Un laboratorio multimediale dedicato a Carlo Emilio Gadda: il modello e i primi dati implementati in formato XML.

A multimedial laboratory dedicated to Carlo Emilio Gadda: the model and a partial XML implementation

Maria Luigia Ceccotti, Manuela Sassi, Gabriella Pardelli

Istituto di Linguistica Computazionale del CNR, Pisa

Abstract

Con questa comunicazione presentiamo un progetto CNR '*L'Archivio elettronico delle Opere di Carlo Emilio Gadda: supporti lessicografici e bibliografici in XML'*.
Il testo è composto di due sezioni: nella prima è presentato l'Archivio elettronico delle Opere di Carlo Emilio Gadda, realizzato presso l'ILC, nella seconda gli obiettivi del progetto e i risultati conseguiti nei primi mesi di attività.

In this paper we present a project of the Italian National Council of Research titled "Gadda's Electronic Archive: Lexicographical and bibliographical Tools in XML".
The text is made of two sections: in the first, we present Gadda's Electronic Archive, implemented at the ILC, and in the second, we show the project's objectives and the results achieved in the first months of work.

1. Prima sezione

La realizzazione dell'Archivio Elettronico delle Opere di C.E.Gadda è stata avviata nel luglio 1994 partendo dal materiale codificato per la fotocomposizione che la Garzanti Editore s.p.a. ha messo a disposizione. Si tratta dell'edizione diretta da Dante Isella, collana I Libri della Spiga, 1988-93 [Gad 88-93].
I file predisposti per la fotocomposizione sono stati 'ripuliti' dai codici approntati per la stampa dei volumi garzantiani e sono stati sottoposti ad un faticoso lavoro redazionale che ha trasformato i testi piatti (raw texts) in testi strutturati, predisposti per il DBT [Pic 97] , un sistema interattivo testuale, che tramite funzioni accessibili mediante appositi menu, permette la lettura e la ricerca testuale in un *archivio*, (costituito da uno o più testi) ovvero in un *corpus*, cioè in un insieme di archivi.
Il lavoro redazionale è consistito, dopo l'eliminazione dei comandi per la fotocomposizione, nell'inserimento manuale e/o automatizzato di 'codici'. Riferiamo qui brevemente i codici applicati e ciò che il loro corretto inserimento genera in fase di consultazione del singolo archivio o del corpus:

codice di riferimento logico: assegna un identificatore ad ogni unità logica del testo, unità logica che può coincidere con tutto il testo. E' il codice che trasforma un file, contenente un testo in machine-readable form, in un file-input-DBT ed è quindi l'unico codice di questo sistema proprietario che è obbligatorio inserire; l'identificatore sarà attribuito dal sistema ad ogni forma che fa parte dell'unità logica da esso etichettata;

codice di riferimento topografico:	visualizza la posizione fisica (numero di pagina e di riga) dell'occorrenza di una parola nel testo cartaceo; forma un binomio con il codice precedente e insieme costituiscono l'informazione fondamentale nella fase di text retrieval;
codice maiuscola:	distingue la parola iniziante con maiuscola dopo il punto dalla parola iniziante con maiuscola per norma o per volontà dell'autore del testo;
codice legame:	tratta come unità lessicale più stringhe di caratteri;
codice segno speciale:	disambigua segni che il sistema deve valutare differentemente; il trattino ad esempio può avere almeno tre funzioni: quella di trait d'union, di punteggiatura e di divisione di parola a fine riga;
codice personaggio:	attribuisce ad un personaggio la sua parte di testo;
codice poesia:	segnala l'inizio e la fine di un testo poetico;
codice linguaggio:	costruisce sottoinsiemi di lingue naturali, di linguaggi speciali o di parole che si vogliono raggruppare per una particolare caratteristica;
codice data, numero, sigla:	riconosce numeri, date, sigle;
codice nota:	inserisce delle icone, che attivate visualizzano il testo di note;
codice immagine:	segnala la presenza nel testo cartaceo di immagini, e permette di visualizzare il contenuto dei relativi *file-album* precedentemente creati.

Il DBT, quindi, permette la consultazione di testi preparati secondo regole di codifica (proprietarie) e fornisce, attraverso un'interfaccia abbastanza semplice, funzioni di ricerca full-text ed altre più sofisticate. Oltre alle funzioni di ricerca full-text più semplici, di facile comprensione e utilizzo e comuni ad altri sistemi di full-text retrieval, il sistema ne permette altre, orientate a fini linguistici, che difficilmente si trovano in altri sistemi di interrogazione.
La ricerca di parole permette, ad esempio, di ottenere la lista delle forme

- ✓ che hanno una o più vocali accentate
- ✓ che hanno una determinata vocale accentata,
- ✓ che hanno una determinata lunghezza e/o che al loro interno presentino sequenze di caratteri definiti o parzialmente indefiniti (es.: tutte le parole di 5 lettere, oppure tutte le parole che contengano 'g$i' dove il $ indica il carattere jolly che vale qualunque lettera dell'alfabeto, etc.).

