzioni di IR che determina l'efficienza del modello. Si puo' dire che un modello di utenza è realizzato solo quando la trasparenza del funzionamento di un sistema lo fa sentire all'utente come suo proprio [MON84].

Quindi, in conformità con l'ipotesi dell'esistenza di un approccio preferenziale dell'utente ad un certo tipo di informazione abbiamo distinto quello diretto alla gestione, quello diretto alla ricerca di dati strutturati e quello diretto al recupero semantico delle informazioni. In secondo luogo, per garantire il principio della flessibilità abbiamo evidenziato in ogni approccio alla ricerca il significato della gamma delle opzioni disponibili ed agevolato il passaggio tra un approccio e l'altro. In questo ambito si è poi tenuto conto della distinzione tra utente casuale ed esperto garantendo per certi tipi di indagine livelli operativi di maggiore o minore difficoltà.

## 2. Modello di IR e schema dell'interfaccia

Con questa indagine preliminare si sono individuate le tre funzioni principali da attuare e rendere disponibili con l'interfaccia: il recupero dell'informazione diretto ai dati strutturati, quello diretto al contenuto semantico dell'informazione ed infine la gestione (fig.2).

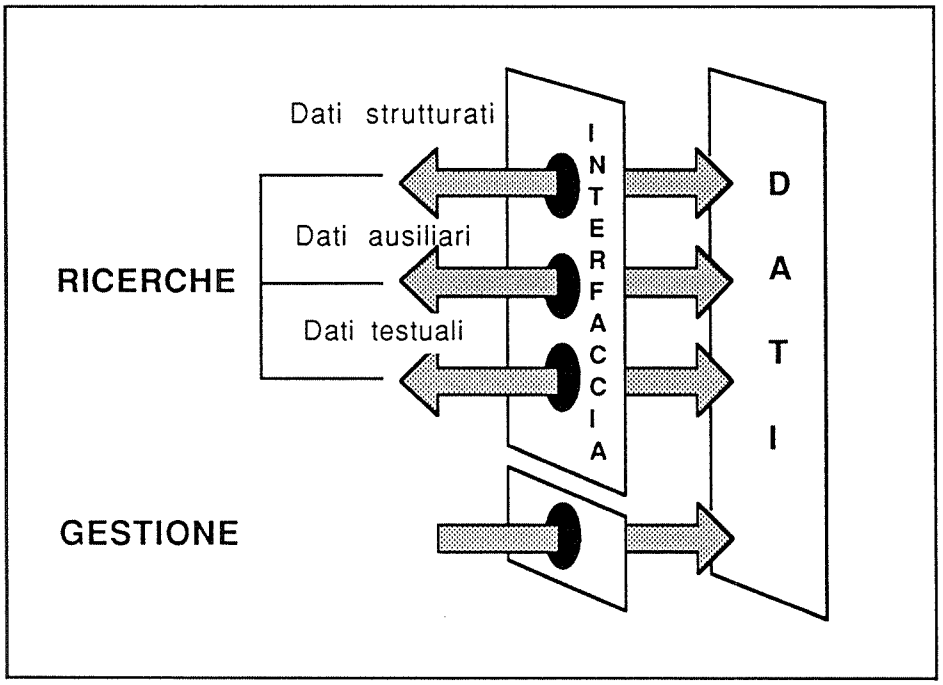


**Fig 2**: *Schema di interfaccia*

L'ipotesi del recupero semantico delle informazioni, formulata in base all'indagine sui dati e sull'utenza, prevede l'uso di due tecniche di IR complementari [SAL86], una di tipo full-text ed un'altra che permette di scorrere la struttura di un linguaggio documentario per l'esplorazione della base di dati. Le due modalità di accesso sono state realizzate in modo da soddisfare il criterio della flessibilità dell'approccio.

Nella ricerca 'full-text' la base conoscitiva è rappresentata da una descrizione del contenuto informativo dei progetti in linguaggio naturale. Si è convenuto che i campi più rappresentativi in questo tipo di descrizione fossero quelli relativi al titolo, all'obiettivo ed ai termini liberi. A partire da questi punti di accesso, l'utente interroga il sistema esprimendo le domande in linguaggio naturale.

Per facilitare l'utente sono previsti tre diversi tipi di approccio alla ricerca.

Con il primo, denominato 'parla di', si dà la possibilità di effettuare una ricerca sul contenuto semantico del progetto. Questo tipo di interrogazione, pensato per una utenza non specialistica, dà una possibilità di ricerca anche all'utente che non conosce il sistema e gli permette di esprimere la richiesta con le proprie parole. In questo caso si può parlare di una autentica interrogazione in linguaggio naturale perchè non si fa ricorso a linguaggi che solo esteriormente sono naturali ma che di fatto sono o frutto di una manipolazione basata sull'impiego di operatori booleani o un insieme di termini o espressioni utilizzati per la rappresentazione dei documenti.

Il tipo di interrogazione previsto nel secondo approccio alla ricerca si può definire strutturato in quanto si compone di due passaggi: la selezione del campo da indagare e la formulazione dell'argomento con l'impiego di operatori booleani. Questo tipo di ricerca più complessa della precedente è utile nei casi in cui l'utente ricorda in maniera approssimativa qualche parola letta in un campo testuale. Diversamente dal primo caso, il sistema effettua confronti parziali sui dati solo con il consenso dell'utente. Grazie a quest'ultima possibilità e alla maggiore specificità della richiesta, il grado di precisione ottenibile è più alto rispetto al primo caso.

