Estrazione terminologica e gestione della conoscenza: il caso CNR[1]

Roberto Guarasci[2]

***ABSTRACT**: The reform of the National Research Council (C.N.R.) has introduced a new typology for the management of research planning activities, which are now seen as operating transversely as opposed to their being organised by single institutes. The project which is now underway, which is the result of a public/private partnership, hypothesises the use of the term clustering for the organisational analysis and functional reunification of the research policies of the Council. Experts in computer studies, documentation and terminology, working in synergy, have begun to produce the first significant results, which have already taken concrete form in the creation of the C.N.R.'s documentary management system, and in the activation of an Integrated System for the Management of Research Policy (SIGLA).*

La crescente produzione legislativa nazionale in ordine alla necessarietà ed obbligatorietà della produzione di documenti non cartacei ed un parallelo, seppur più tranquillo, trend europeo nello stesso senso hanno fatto vacillare - nell'ultimo decennio - molte certezze disciplinari, incrinato molti consolidati curricula formativi e messo in luce la necessità di convergenze tra discipline che non avevano, nel passato, mai costruito dei legami di organica collaborazione. Questa esigenza, particolarmente evidente, quando si approcciano aspetti specifici dell'analisi testuale è stata, ad esempio, concretizzata in Francia con la costruzione del TIA, gruppo di ricerca su Terminologia e Intelligenza Artificiale, frutto della convergenza di terminologi, linguisti, documentalisti e specialisti ICT[3].
L'approccio multidisciplinare alla classificazione della conoscenza, sostanzialmente nuovo per il panorama italiano, permette di presentare una vista globale del problema postulando la collaborazione di diverse expertises in relazione ai successivi momenti di costruzione dell'ambiente di Knowledge.
Un approccio siffatto permette, inoltre, una riappropriazione umanistica di ambiti, per troppo tempo e per nostra colpa, esclusivo appannaggio di altre discipline non sempre dotate delle necessarie e specifiche competenze ma, sicuramente, titolari di una pretesa omnicomprensiva difficilmente giustificabile a livello teorico. In questo senso ed in questa direzione l'idea di utilizzare metodiche proprie della documentazione, della linguistica computazionale e dell'analisi terminologica per la costruzione di un sistema adattivo semantico finalizzato alla classificazione delle ricerche del Consiglio Nazionale delle Ricerche assume particolare significato e valenza progettuale[4].

Il Decreto Legislativo n. 127 del 4 giugno 2003[5] ha inteso disporre una radicale modificazione del Consiglio Nazionale delle Ricerche al fine di "*promuovere e di collegare realtà operative di eccellenza, di evitare duplicazioni per i medesimi obiettivi, di assicurare il massimo livello di*

[1] Relazione al convegno dell'Associazione Italiana di Terminologia «Terminologie Specialistiche e tipologie testuali», Milano 26-27 maggio 2006, Università Cattolica;
[2] Roberto Guarasci è professore ordinario di Documentazione presso l'Università degli Studi della Calabria.
[3] NATHALIE AUSSENAC GILLES - ANNE CONDAMINES, *Documents èlectroniques et constitution de ressources terminologiques ou ontologiques*, in: *Information - Interaction - Intelligence*, vol 4 (2004), n.1, pp. 75-93.
[4] Il gruppo di progetto è coordinato dal prof. Giovanni Adamo (Illesi-CNR) e ne fanno parte, oltre al CNR, l'Università della Calabria - Laboratorio di Documentazione, l'Università di Catania, l'ISUFI dell'Università di Lecce e L'Università di Bologna, Scuola Superiore Interpreti e Traduttori di Forlì.
[5] d.lgs. 4 giugno 2003 n. 127, pubblicato sulla G.U.R.I. Serie Generale n. 129 del 6 Giugno 2003