Il sistema offre inoltre altri strumenti più sofisticati: ad esempio lo studio delle reggenze di un verbo può far riferimento alla lista di tutti i contesti delle forme verbali, ordinati alfabeticamente sulla parola seguente o su quella antecedente; oppure per lo studio di documenti manoscritti (o su altri supporti come nel caso delle iscrizioni lapidee) si possono visualizzare un'immagine e il testo corrispondente, partendo da una parola o da un'immagine. E' possibile anche il calcolo delle cooccorrenze (mutual information) sulla base di una o più parole date, con possibilità di variare i parametri (contesto destro/sinistro, soglia di frequenza, stopwords) del calcolo statistico. Non ci soffermiamo qui su varie applicazioni collaterali, esterne quindi all'ambiente di consultazione, che vanno dalla produzione di vari tipi di indici all'analisi automatica o assistita (lemmatizzazione/tagging), alla creazione di archivi bilingui sincronizzati tramite motori di analisi morfologica plurilingue (italiano, inglese, francese, spagnolo, latino).

La costruzione dell'Archivio elettronico di Gadda è stata realizzata quindi con i vantaggi ed i vincoli di questo sistema proprietario, vincoli che, in pratica, abbiamo pensato di 'interfacciare' con la nostra esperienza in modo da produrre risultati che siano utili strumenti di studio: dopo il 1997 [CNR 97] ci siamo trasformate infatti da redattrici a fruitrici di questo database producendo degli strumenti lessicografici, cartacei ed elettronici.

2. Seconda sezione

Giorgio Bocca nell'articolo 'Cari credenti del Web, vi regalo qualche dubbio' apparso su L'Espresso del 24 febbraio scorso dimostra il suo 'pessimismo cosmico' ironizzando nei confronti di chi crede che 'Fuori dal Web, fuori dalla rete informatica, non si vive', e che valga 'L'efficienza prima di tutto, la comunicazione e la pubblicità prima della produzione'. Concordare con questo punto di vista non esclude a nostro parere il tentativo di 'imparare' a muoversi con un minimo di competenze nella società caratterizzata dalla rivoluzione informatica e dal 'presunto' mercato globale. Questo per introdurre brevemente le motivazioni che ci hanno convinto dell'opportunità di studiare perchè e come proporre 'indirettamente' il 'prodotto elettronico' Gadda al pubblico attraverso la rete.

Se per il DBT si prevede il suo inserimento in Internet [Pic 99], nell'ottica dell'integrazione di due tecnologie, web e basi di dati, il progetto di costruire in rete un sito dedicato alle opere di/su Gadda in cui implementare un modello di laboratorio culturale, in cui elementi iniziali fossero alcuni brani gaddiani, nostre pubblicazioni, dati bibliografici, laboratorio da arricchire in modo interattivo con il contributo del lettore di Gadda, studioso, studente, curioso, ci è sembrato un passo obbligatorio anche per divulgare l'attività finora svolta [Cec 00].
Documentandoci su Internet, sul più ricco archivio di documenti testuali e ipertestuali, ci siamo convinte che ci apprestavamo di sicuro ad una fatica degna di Sisifo se avessimo affrontato l'accesso ad Internet come la maggior parte dei suoi utenti che hanno creato siti web con approcci empirici e privi di una valida metodologia per quanto riguarda l'analisi e la progettazione.

Abbiamo imparato che il futuro della RETE è stato avviato dagli addetti ai lavori nel corso della più importante manifestazione a livello internazionale di Internet, la Internet World, dove, nel dicembre del 1996, è stata proposta una rivoluzionaria tecnologia: la 'information push', che privilegia l'utente spettatore rispetto all'utente-navigatore in quanto un sito è un canale che trasmette dati automaticamente all'utente interessato a riceverli, utente, che, in questa rivoluzione, diventa soggetto, spettatore attivo che riceve dati selezionati ab origine a seconda delle sue richieste, che possono essere dinamicamente sollecitate dall'interazione con l'informazione ricevuta.

Abbiamo imparato che, nella gestione delle informazioni in un sito web, è opportuno mutuare dalle BD la distinzione dei seguenti tre livelli, quello della struttura dei dati, quello dei dati e quello della visualizzazione dei dati sì da poter utilizzare gli stessi dati per scopi e ambienti di lavoro differenti.

Abbiamo imparato che se l'HTML, il linguaggio base del WWW, è stato realizzato principalmente per la visualizzazione dei dati, l'XML, il cui progenitore è il linguaggio SGML, è un metalinguaggio che offre, oltre ai dati, informazioni di tipo strutturale e semantico sui dati.

Abbiamo deciso di conseguenza che se l'obiettivo finale, quasi certamente utopico, è la creazione di un sito-Fondazione in cui consultare testi di/su Gadda, disegni, fotografie, bibliografie, cataloghi, in breve tutto ciò che può aiutare a studiare questo scrittore difficile, l'obiettivo raggiungibile poteva essere la attivazione di un sito 'biglietto da visita' dell'attività finora svolta.