Il terzo approccio, consigliato ad utenti che abbiano familiarità con linguaggi documentari, utilizza i termini liberi che i ricercatori hanno assegnato alle loro ricerche.

L'uso di questo strumento convenzionale permette all'utente di raffinare le proprie ricerche in un ambiente già selezionato dall'interrogazione eseguita sul descrittore prescelto. Si ha qui il più alto grado di precisione ottenibile con questo sistema perchè i documenti recuperati sono tutti rilevanti grazie alla seconda selezione effettuata.

Le modalità della ricerca sopradescritta non sono inquadrabili in una strategia; l'utente si serve delle possibilità indicate dal menù e sceglie quella che meglio si attaglia alle esigenze del momento. Le difficoltà che incontra nella ricerca non sono riferibili alla scansione di livelli operativi più o meno complicati, ma sono piuttosto di ordine psicologico dovute all'assoluta libertà in cui si trova quando formula l'argomento da indagare oppure all'inesperienza nell'uso dei connettivi logici.

La ricerca guidata da un linguaggio documentario ha caratteristiche molto diverse perchè consente di allargare o raffinare la ricerca attraverso una opportuna strategia di esplorazione basata sulla struttura concettuale del linguaggio.

Nel nostro caso l'inadeguatezza del sistema di indicizzazione usato nelle schede di compilazione, inadatto ad impostare interrogazioni a più livelli, non consentiva questo tipo di ricerca. Si è quindi introdotto nel sistema di indicizzazione un linguaggio controllato autorevole quale elemento di normalizzazione e strutturazione del vocabolario dei termini liberi nella convinzione che l'organizzazione terminologica fosse essenziale anche ad una buona rappresentazione grafica [TOD82] e che questa fosse di ausilio all'utente. Il sistema di indicizzazione adottato è composto dalle categorie generali e da quelle sottostanti di livello più specifico dello schema di classificazione dell'ACM Computing Reviews, suggerito come utile strumento di rappresentazione e di interrogazione per basi documentarie in ambito informatico [BAL77].

Con questo inserimento, che ha comportato l'indicizzazione manuale di un campione di record, il vocabolario dei termini liberi, terzo livello nella struttura del linguaggio, si è ordinato automaticamente in sottoinsiemi associati alle varie categorie dello schema, creando il presupposto per una precisa strategia di ricerca. Questa strategia permette di scorrere la struttura gerarchica del vocabolario ed il reticolo semantico che si genera con l'assegnazione dei descrittori da parte dei ricercatori [VEN90]. Devono ora essere

specificate le modalità di gestione della ricerca, garantiti i margini di autonomia dell'utente nel controllo dell'indagine ed introdotti gli accorgimenti per favorirlo durante la navigazione.

Si tratta di combinare una fase di ricerca 'guidata' ad una che consenta di allargare la sfera di autonomia dell'utente. La prima si basa sulla rappresentazione gerarchica della conoscenza presente nella base di dati, ed ha inizio con la visualizzazione del vocabolario controllato che evidenzia sia la base conoscitiva sia la divisione concettuale delle aree di interesse informatico. Da questa base di partenza, adatta anche ad un utente casuale, il percorso è segnato dalla struttura ad albero del vocabolario che indirizza l'indagine dall'alto verso il basso lungo i tre livelli della sua articolazione, fino a raggiungere i termini liberi che rappresentano lo strumento di interazione più efficace con la base referenziale. In ogni punto di qualsiasi livello è possibile prendere visione dei termini del livello successivo ad esso connessi ed esaminare la porzione della base di dati referenziale che ad essi si riferisce. La possibilità di ritornare al vertice dell'area di interesse indagata e la visibilità dei nodi che vi sono collegati consente all'utente di cambiare percorso.

Maggiore autonomia di conduzione della ricerca si riscontra lungo il reticolo delle associazioni semantiche tra descrittori. La conoscenza incorporata nella base di dati è, in questo caso, rappresentata dal reticolo in cui i nodi rappresentano i termini generali, quelli più specifici e quelli liberi, e gli archi le connessioni tra essi. Il legame tra le categorie concettuali o i termini liberi non rispetta qui nessun vincolo schematico, ma indica solo la condivisione di qualche aspetto semantico generato automaticamente dall'assegnazione dei descrittori nella fase di rappresentazione dei progetti. Il percorso trasversale nel reticolo consente il passaggio da un ambito informatico ad un altro in modo trasparente e dipende unicamente dalla scelta che l'utente compie tra le categorie o descrittori che incontra sul percorso. In ogni punto infatti si potranno produrre 'viste' su insiemi di descrittori associati a quello interrogato e allargare o restringere la ricerca [VEN86]. Questo effetto 'a viste', che da una parte interrompe la presentazione lineare del vocabolario e dall'altra favorisce la prosecuzione della ricerca, utilizza la tecnica del 'rank' [BAL83] e può essere usato per ottenere una gamma di associazioni gerarchiche o

reticolari: categorie di II livello ACM associate al I livello per pertinenza, termini liberi associati ad una precisa categoria, termini liberi associati tra loro perchè condividono un qualche aspetto semantico con il termine indagato.