flessibilità, di autonomia e di efficienza, nonché una più agevole stipula di intese, accordi di programma e consorzi … (art. 1)". Ciò ha comportato, conseguentemente, un sostanziale ripensamento dell'intera rete scientifica dell'Ente, con la costituzione di 11 Dipartimenti[6], corrispondenti ad altrettante macro-aree di ricerca, la soppressione delle sezioni[7] e l'organizzazione per progetti e commesse[8] delle attività di ricerca, nell'idea di costruire un sistema a matrice "nel quale si incrociano gli Istituti - con funzioni di svolgimento delle attività di ricerca e di gestione dinamica delle competenze - e i Dipartimenti - con funzioni programmatiche - che commissionano la ricerca sia agli Istituti di competenza, sia ad altri Istituti del CNR, sia all'esterno"[9]. Progetti e commesse sono però, quasi sempre, trasversali e multidisciplinari e comportano, quindi, la costruzione di gruppi di ricerca che non tengono necessariamente conto della partizione che caratterizza - ad oggi - l'articolazione geografica della rete scientifica dell'ente. A regime un assetto così configurato dovrebbe portare ad una qualche forma di equilibrio e di coerenza tra la domanda di ricerca, espressa dai Dipartimenti, e l'offerta di ricerca, proposta dagli Istituti, dando contestualmente l'avvio ad una riorganizzazione funzionale che potrà essere definita dopo un primo periodo di funzionamento del sistema.

A supporto di questo modello organizzativo è stato creato SIGLA[10] - *Sistema Informativo per la Gestione delle Linee di Attività* - con il compito di essere non solo funzionale alla gestione finanziaria e contabile delle attività, ma di costituire un più complesso sistema di rappresentazione e registrazione di eventi, monitoraggio e valutazione dei risultati.
Più nel dettaglio, SIGLA è un sistema informativo integrato che supporta:

1) il processo di programmazione triennale;
2) il processo di proposizione e negoziazione delle commesse;
3) la produzione dei budget (preventivi e consuntivi) e la loro gestione;
4) la produzione dei bilanci finanziari (preventivi e consuntivi) e la loro gestione;
5) la produzione del Conto Economico e dello Stato Patrimoniale;
6) la verifica del conseguimento degli obiettivi programmatici e la verifica dei risultati.

Esso è funzionalmente interconnesso con il sistema di gestione documentale dell'ente[11] che collega l'intera struttura di ricerca diffusa sul territorio nazionale. In quest'ultimo sistema, al fine

[6] Agroalimentare, Scienze della Vita, Medicina, Materiali e dispositivi, Progettazione molecolare, Terra e Ambiente, Teconologie dell'Informazione e delle Comunicazioni, Sistemi di Produzione, Identità culturale, Patrimonio culturale, Energia e Trasporti.

[7] Il regolamento di organizzazione e funzionamento del CNR, emanato in attuazione del decreto di riordino, non prevede l'esistenza di "sezioni territorialmente distinte" degli Istituti che, secondo il previgente ordinamento, erano articolazioni degli stessi istituti che costituivano "autonomo centro di spesa" ai fini della gestione delle attività. A far data dal 1 giugno 2005, con l'entrata in vigore dei regolamenti e, in funzione di quanto disposto con il provvedimento del Presidente del CNR n. 35 del 31 maggio 2005, le sezioni hanno cessato di esistere e l'Istituto è una struttura unica sotto la responsabilità del direttore.

[8] Si definisce *commessa* il segmento di attività che un organo esecutore (ad es. l'Istituto) concorda di svolgere per conto di un committente esterno - il capo progetto - il cui riferimento è il Dipartimento.

[9] *La Gestione per commesse delle attività di ricerca e sviluppo del CNR*, DOC.CDA (04) 16.

