Archeologia e Calcolatori
7, 1996, 721-734

EDIZIONE E ANALISI INFORMATICA DI TESTI: STANDARD INTERNAZIONALI PER LA CODIFICA DEI DATI TESTUALI

1. Testi e informatica

«È ormai tempo di domandarsi se il computer sia stato usato in modo tale da modificare significativamente la nostra prospettiva dell'universo letterario, o se invece sia stato usato soltanto per mostrare, con profilo più preciso e dati più consistenti, ciò che già ritenevamo esservi contenuto»[1]. Con questo invito a esaminare la reale portata innovativa conseguente all'adozione dello strumento informatico nello studio dei testi, si apriva un articolo di Susan Wittig, pubblicato nel 1978 dalla rivista «Computers and the Humanities»: *The computer and the concept of text*. Considerare un testo «un'entità lineare, un'opera definitivamente compiuta – prosegue la Wittig – saldamente ancorata alla sua rappresentazione grafica, la pagina stampata; autonomamente indipendente da ogni altra entità, significativa in sé»[2] è dovuto all'influsso di modelli linguistici e critici quali la grammatica strutturalista-formalista e il New Criticism.

Qualche tempo prima era apparso sulla stessa rivista un articolo di Todd K. Bender, *Literary texts in electronic storage: The editorial potential*, al quale la stessa Wittig fa riferimento. «Il "vero" deposito di informazione - afferma Bender – sono i dati elettronici. Ogni espressione a stampa di quei dati è soltanto uno fra i tanti formati provvisori, incompleti e arbitrari che è possibile ottenere dall'informazione che esiste nella memoria elettronica nella sua forma più piena e flessibile. Piuttosto che usare l'apparato elettronico semplicemente per imitare le caratteristiche del deposito dei dati a stampa, dovremmo cercare di capire in quali modi l'apparato elettronico possa essere uno strumento migliore della carta per preservare il nostro patrimonio letterario»[3].

Illustrando il suo lavoro su Joseph Conrad, Bender parla di una banca

[1] «Now [...] it is time to inquire whether the computer has been used in ways that significantly alter our view of the literary universe, or whether it has merely been used to show, in more distinct outline and with more substantiating data, what we already knew to be there», Wittig 1978, 211.

[2] «a linear entity; [...] a one-time, completed work [...] firmly confined to its graphic representation, the printed page; [...] autonomously independent of any other entity, [...] meaningful in and of itself», *Ibid.*, 211-212.

[3] «The "real" repository of information is the electronic data. Any printed expression of that data is merely one among many possible provisional, incomplete, and arbitrary formats of information which exists in its fullest and most flexible form in electronic

dati in grado di accogliere differenti *stati* o *versioni* di un medesimo testo e fa riferimento a una dimensione orizzontale (*length*), nel suo caso l'insieme delle opere a stampa di Conrad registrate in formato elettronico, alla quale si affianca una dimensione verticale (*depth*), che consiste nelle differenti versioni registrate per ciascuna opera. «Sull'asse orizzontale, possiamo mettere a confronto il vocabolario o i modelli sintattici di un romanzo iniziale con quelli attestati in un'opera più tarda. Nello spaccato verticale, possiamo verificare come Conrad abbia modificato l'ortografia e l'interpunzione per un'edizione inglese rispetto a un'edizione americana, o come abbia perfezionato il suo stile o cambiato le sue idee con il variare delle revisioni»[4].

La rappresentazione di testi in formato elettronico acquista una valenza *dinamica* – in quanto svincola il testo dalla staticità della riproduzione a stampa – ma anche *multidimensionale*. Grazie alle potenzialità dello strumento informatico, si realizza infatti una "nuova" lettura dei testi, che è stata opportunamente definita «sintetica» o «sinottica»[5]. Adeguati algoritmi di interrogazione o di elaborazione di liste di concordanza presentano simultaneamente al lettore tutti i luoghi del testo in cui una determinata parola ricorre, offrendo anche la possibilità di verificarne la differente funzione sintattica all'interno di proposizioni diverse (anche nei casi di omografia) e il diverso significato che forme graficamente identiche esprimono, nei casi di ambiguità semantica o di omonimia, favorendo inoltre lo studio delle variabili (i *loca variantia* e le *variae lectiones* dei filologi).