Altro accorgimento, introdotto per facilitare l'utente in ogni fase della ricerca, è quello della visibilità del 'contesto corrente' espresso dalle categorie concettuali (del I e II livello ACM) coinvolte nell'interrogazione. La conoscenza dell'ambito concettuale nel quale l'utente si sta muovendo evita la perdita di controllo dell'indagine e offre l'occasione per puntare in direzioni non usuali o perlomeno non previste.

Un disegno di IR di questo tipo, impostato su modelli di dati funzionali a precise tecniche di recupero, ha comportato una struttura di interfaccia articolata in moduli. A ciascuno di questi corrisponde un archivio virtuale al quale si accede da un menù modellato sul tipo di approccio.

## 3. Scelta del sistema di gestione

Il procedimento seguito nella progettazione del sistema ha coinvolto innanzitutto la modellizzazione dei dati in conformità alla diversità di gestione esistente tra un DBMS ed un IRS, illustrata efficacemente nel diagramma [AGO90] esposto nella fig.3.

Mentre la funzione di recupero deterministico dei dati può essere svolta con le funzioni base di qualsiasi DBMS, la parte relativa all'IR richiede una trattazione separata mirata allo sviluppo di questa applicazione non tradizionale nei DBMS.

Si è operata quindi una integrazione tra il modello di dati utili per il recupero semantico delle informazioni e quello per il recupero dei dati strutturati onde realizzare il modello di dati complessivo e lo schema concettuale. Dall'esito di questo procedimento sono dipese la scelta del sistema di gestione e le modalità di esecuzione.

Questa non è stata nè semplice nè immediata per la difficoltà di reperire tra i sistemi presenti sul mercato quelli che permettono la descrizione del contenuto semantico di un documento mediante un sistema di indicizzazione gerarchizzato e quindi la possibilità di effettuare le ricerche navigando lungo la struttura del linguaggio documentario. Infatti, da un'analisi preliminare dei prodotti presi in esame, risultava da un lato che i DBMS

disponibili erano in uno stadio di ingegnerizzazione avanzata, ma non consentivano applicazioni non tradizionali, e che dall'altro lato i sistemi di IRS disponibili non erano altrettanto consolidati sul mercato o non offrivano le funzioni necessarie a questo tipo di applicazione e la possibilità di programmarle.

| | Gestione dati | Information Retrieval |
|---|---|---|
| 1 - Modalità di risposta del sistema | | |
| 1a - Modo in cui il sistema risponde ad una frase di richiesta di informazioni | Mediante un confronto esatto | Mediante un confronto parziale o il confronto migliore che riesce a realizzare |
| 1b - Unità ricercate | Che soddisfano il confronto esatto | Unità pertinenti la richiesta informazioni |
| 1c - Specificazione della richiesta di informazioni | Completa | Incompleta |
| 2 - Modello di recupero dati sul quale si fonda il sistema | | |
| 2a - Modello di descrizione | Deterministico | Probabilistico |
| 2b - Classificazione | Monotematica | Politematica |
| 3 - Criterio del successo: correttezza o utilità | | |
| Errore di risposta | Sensibile | Non sensibile |
| 4 - Linguaggio di interrogazione e interfaccia fra utente e sistema | | |
| Linguaggio di interrogazione | Artificiale | Naturale (si tende a) |

*Tabella: Diversità fra l'information retrieval e la gestione dati*

Fig. 3: *Diversità tra l'IR e la gestione dati.*

Si è preso quindi in considerazione la possibilità di utilizzare la tecnologia standard di un RDBMS (in questo caso il dBase III plus) con i vantaggi che un sistema relazionale consolidato comporta e di realizzare una applicazione per ottenere quella 'estesa indicizzazione' necessaria a risolvere l'aspetto non tradizionale della applicazione. L'intento di accrescere la capacità espressiva del dBase III plus con la definizione di nuovi tipi e modalità di accesso ha comportato un notevole sforzo di programmazione.

## 4. Schema concettuale

Una volta identificati i requisiti informativi della applicazione, si è passati alla costruzione di un modello globale astratto. Lo schema concettuale descrive, attraverso un linguaggio formale, l'interazione delle diverse realtà coinvolte nell'applicazione,

consentendone una visione astratta. Il linguaggio formale che si è utilizzato è l'Entity Relationship Model (ER) il quale presenta una estrema semplicità di rappresentazione delle entità e relative associazioni ed una effettiva potenza semantica ottenuta grazie ad una consolidata tecnica diagrammatica estesa anche per la rappresentazione di dati tipo testo [AGO89].

In questa ottica i tre principali tipi di dato individuati sono stati suddivisi in entità rappresentanti le minime unità concettuali per ciascun tipo di dato [AMA91]. I dati strutturati sono stati scissi nelle entità **Unità Operativa** e **Sede Contraente** che contengono rispettivamente informazioni relative all'Istituto che effettua la ricerca e a quello che la commissiona, e nelle entità **Collaborazioni** e **Collaboratori** che contengono rispettivamente i dati relativi agli enti ed alle persone che collaborano alla realizzazione del progetto di ricerca. Tali entità, come è chiaramente comprensibile, rappresentano realtà distinte il cui unico punto di unione è dato dal progetto di ricerca di cui tutte fanno parte. E' stata così introdotta una entità centrale **Progetto** che attraverso il valore della chiave primaria "numero del progetto" consente di porre in relazione tra loro tutte le entità dello schema concettuale.