[10] R. GUARASCI, M. LANCIA, J. F. MASCARI, F. TUZI, R. PUCCINELLI, A. ROVELLA, *SIGLA: a flexible portal and reusable ERP system for Research Management,* Orlando (USA), 2005

[11] Con deliberazione 57/2004 del 23 dicembre 2004 il CDA del CNR ha definito la strutturazione delle Aree Organizzative Omogenee con le caratteristiche e le specificità di cui all'art. 50 del D.P.R. 445/00; Con deliberazione n. 21/2005 del 9 febbraio 2005, ha approvato due accordi di collaborazione con l'Università della Calabria, la società Prisma Engineering S.r.L. e la società FileNet Italy S.r.l. per la realizzazione di un sistema di protocollo informatico e di gestione dei flussi documentali da integrarsi con il Sistema Informativo per la gestione delle linee di attività (SIGLA);
Con deliberazione n. 39/2005 del 6 aprile 2005 è stato approvato il titolario per la classificazione degli atti dando, contestualmente, mandato al Presidente di istituire con proprio decreto il servizio per la gestione informatica dei documenti, dei flussi documentali e degli archivi ai sensi dell'art. 61 del DPR 445/00. Con provvedimento del Presidente n. 44/2005 del 21 giugno 2005 è stato istituito, presso la Direzione generale, dell'Ufficio per la gestione informatica dei flussi documentali e degli archivi e nominato il relativo responsabile; A decorrere dall'8 settembre 2005 il sistema è stato avviato limitatamente alle AOO Amministrazione Centrale e Presidenza. Il 28 febbraio 2006 è iniziato il collegamento dei 107 istituti che si è concluso il 30 marzo dello stesso anno.

anche di dare pratica attuazione a quanto previsto dalla normativa vigente in tema di dematerializzazione e gestione della documentazione non cartacea, è prevista, oltre all'obbligatoria protocollazione degli atti - ivi inclusa la documentazione di ricerca - la classificazione degli stessi mediante l'utilizzazione di un titolario e la scansione ottica di quanti risultino essere redatti su supporto cartaceo. Il recupero dell'informazione, oltre che mediante gli obbligatori indicatori numerici, avviene, previa tokenizzazione e stemming, limitatamente alle stringhe di testo digitate nei campi della base di dati. Ciò anche in considerazione del fatto che la produzione documentale è in massima parte su carta e l'obbligo - previsto nel manuale di gestione - di allegare il file ai documenti scanditi, anche in assenza di firma digitale, non ha sortito - fin'ora - effetti di rilievo rendendo di difficile applicazione altri sistemi di retrieval. La necessarietà di circoscrivere la ricerca in una catena composta da un limitato numero di sintagmi, la cui digitazione non è sicuramente scevra da un variabile grado di arbitrarietà in dipendenza dell'operatore umano, oltre ai noti limiti caratteristici della tipologia di retrieval utilizzata, ha indotto a prefigurare ipotesi più performanti in ordine alle logiche di interrogazione. Un primo tentativo è stato quello di richiedere l'inserimento di una tripletta di parole chiave - a testo libero - ad ognuno dei redattori dei documenti. L'artifizio, i cui limiti sono noti, non si prefiggeva di essere necessariamente una soluzione ottimale quanto di valutare se il contesto specialistico di utilizzazione, pur se multidominio, riusciva in tutto o in parte ad ovviare alla genericità dei termini di solito introdotti in contesti similari. I risultati sono stati, sostanzialmente, ambivalenti. Da un lato il livello di rispondenza delle keyword introdotte rispetto ai testi di origine è risultato soddisfacente ma relativo al solo tema principale delle ricerche, senza, quindi, tenere in nessuna considerazione l'articolazione interna delle stesse. Quest'ultima informazione risultava, però, spesso più significativa della prima ai fini della comprensione delle linee di attività e della ricaduta industriale dei risultati, anche in considerazione del fatto che proprio la citata logica delle aggregazioni trasversali per progetti e commesse obbligava ad un elevato grado di complessità interna. Ulteriore passaggio è stato la costruzione di liste predefinite di termini con implementazione controllata. Partendo dal presupposto che le attività di ricerca hanno, in genere, una durata triennale e che, quindi, il corpus dei testi delle commesse dell'anno 2005 è sovrapponibile di almeno l'80% rispetto a quello dell'anno in corso si è provato ad estrarre da tale corpus - costituito dalle circa 500 commesse CNR dell'anno 2005 - delle liste di termini mediante un procedimento di analisi concettuale e di lemmatizzazione morfologica delle stringhe di soggetto ottenute alla fine del processo. Ciò ha prodotto 998 termini di indicizzazione la cui significatività era sicuramente maggiore rispetto alle parole chiave digitate direttamente dagli utenti ma che - slegate dal contesto di riferimento - non erano sicuramente esaustive rispetto ai risultati attesi. Anche l'ipotesi di considerare i termini estratti come segmenti di un thesaurus implementando quindi, per ognuno di essi, l'insieme delle relazioni contenute in un vocabolario di indicizzazione precoordinato non è riuscita ad avere pratica attuazione a causa della più volte ricordata tipologia multidominio dei documenti trattati e della conseguente indisponibilità delle relative tassonomie. Inoltre, nel frattempo, era anche intervenuta una ulteriore richiesta del management dell'ente di organizzare le voci di indicizzazione in una visione spaziale tridimensionale nella quale gli assi XYZ avrebbero dovuto rappresentare, rispettivamente, l'*oggetto*, l'*obbiettivo* e l'*approccio* metodologico della commessa. Ciò al fine di permettere un'immediata verifica dei casi di sovrapposizione/duplicazione di attività e la trasformazione di quelli riscontrati in attività sinergiche mediante l'eventuale modifica di una delle variabili[12].