Gli spunti di riflessione su cui ci siamo soffermati ci inducono ancora una volta a considerare quanto un approfondimento teorico e metodologico del concetto stesso di «testo» possa avere riflessi determinanti sull'elaborazione informatica di dati testuali. In verità, scorrendo le fonti bibliografiche, si deve constatare che la letteratura relativa agli aspetti applicativi è quantitativamente assai più cospicua rispetto ai contributi teorici e metodologici. E il dato non sorprende, se si considerino le difficoltà che derivano agli studiosi umanisti dall'applicazione di tecniche e metodi poco familiari al loro lavoro. Ci sembra però che negli ultimi anni sia da registrare un'inversione di tendenza. Per verificarlo, ci soffermeremo a considerare l'evoluzione dell'analisi

memory. [...] Rather than use electronic equipment merely to imitate the features of printed data storage, we should try to see in what ways it can be a better medium than paper for preserving our literary heritage», BENDER 1976, 194-195.

[4] «On the horizontal axis, we can compare vocabulary or syntactical patterns in an early novel to those found in a later work. In the vertical stack, we can see how Conrad modified his spelling and punctuation for an English edition as opposed to an American edition, or how he refined his style or modified his ideas in varying revisions», *Ibid.*, 196.

[5] TOMBEUR P. 1977, *Informatique et étude de textes. Pour une meilleure connaissance du vocabulaire médiolatin*, «Archivum Latinitatis Medii Aevi (Bulletin Du Cange)», XL, 131; MARZULLO B. 1968, *A proposito di una concordanza*, «Quaderni dell'Istituto di Filologia greca dell'Università di Cagliari», III, 7; GUILBAUD G.Th. 1982, *Statistique et philologie*, in M. FATTORI, M. BIANCHI (eds.), *Res. Atti del III Colloquio Internazionale del Lessico Intellettuale Europeo (Roma, 7-9 gennaio 1980)*, Roma, Edizioni dell'Ateneo, 18.

informatica di testi, con particolare riferimento ad alcune esperienze italiane, rivolgendo la nostra attenzione all'analisi non solo di opere letterarie ma anche di altre forme di produzione culturale.

2. La situazione italiana

Se si eccettuano alcuni tentativi di traduzione automatica compiuti nel corso della Seconda guerra mondiale, l'intuizione di applicare le macchine calcolatrici e poi i calcolatori digitali al trattamento di testi in lingua naturale si deve a Roberto Busa che, dopo aver condotto alcuni esperimenti sul III canto dell'*Inferno* dantesco e sugli inni liturgici di Tommaso d'Aquino, si è dedicato per lungo tempo all'analisi dell'intero *corpus* tomistico. A pochi anni di distanza da questi primi esperimenti, un numero dell'*Almanacco letterario Bompiani* (Milano 1962) veniva dedicato a esaminare il rapporto tra elettronica e letteratura.

A metà degli anni Sessanta, l'Accademia della Crusca poneva in cantiere lo spoglio della produzione letteraria italiana, per la redazione di un *Grande dizionario della lingua italiana*. Dell'ingente numero di opere memorizzate sono stati pubblicati alcuni volumi di concordanze e negli ultimi anni ha preso corpo l'idea di pubblicare un *Tesoro della lingua italiana dalle origini fino al 1375*, che il Centro di studio per l'Opera del Vocabolario Italiano prevede di condurre a termine entro il 2021, anno in cui si celebrerà il settimo centenario della morte di Dante. Già nel corso del 1996 si potrà accedere via Internet al patrimonio lessicale raccolto, che conta attualmente 14 milioni di forme. Parallelamente a questo progetto, e in qualche modo raccordate con esso, D'Arco Silvio Avalle ha realizzato le *Concordanze della lingua poetica italiana delle origini*.

Sul modello del progetto della Crusca, e anzi per ampliarne e integrarne gli ambiti disciplinari, prendevano consistenza altri due progetti del Consiglio Nazionale delle Ricerche: il *Vocabolario Giuridico* dell'Istituto per la Documentazione Giuridica di Firenze, e il *Lessico filosofico dei secoli XVII e XVIII* del Lessico Intellettuale Europeo di Roma. Del primo, che non ha raggiunto una realizzazione editoriale, è oggi a disposizione degli studiosi l'archivio lessicografico. Accanto ad esso, non si può omettere di ricordare i monumentali volumi del *Legum Iustiniani Imperatoris Vocabularium*, realizzati in collaborazione con l'Università fiorentina.