Una particolare attenzione è stata rivolta ai dati di tipo testuale sui quali è attivata una ricerca di tipo full-text. In questo ambito, oltre alle entità **Titolo**, **Obiettivo** e **Parole Chiave**, che rappresentano la parte testuale di un progetto, è stata creata una entità **Stop-List** contenente una lista di termini (quali preposizioni, articoli, avverbi, ecc.) che non hanno alcuna rilevanza semantica ai fini della ricerca e che quindi possono essere automaticamente esclusi qualora contenuti nella richiesta dell'utente. Inoltre sono state individuate delle entità rappresentanti gli indici (a liste invertite) per permettere il recupero dei singoli termini contenuti nei testi [SPR87]. Tali entità sono legate per mezzo di una associazione di tipo N:M con i testi contenenti le parole degli indici.

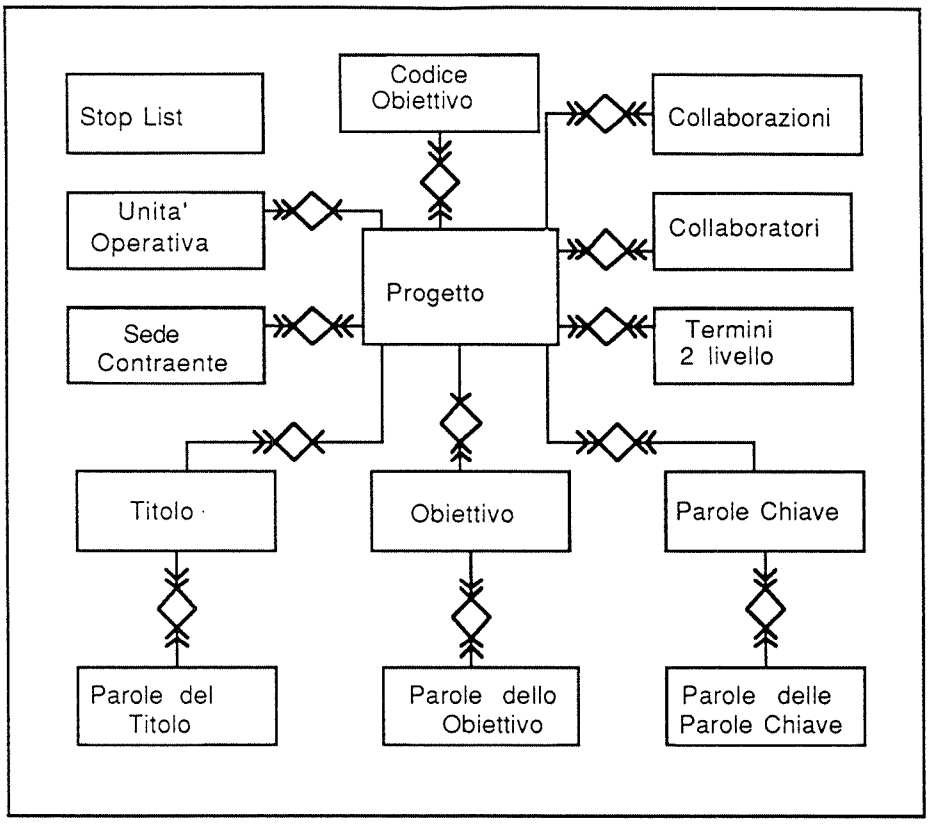


**Fig 4**: *Schema concettuale*

Per favorire il recupero semantico dell'informazione, sono stati introdotti i dati ausiliari rappresentati da entità contenenti i 'subject headings' dell'ACM ed i codici di classificazione di un vocabolario controllato che identificano gli obiettivi dei progetti.

Il passo successivo è consistito nella progettazione fisica della struttura dei singoli archivi e delle procedure che permettono una loro interazione ed interrogazione.

## 5. Ricerca di un documento

Una delle principali caratteristiche di PROGEST [AMA91] è la semplicità di funzionamento della sua interfaccia.

Essa riunisce in sè una duplice funzione di guida nei confronti dell'utente:

a) da un lato, grazie ad un "help" in linea, evita la consultazione del manuale per l'uso dei comandi accessibili ad ogni passo;

b) dall'altro, fornisce una ricerca di tipo interattivo che guida l'utente in una esplorazione graduale e non ripetitiva della base di dati in modo da permettere una migliore definizione delle sue necessità e favorire così il recupero solo dei documenti effettivamente interessanti.

Una prerogativa di tutti i menù dell'interfaccia è data dalla possibilità di:

1) Ritornare al menù precedente;

2) Ritornare al primo menù;

3) Selezionare le opzioni mediante i tasti freccia ed ENTER.

Il primo menù (fig.5) che viene presentato all'utente permette la scelta fra tre ambienti, contraddistinti dal tipo di ricerca che vi si effettua; ciascuno di questi opera su dati omogenei per contenuto di informazione:

- RICERCA SUL CONTENUTO: tipica degli IRS;
- RICERCA SUI DATI STRUTTURATI: tipica dei DBMS.;
- GESTORE BASI DI DATI: funzionalità di gestione e manutenzione della banca dati.