Una commessa: *Studio chimico-fisico delle discariche per l'analisi dei terreni* … verrebbe così

[12] A questo punto è, però, necessario fare una brevissima digressione sulla tipologia delle ricerche attualmente in essere al CNR per capire appieno la ratio e la finalità della richiesta:"…in applicazione della legge di riforma le attività dell'ente sono classificabili in tre direttrici, fortemente correlate agli Assi previsti dal Programma Nazionale di Ricerca: attività di ricerca spontanea a tema libero (curiosity driver); attività di ricerca finalizzata allo sviluppo delle competenze; attività di ricerca, sviluppo e dimostrazione finalizzate alla realizzazione degli obbiettivi posti dalle aree tematiche a carattere strategico sulle quali l'ente è prioritariamente impegnato". *La Gestione per commesse delle attività di ricerca e sviluppo del CNR* - DOC.CDA (04) 16, p. 2.

segmentata in studio delle discariche (obbiettivo) chimico/fisico (approccio) analisi dei terreni (oggetto).

In una visione euclidea ogni commessa dovrebbe quindi essere riconducibile ad un punto (x,y,z) nello spazio tridimensionale; più commesse ricadenti in un intervallo determinato costituiscono un *cluster*[13]. I Cluster tematici così costruiti vanno a rappresentare le future e ipotetiche aggregazioni organizzative interne dell'ente.