Del *Lessico filosofico* sono stati pubblicati i primi due fascicoli relativi all'ambito linguistico latino (Vol. I, 1 «a - aetherius», Roma, Edizioni dell'Ateneo, 1992; Vol. I, 2 «aetherius - animositas», Firenze, Leo S. Olschki, 1994), ed è in corso la stampa del terzo. Nella collana «Lessico Intellettuale Europeo» sono inoltre apparsi volumi di concordanze e indici di singole opere, e i lessici d'autore di Giordano Bruno e Francesco Bacone. Negli anni Settanta e nei primi anni Ottanta le imprese fin qui citate applicavano le tecniche e i pro-

grammi di trattamento messi a punto dai ricercatori dell'Istituto di Linguistica Computazionale del CNR, con sede a Pisa e diretto da Antonio Zampolli.

Vale la pena di sottolineare lo stimolo e l'attrattiva esercitati dalla grande potenzialità di "memoria" offerta dai calcolatori elettronici nella progettazione e nell'impianto di queste grandi imprese documentarie[6].

Un momento importante di transizione e insieme di ripensamento si è determinato agli inizi degli anni Ottanta, potremmo dire in coincidenza con la progressiva diffusione dell'«informatica distribuita». Si registrava in quel periodo una situazione di stallo nelle maggiori imprese; nello stesso tempo, peraltro, i sintomi di tale crisi si avvertivano anche a livello internazionale. Ci si trovava, infatti, a constatare che l'adozione su vasta scala di tecniche informatiche ispirate a criteri di massima rappresentatività, e spesso di esaustività, finiva per rendere difficile e forse impossibile il dominio della mole di risultati prodotti dalle elaborazioni elettroniche, soprattutto nel settore della documentazione lessicografica. Cominciavano inoltre a diffondersi con capillarità sempre crescente i *personal computers*, macchine di dimensioni notevolmente ridotte eppure dotate di risorse di calcolo sempre più elevate.

Risalgono proprio a quegli anni una serie di iniziative e di attività di ricerca, più o meno istituzionalizzate, che si proponevano di affrontare dichiaratamente l'aspetto metodologico delle applicazioni dell'informatica agli studi umanistici: il Gruppo Interdisciplinare di Ricerche per la Computerizzazione dei Segni dell'Espressione (GIRCSE) dell'Università Cattolica di Milano (1982); il Gruppo di ricerca «Informatica e Discipline Umanistiche» dell'Università di Roma La Sapienza (1983); il Gruppo nazionale di coordinamento del CNR per le «Concordanze della Lingua Italiana Poetica dell'Otto/Novecento» (CLIPON, 1983-1984), il Centro Interdipartimentale di Servizi Informatici per le Scienze Sociali e Psicologiche e le Discipline Umanistiche dell'Università di Torino (CISI, 1987). Tra i contributi più significativi che questi gruppi di ricerca hanno arrecato a un corretto sviluppo degli studi di informatica umanistica, sono da segnalare alcuni convegni o incontri di studio che sanciscono una nuova tendenza nei contributi presentati dai partecipanti: quella di privilegiare l'aspetto metodologico delle loro ricerche, dando per scontati o ponendo in secondo piano gli elementi di carattere applicativo[7]. Contemporaneamente, la stessa tendenza andava affermandosi in ambito internazionale.

[6] È da segnalare, a questo proposito, un Convegno Internazionale organizzato dalla Fondazione IBM Italia con il patrocinio dell'Accademia Nazionale dei Lincei, svoltosi a Roma nei giorni 2-3 dicembre 1993. Il Convegno ha esaminato i molteplici aspetti legati al problema della conservazione del patrimonio informativo, con particolare riferimento ai supporti, agli strumenti e alle tecniche che rendono l'informazione fruibile nel tempo. Gli *Atti* sono raccolti in Gregory T., Morelli M. (eds.) 1994, *L'eclisse delle memorie*, Roma-Bari, Laterza.

[7] Cfr.: Savoca 1986; Savoca 1989; Gallino 1991; Fondazione IBM Italia 1992; Orlandi 1993. Un'ampia e documentata rassegna delle attività di ricerca in Italia è

Tra i frutti più attesi e rilevanti della nuova e più attenta considerazione riservata agli aspetti teorici e metodologici, è da segnalare l'elaborazione di alcuni strumenti (manuali, bibliografie, programmi integrati) che, contribuendo alla sistemazione organica della materia, attestano il consolidamento di un settore di studio divenuto ormai peculiare, l'Informatica umanistica[8].

3. La codifica dei dati testuali

La seconda metà degli anni Ottanta segna anche una fase di approfondimento e di rielaborazione dei criteri di preparazione della *machine readable form* di un testo.