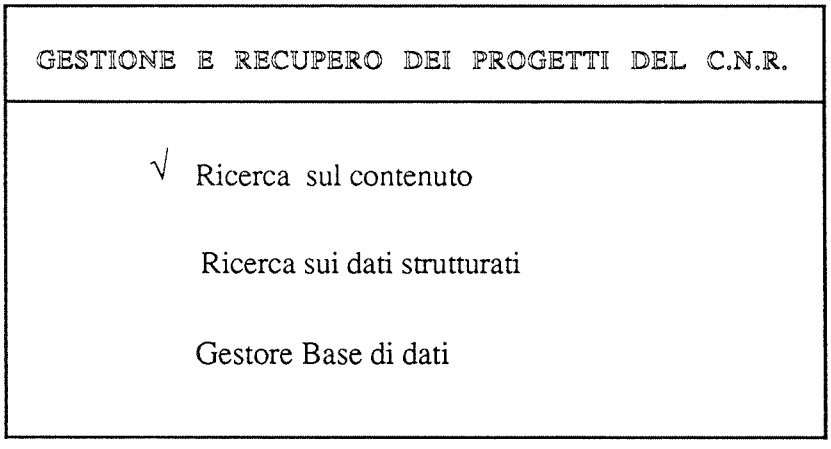


**Fig 5**: *La finestra di presentazione del sistema*

Le funzionalità di gestione e manutenzione della banca dati sono state isolate dal nucleo del sistema (costituito dalle operazioni di ricerca) tenendo conto della loro specificità e particolarità di realizzazione.

Selezionando dal menù di fig.5 la voce **Ricerca sul contenuto** entriamo nell'ambiente più propriamente di IR.

Esso, come si può vedere dal menù in fig.6, permette:

- **Ricerche di tipo full-text** (ottenibili dalla selezione di RICERCHE A PARTIRE DALLE PAROLE o RICERCHE SU CAMPI SPECIFICI)
- **Ricerche mediante uno schema di classificazione** (ottenibili dalla selezione di RICERCHE A PARTIRE DALLO SCHEMA).

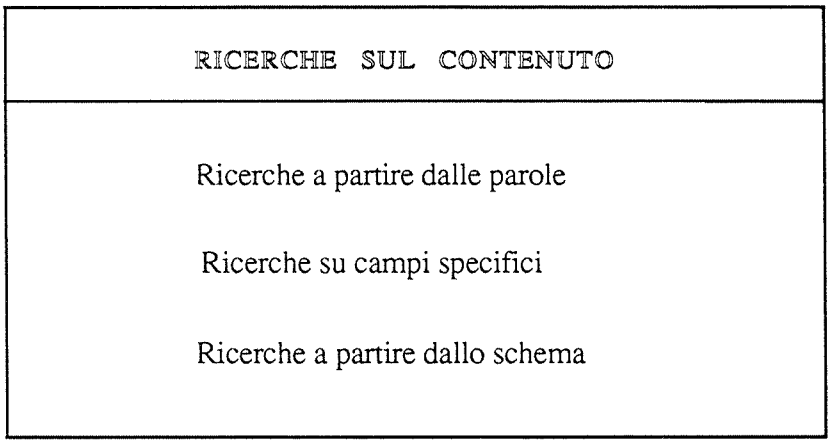


**Fig 6**: *La finestra per la scelta del tipo di ricerca testuale*

A) Ricerche di tipo "full text": in questo tipo di ricerca l'utente può interrogare il sistema direttamente con il proprio linguaggio al fine di esprimere al meglio le proprie necessità. Il successo della ricerca dipende in gran parte dalla capacità di pensare in quale modo un particolare concetto può essere stato espresso nella base di dati.

La richiesta posta al sistema consiste in un insieme di frasi che descrivono il problema senza alcun tipo di restrizione sul vocabolario; infatti sarà il sistema che si farà carico di scartare i termini non significativi grazie alla presenza di una lista aggiornabile di "stop words".

Per facilitare l'utente abbiamo individuato e realizzato diversi tipi di approccio alla ricerca di tipo "full text":

A.1) ricerca sul contenuto semantico del progetto denominata *parla di* (ottenibile selezionando dal menu di fig.6: RICERCHE SUL CONTENUTO) che viene resa disponibile quando l'utente conosce solo l'argomento sul quale vuole impostare l'indagine, ma non conosce la struttura interna dei dati su cui opera. In questo caso è il sistema a farsi carico del recupero dei documenti che trattano quell'argomento in almeno uno dei campi testuali.

Il tipo di interrogazione che viene fornito in questo ambiente è il più semplice tra quelli permessi (infatti nella formulazione della query non devono essere specificati operatori di tipo booleano) poichè è orientata verso un tipo di utenza non specialistica. L'utente è libero di digitare frasi complete, parole, o parti di parole.

Il sistema è programmato in modo da poter effettuare dei confronti parziali sui dati in maniera trasparente all'utente, al fine di non recuperare solo le informazioni direttamente richieste, ma fornire anche una serie di riferimenti a documenti che potrebbero contenere ciò che l'utente sta cercando.

Si osservi che il grado di precisione ottenibile con questo tipo di ricerca è in generale basso, cioè sono pochi i documenti rilevanti fra quelli recuperati.

A.2) ricerca realizzata sui campi testuali **titolo**, **obiettivo** e **parole chiave**. Questo tipo di ricerca si ottiene selezionando la voce RICERCHE SU CAMPI SPECIFICI dal menù di fig.6 ottenendo così il menù di fig.7.