La risposta a questa esigenza incontrava un limite negli "…attuali metodi statistici e strumenti di data mining che estraggono da dati e informazioni dei pattern ricorrenti considerando solo i loro attributi e non la conoscenza di dominio. Tali sistemi non sono quindi in grado di richiamare parte della conoscenza assimilata in precedenza per disambiguare il contenuto di parole e frasi, assegnando nuovi significati al testo non direttamente ricavabili dal suo contenuto."[14] Oltre a ciò la costruzione dei cluster tridimensionali postulava l'assegnazione di pesi ai collegamenti tra termini, pesi che, in ragione dei corpora di origine - dovevano necessariamente essere automaticamente aggiornati con il variare del contenuto dei documenti utilizzati o con l'aggiunta di nuovi testi.
Una soluzione al problema potrebbe essere costituita da un formalismo di rappresentazione della conoscenza come rete associativa[15]. La propagazione di un segnale di attivazione che si diffonde attraverso i nodi semantici di una rete, nodi costituiti dai significati che un termine riveste all'interno di un testo o di una porzione di esso, fino a quando il suo valore non diventa stabile dopo essersi amplificato e/o accumulato a seconda dei pesi dei collegamenti attraversati permette di individuare i nodi maggiormente attivati definendo quindi il contesto di utilizzo dei termini nel documento di riferimento.
Così l'analisi del testo di una commessa: Materiali e processi per l'energetica, individua i nodi corrispondenti ai termini: "materiali", "processi" ed "energetica" e, mediante la diffusione di un segnale di attivazione, può arrivare ad attivare i nodi "combustione" e "alte temperature" ai quali è collegato il nodo "materiali", mentre il livello del segnale di attivazione in presenza del nodo "anemometria" risulterebbe basso, mentre potrebbe attivarsi maggiormente nel caso della commessa "materiali per l'energia eolica". Il processo di assegnazione del significato di dominio è da intendersi - comunque - semi automatico in quanto è possibile verificare la compresenza - a parità di livello del segnale - in due sottodomini semantici anche, pur se non necessariamente, alternativi. Il vantaggio del sistema in presenza di testi multidominio e non completamente strutturati, anche se in parte preformattati, è la sua indipendenza dal parsing sintattico che consente un alto livello di adattività ai vari linguaggi specialistici utilizzati dai redattori. L'individuazione di nodi univoci con pesi determinati non è funzionale solo alla costruzione dei cluster ma, andando a prefigurare non solo l'ontologia di dominio, come somma dei nodi rilevanti, ma anche delle pseudo-ontologie per singola entità testuale, permette la costruzione non solo del profilo del redattore sulla base della query, ma anche di un profilo del documento e della sua relazione rispetto ai tre assi individuati[16].

[13] Cfr STEPHANE LAPALUT, *Text Clustering to support Knowledge Acquisition from Documents*, Rapport de Recherche INRIA n. 2639, Agosto 1995.
[14] LUIGI LELLA, *Oltre il Datamining: Verso l'estrazione della conoscenza tacita*, www.itconsult.it.
[15] J. L. MC CLELLAND - D.E. RUMELHART, *Parallel distributed processing*, Cambridge, MIT Press, 1986. Altra ipotesi per la strutturazione del sistema di classificazione è rappresentata dall'uso di una SOM (Self Organizing Map) rete neurale non supervisionata.
[16] L'analisi dei testi italiani per l'identificazione ed annotazione della terminologia di dominio a partire dai testi specialistici delle ricerche è affidata ad "AnIta" (Analizzatore dell'Italiano) un modulo software ideato dall'Istituto di Linguistica Computazionale del CNR e già utilizzato con successo in Pekita: Personalized Knowledge In The Air - Strutturazione dinamica del reperimento e della fruizione della conoscenza senza vincoli di spazio e di tempo, progetto di ricerca congiunto tra l'Università della Calabria e Siemens Italdata. Cfr: ROBERTO BARTOLINI, ALESSANDRO LENCI, SIMONE MARCHI, SIMONETTA MONTEMAGNI, VITO PIRRELLI, *Text-2-Knowledge, Acquisizione semi automatica di ontologie per l'indicizzazione semantica di documenti,* Rapporto di lavoro Pekita, Agosto 2005. Cfr anche FELICE DELL'ORLETTA, ALESSANDRO LENCI, SIMONE MARCHI, SIMONETTA MONTEMAGNI, VITO PIRRELLI, *Text-2-Knowledge: uno strumento linguistico-computazionale per l'estrazione di conoscenza da testi*, XL Congresso della Società di Linguistica Italiana, Vercelli, 2006; S. FEDERICI, SIMONETTA MONTEMAGNI, VITO PIRRELLI. (1998a),

Le novità vere - come dicevamo all'inizio - sono nell'approccio multidisciplinare, del quale le soluzioni tecnologiche sono naturali emanazioni, nell'ambito di applicazione - unico vero ambiente multidominio nel panorama della ricerca nazionale ed, in ultimo, nelle finalità organizzative interne, non certo comuni come output dichiarato delle attività di document management o di estrazione terminologica.

"Chunking Italian: Linguistic and Task-oriented Evaluation", in *Proceedings of the First International Conference on Language resources and Evaluation*, Granada, Spain, 1998.