Se non possiamo mettere a disposizione l'archivio elettronico, possiamo, per ora, proporre i risultati da noi conseguiti mediante il suo utilizzo insieme con altri strumenti di lavoro (dove ovviamente lavoro=studio) non in una biblioteca, in uno studio tradizionale ma tramite il computer, adeguandoci *cum grano salis* alla moda corrente, il web, per contribuire allo studio della complessità del mondo di Gadda.
Dei quattro nostri report disponibili attualmente in rete mentre il primo [Cec 97] è la descrizione abbastanza dettagliata di come è stato da noi realizzato l'Archivio di Gadda, gli altri tre [Cec 98-99b] contengono essenzialmente dati che sono stati da noi estratti dall'Archivio e pubblicati dopo un attento controllo. L'inserimento di questi tre report è il risultato di un compromesso da noi attuato: le introduzioni dei lavori sono state implementate in HTML in quanto era sufficiente la loro visualizzazione mentre i dati (liste di migliaia di record) sono stati inseriti utilizzando le specifiche del formato XML

Per quanto riguarda il recupero di informazioni su Gadda via Internet abbiamo utilizzato per ora un motore di ricerca commerciale (Excite) e abbiamo messo a disposizione degli auspicabili 25 visitatori del sito alcuni dei tanti link che abbiamo selezionato, proponendoci in breve termine di strutturarli a seconda del dato che contengono (che può essere un semplice accenno a Gadda in un importante testo o l'avviso di uno spettacolo teatrale dedicato a Gadda)

Abbiamo aperto questa sezione citando Giorgio Bocca, permetteteci di chiuderla citando Massimo Riva che sempre su L'Espresso del 24 febbraio scorso nell'articolo 'Ecco a voi i nuovi padroni' scrive tra l'altro:

> "Il punto cruciale è che la globalizzazione dell'economia mondiale ha posto la telematica e i processi di innovazione tecnologica in cima e al centro della competizione mercantile. Per giunta, con una velocità di penetrazione impensabile in altri settori perché legata non alla produttività di costose immobilizzazioni tecniche, ma alla immateriale e mobilissima capacità inventiva della mente umana. Negli Stati Uniti - ha ricordato Carlo De Benedetti in un sua recente conferenza alla London School of Economics di Londra - le imprese tendono ormai a pensare in termini di anni-rete, intendendo con questa definizione un quarto di anno finanziario tradizionale. Il che equivale a ritenere che alla rivoluzione telematica si attribuisce il potere di produrre in una sola generazione i mutamenti intervenuti in quattro nel corso della gloriosa rivoluzione industriale verificatasi durante il Novecento. ... Insomma, una nuova economia batte alle porte e scosse violente modificano la gerarchia del potere finanziario. Ma le virtù e soprattutto i vizi dell'umano agire sono quelli di sempre, per cui il naturale assillo dell'incertezza resta il tema dominante."

Bibliografia

[Cec 97] Ceccotti M.L., Sassi M., "L'Archivio elettronico delle Opere di C.E.Gadda: come è stato costruito, come si consulta", ILC, Pisa, S.T.A.R., 1997.

[Cec 98] Ceccotti M.L., Sassi M., "Apax in Gadda - Un Indice Inverso", ILC-CNR, Pisa, S.T.A.R., 1999.

[Cec 99a] Ceccotti M.L., Sassi M., "Forme accentate in Gadda - Un Index locorum", ILC-CNR, Pisa, S.T.A.R., 1999.

[Cec 99b] Ceccotti M.L., Sassi M., "Alla ricerca dei termini gaddiani. Una pre-concordanza", ILC-CNR, Pisa, S.T.A.R., 1999.

[Cec 00] Ceccotti M.L., Sassi M., Pardelli G., "Il soccorso informatico per lo studio di uno scrittore difficile, C.E. Gadda", Didamatica, Cesena 4-5-6 maggio 2000.

[CNR 97] Convegno "Uno strumento informatico per la lettura e l'analisi delle *Opere* di Carlo Emilio Gadda", CNR-ILC-P.F. MADESS-Università di Roma Tor Vergata e Roma III, 14 novembre 1997, Aula Marconi del CNR, Roma.

[Gad 88] Gadda, C.E., "Romanzi e racconti I", Collana I Libri della Spiga, Garzanti, Milano 1988.

[Gad 89] Gadda, C.E., "Romanzi e racconti II", Collana I Libri della Spiga, Garzanti, Milano 1989.

[Gad 90] Gadda, C.E., "Saggi Giornali Favole I", Collana I Libri della Spiga, Garzanti, Milano 1990.

[Gad 91] Gadda, C.E., "Saggi Giornali Favole II", Collana I Libri della Spiga, Garzanti, Milano 1991.

[Gad 93] Gadda, C.E., "Scritti Vari e Postumi", Collana I Libri della Spiga, Garzanti, Milano 1993.

[Pic 97] Picchi, E., "DBT 3 - Data Base Testuale: Guida all'uso", versione 3.1. Lexis Ricerche s.r.l. su licenza del C.N.R., Roma, 1997, 96 p.

[Pic 99] Picchi, E., "Informatica e scienze umane: Procedure di analisi testuale" in Parola e Immagine, a cura di M.A. Zanetti, La Nuova Italia Editrice, Firenze, 1999.