In questo caso, a differenza del precedente "parla di", viene specificato, oltre all'argomento della ricerca, anche il campo in cui esso deve essere contenuto. Il tipo di interrogazione che viene fornito in questo ambiente è più complesso rispetto a quello precedente in quanto nella formulazione della query possono essere specificati operatori di tipo booleano.

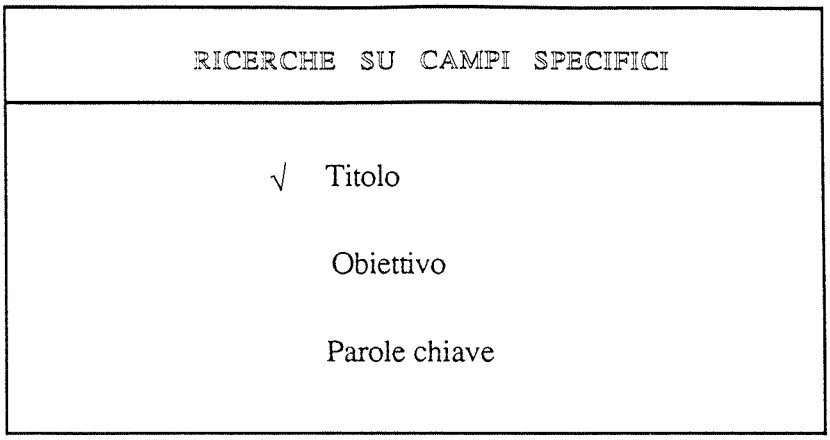


**Fig 7**: *La finestra per la selezione dei campi*

Scegliendo l'opzione TITOLO o OBIETTIVO il tipo di ricerca che viene realizzato è analogo, e viene visualizzata, nel caso del TITOLO, la seguente maschera per l'interrogazione.

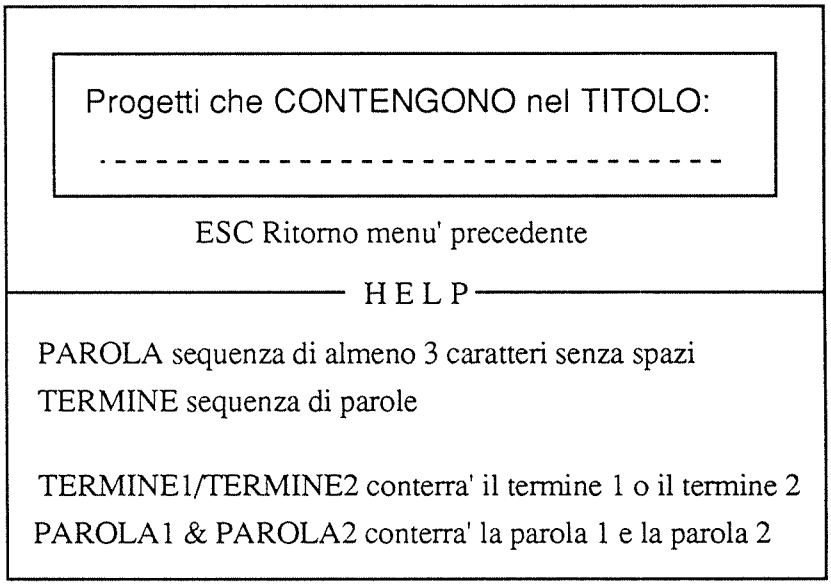


**Fig 8**: *La finestra per la formulazione dell'interrogazione*

A.3) Scegliendo l'opzione PAROLE CHIAVE dal menù di fig.7, appare uno schermo simile a quello di fig.8 e si attiva la **ricerca sui termini liberi**. Il tipo di interrogazione è analoga a quella vista nel primo caso anche se viene richiesto il campo su cui si deve realizzare la ricerca.

Inizialmente si effettua la ricerca parziale sui termini in modo da recuperare un insieme di documenti, e successivamente viene data all'utente la possibilità di scegliere, fra i dati recuperati, quelli che corrispondono meglio alle sue esigenze informative, restringendo quindi l'insieme dei progetti reperiti.

B) ricerche guidate da un linguaggio di indicizzazione: con questo strumento, qualsiasi utente, anche occasionale, ha a disposizione un sistema di informazioni utili per iniziare e definire una strategia di ricerca.

Per poter effettuare questo tipo di ricerca è necessario selezionare la voce "RICERCHE A PARTIRE DALLO SCHEMA" dal menù di fig.6.

Quando il modulo viene attivato l'utente prende visione delle categorie del I livello dell'ACM che può interrogare. Dopo questo passo può continuare la navigazione lungo l'albero passando ai termini del secondo livello associati alla categoria selezionata.

Una volta visualizzati i termini del secondo livello, interroga il sistema selezionando uno di questi termini, oppure prende visione dei termini liberi associati al termine selezionato e quindi nel secondo caso:

- procede all'interrogazione sul termine libero
- visualizza le categorie del secondo livello associate al termine indagato e le scorre per effettuare nuove interrogazioni.

Da ciascuno dei punti sopra descritti è possibile risalire nell'albero la cui radice è il termine del primo livello selezionato e le foglie sono rappresentate dai termini liberi.

Per evitare che l'utente "si perda" durante una indagine, è stata fornita la possibilità di visualizzare il *contesto corrente* relativo all'argomento oggetto dell'interrogazione, ovvero di vedere l'insieme delle scelte effettuate sia sui termini dell'ACM, sia sui termini liberi (il cammino attuale lungo l'albero di ricerca); tale contesto è rappresentato mediante le due categorie ed i termini liberi coinvolti nell'interrogazione.

Inizialmente il contesto corrente contiene l'universo dei progetti, non essendo ancora stato effettuato nessun tipo di selezione.

Il primo sottoinsieme dell'universo dei progetti viene individuato nel passaggio dai termini del I ai termini del II livello dell'ACM restringendo il campo di interesse ai progetti classificati con un particolare termine del primo livello.

Ad esempio, supponendo di scegliere "Computing methodologies" come termine del primo livello il contesto corrente sarà:

TERMINE DEL 1° LIVELLO --> COMPUTING METHODOLOGIES
TERMINE DEL 2° LIVELLO --> NON SELEZIONATO
PAROLE CHIAVE --> NON SELEZIONATE

Tale contesto continua ad essere valido fino a quando non verrà selezionato un diverso termine del I livello.

Il secondo sottoinsieme dei progetti viene individuato selezionando un termine del secondo livello ed attivando il passaggio dai termini del II livello dell'ACM ai termini liberi.

Ad esempio supponendo di scegliere "Artificial intelligence" come termine del secondo livello il nuovo contesto corrente sarà:

TERMINE DEL 1° LIVELLO --> COMPUTING METHODOLOGIES
TERMINE DEL 2° LIVELLO --> ARTIFICIAL INTELLIGENCE
PAROLE CHIAVE --> NON SELEZIONATE

Tale contesto continua ad essere valido fino a quando non verrà selezionato un diverso termine del II livello.

Il terzo ed ultimo sottoinsieme dei progetti viene individuato con la selezione di uno dei termini liberi. Supponendo di scegliere "Logica" avremo:

TERMINE DEL 1° LIVELLO --> COMPUTING METHODOLOGIES
TERMINE DEL 2° LIVELLO --> ARTIFICIAL INTELLIGENCE
PAROLE CHIAVE --> LOGICA

Tale contesto continua ad essere valido fino a quando non verrà selezionato un diverso termine libero o si percorrerà un nuovo cammino lungo l'albero di ricerca.

Il *contesto corrente* rappresenta inoltre un potente strumento di ricerca in quanto permette di specificare ulteriormente la richiesta restringendo l'insieme dei documenti recuperati al passo precedente.

C)Ricerca su dati strutturati (amministrativi): dal menù di fig.5 entriamo nell'ambiente delle ricerche su dati strutturati. In questo caso, i progetti sono recuperati solo in base ad un confronto esatto degli attributi specificati nella richiesta.

Gli attributi sui quali può essere attivata la ricerca sono presentati all'utente su tre pagine che possono essere scorse ciclicamente con i tasti PG-UP e PG-DWN.

Per poter effettuare una ricerca è sufficiente inserire nel campo interessato i dati della interrogazione ed attivare la ricerca premendo CTRL-END.

Vengono così visualizzati, a partire dalla pagina contenente il campo su cui si è fatta la ricerca, i progetti trovati. Al termine, è tenuta memoria delle richieste specificate precedentemente al fine di poter realizzare una "selezione in cascata" restringendo l'insieme dei documenti recuperati. Con la definizione "ricerca in cascata" si intende infatti un tipo di ricerca, a partire da un insieme di documenti recuperati in una precedente interrogazione, in cui si specificano i valori di attributi non ancora considerati.

La **visualizzazione dei progetti recuperati** avviene a partire dalla pagina contenente il titolo o l'obiettivo a seconda del campo su cui è stata effettuata la ricerca.

Per ottenere una buona leggibilità del contenuto di un progetto, lo si è diviso su più pagine. La prima ad essere presentata all'utente è quella relativa agli attributi sui quali è stata fatta l'interrogazione. Nella stessa pagina viene mostrato il numero progressivo del progetto sulla totalità dei progetti reperiti.

Due tipi di scorrimento sono possibili:

- uno permette la visualizzazione di tutte le informazioni relative ad un singolo progetto recuperato; tali informazioni sono suddivise in quindici pagine video numerate progressivamente ed a scorrimento ciclico: dopo la quindicesima è visualizzata di nuovo la prima;
- l'altro consente la visualizzazione della pagina di ingresso di tutti i progetti recuperati con una determinata interrogazione.

Ogni pagina che appare sullo schermo è concettualmente divisa in tre parti: sulla prima riga dello schermo appare il numero progressivo di pagina e quello che identifica il progetto esaminato, nella parte centrale i dati e sull'ultima riga la lista di opzioni operanti in quella pagina. I comandi sono scritti per esteso nell'ultima riga dello schermo e la loro escuzione avviene mediante la pressione della lettera iniziale di ciascuno. Inoltre, i dati relativi al progetto correntemente visualizzato possono essere stampati su carta in un formato scheda.

## 6. Conclusioni

In questo lavoro sono stati introdotti e risolti i problemi connessi alla creazione di un ambiente specifico di IR integrato con quello tipico della gestione dei dati e sono state messe in evidenza le caratteristiche delle tecniche di interazione per operare in questo ambiente. In particolare si è voluto evidenziare un approccio simbolico che consenta all'utente di esprimere direttamente i concetti tramite le parole chiave. L'ambiente a indici cosi' realizzato, oltre a favorire una maggior velocità di recupero, permette l'uso di tecniche per la produzione di una vasta gamma di rappresentazioni visive e quindi di strategie di recupero. La generazione di display ottenuta a partire da punti di vista diversi conferisce loro un alto grado di significatività mentre la possibilità di manipolazione diretta facilita la formulazione delle query ed i meccanismi di orientamento.

## 7. Estensioni previste

E' in corso di realizzazione l'estensione, in ambito multidisciplinare, delle funzionalità di recupero basate sull'impiego della struttura di uno schema poligerarchico. Lo scopo di questo schema, in via di preparazione [NEG90], è quello di stabilire, mediante una correlazione tra discipline diverse, una gerarchia di classi e realizzare dei legami di coordinazione e subordinazione di elementi in base a criteri diversi, individuando le cosiddette 'aree comuni'. In questo senso esso rappresenta uno strumento efficace per descrivere la pluridisciplinarietà di un progetto di ricerca, permettendo di osservarlo da diversi punti di vista. L'applicazione in corso prevede l'adozione dei criteri già sperimentati per l'uso dello schema ACM ed in particolare sono già operanti e manipolabili

i display delle categorie di secondo livello dello schema multidisciplinare associate in modo precoordinato a quelle della disciplina sulla quale si sta effettuando l'interrogazione.

Altre estensioni sono il lavoro di completamento e sperimentazione della interfaccia utilizzando un sistema ipertestuale in grado di fornire una maggiore espressività, l'implementazione di un esempio di uso personalizzato del sistema e lo studio per la verifica della portabilità su altri sistemi IR.

## Riferimenti


[ACM90] Guide to Computing Literature 1990. ACM, 1990.

[AGO87] Agosti. M."Evoluzione dei sistemi di I.R." Sistemi e Automazione, n°277, 1987.

[AGO89] Agosti M., Crestani F., Gradenigo G.. "Towards data modelling in I.R." Journal of Information Science Principles & Practice, vol.15, n°6, pp.307-319,1989.

[AGO90] Agosti.M. "Interrogazione e valutazione del recupero delle informazioni." Informatica Oggi, n°58, 1990.

[AMA91] Amaranti R., Pocobelli B. "PROGEST: Un sistema con interfaccia amichevole per il recupero delle informazioni relative ai progetti di ricerca del CNR." Tesi di laurea in Scienze dell'informazione Università degli Studi di Pisa A.A. 1989/1990.

[BAL77] Baldacci M.B., Sprugnoli R.."Recupero dell'informazione blibliografica: un'interfaccia utente-sistema per la definizione delle strategie di ricerca.".pp.129-130. Pisa. Congresso annuale AICA, Pisa, 1977.

[BAL83] Baldacci M.B., Parise M.C. Nardelli A.M.. "Un'interfaccia per l'esplorazione delle conoscenze in un sistema di recupero delle informazioni." IEI CNR, Nota interna B83-20, Dicembre 1983.

[BAL88] Baldacci.M.B. "Rappresentazione e ricerca delle informazioni." La Nuova Italia Scientifica, 1988.

[BIA90] Biagioni S.. "Organizzazione e comunicazione di dati catalografici (un'esperienza)." Convegno 'Linguaggi documentari e Basi di dati'. pp.503-505. CNR, Roma, 3-4 Dicembre 1990.

[CRO87] Croft W.B., Thompson. R.H. "IR: A new approach to the design of document_retrieval systems." Journal of the American Society for Information Science, 38(6),pp.389-404.1987.

[MON84] Monk A. "Fundamentals of Humans Computer Interaction". Academic Press, 1984.

[NAL90] Naldi F.,Carrara P.,Vannini Parenti I. "Le banche dati dei progetti di ricerca C.N.R." Convegno 'Linguaggi documentari e basi di dati'. pp.317-329. CNR, Roma, 3-4 Dicembre 1990.

[NEG90] Negrini G. "Obiettivi e pluridisciplinarità di un progetto di ricerca finalizzato." Convegno 'Linguaggi documentari e Basi di dati'. pp.247-271. CNR, Roma, 3-4 Dicembre 1990.

[SAL83] Salton G., McGill M.J. "Introduction to modern information retrieval". McGraw Hill Book Company, 1983.

[SAL86] Salton G.. "Another look at automatic text retrieval systems". Communications of ACM, vol.29, n°7, pp.648-656. 1986.

[SPR87] Sprugnoli R.. "Le basi di dati". Editori Riuniti, 1987

[TOD82] Todeschini C.. "Sistemi post-coordinati e controllo per soggetto." In 'Documentazione e biblioteconomia' a cura di M.P. Carosella e M. Valenti, F.Angeli, 1982.

[VEN86] Venerosi P.. "CASE STUDY: Attuazione di un sistema di Infomation Retrieval basato sulla revisione della struttura, della presentazione, delle modalità di utilizzo del vocabolario dei descrittori attualmente operante" - IEI CNR, Pisa, 1986

[VEN90] Venerosi P. "L'uso di un linguaggio di rappresentazione in un'interfaccia utente-sistema." Convegno 'Linguaggi documentari e Basi di dati'. pp.272-291. CNR, Roma, 3-4 Dicembre 1990.