Alla consuetudine della «pre-edizione» del testo[9], basata su una prassi empirica, spesso personalizzata e funzionale al tipo di elaborazione prevista da un singolo studioso, subentrano concetti nuovi e forse più affinati: quelli di codifica[10], rappresentazione[11], *markup*[12], *tagging*[13], interpretazione[14] e arricchimento dei dati testuali[15], modello[16] o schema di codifica[17]; tutti tendono ad assicurare una corrispondenza rigorosa tra i dati testuali e la loro versione elettronica, e a facilitarne la decrittazione o la verifica anche da parte di altri studiosi e utenti.

Si cerca, insomma, di evidenziare i meccanismi soggiacenti a quel processo di organizzazione e codificazione dell'informazione veicolata dai testi "composti" a stampa, con implicazioni anche nello studio dei manoscritti e delle trascrizioni della lingua parlata.

È del tutto evidente per il lettore umano che i segni di interpunzione presenti in un testo, l'uso delle lettere maiuscole, la disposizione delle lettere

presentata in Spinosa 1990.

[8] Cfr., a titolo di esempio: Losano 1985-1986; Orlandi 1990; Gigliozzi 1993; Adamo 1994; Picchi 1989.

[9] Cfr. Zampolli A. 1975, *L'elaborazione elettronica dei dati linguistici. Stato delle ricerche e prospettive*, in Accademia Nazionale dei Lincei, *Colloquio sul tema: Le tecniche di classificazione e la loro applicazione linguistica (Firenze, 13 dicembre 1972)*, Roma, 29-30, 32-35.

[10] Cfr. Orlandi 1986.

[11] Cfr. Adamo 1987.

[12] Cfr.: Bryan 1988, 5; Adamo 1992, 364; Sperberg-McQueen, Burnard 1994, 13; Burnard 1995, 42.

[13] Cfr.: Hockey S.M., Walker D.E. 1993, *Developing effective resources for research on texts. Collecting texts, tagging texts, cataloguing texts, using texts, and putting texts in context*, «Literary and Linguistic Computing», 8, 4, 236-242; Sperberg-McQueen, Burnard 1994, 1.

[14] Cfr. EAGLES 1994, 4.

[15] Cfr. Sperberg-McQueen C. M., *Text Encoding and Enrichment*, in Lancashire 1991, 503.

[16] Cfr. Adamo 1987, 59; Ciotti 1994, 220-227.

[17] Cfr. Sperberg-McQueen C. M., *Text Encoding and Enrichment*, in Lancashire 1991, 503; Sperberg-McQueen, Burnard 1994, 1.

e delle parole nella pagina, ma anche gli spazi che separano le parole, l'uso del corsivo o di altri criteri di enfatizzazione, i titoli, la ripartizione in capitoli e paragrafi, i titoli correnti e la numerazione delle pagine sono elementi funzionali alla comprensione dell'informazione contenuta in un testo[18]. Occorre aggiungere che un'attenzione sempre maggiore è riservata all'analisi e alla rappresentazione di quegli elementi che consentono di ricostruire la tradizione dei testi manoscritti e la storia dei testi a stampa, con particolare riferimento agli studi di bibliografia materiale: il supporto scrittorio, gli inchiostri, la dimensione della pagina e la riquadratura della composizione, il tipo e la misura dei caratteri.

Si è affermato che codificare un testo per l'elaborazione elettronica consista, in linea di principio, nel trascrivere un manoscritto dalla *scriptio continua*, ovvero nel rendere esplicito ciò che è congetturale o implicito, insomma nell'indirizzare l'utente su come il contenuto del testo dovrebbe essere interpretato[19]. A mio avviso, si tratta di un'operazione più complessa, proprio in virtù del fatto che la stampa è intervenuta, producendo un processo di organizzazione – e quindi di interpretazione – dell'informazione veicolata dal testo.

Ritengo quindi che l'aspetto più delicato nel "comporre" l'edizione informatica di un testo consista nel preservare quel deposito di informazioni inerenti il supporto tipografico e la veste editoriale del testo in questione, neutralizzandone piuttosto la staticità (questo però è vantaggio implicito nella dinamicità e plasticità del metodo informatico che si intende adottare), ma soprattutto corredando i dati testuali di quelle informazioni che appaiono immediate al lettore umano e che, in mancanza di un'adeguata segnalazione, finirebbero per perdersi in un mare di byte, ovvero farebbero regredire il testo a una forma di *scriptio continua*, seppure magnetica. Mi riferisco, a titolo di esempio, alle citazioni esplicite o implicite di altre opere, ai nomi di persona e di luogo, alle date, ai refusi dell'edizione a stampa, alle parole di espressione linguistica diversa da quella del testo, e quant'altro. Ben si presta, a questo proposito, il criterio di modularità della codifica[20] che informa le *Guidelines* della Text Encoding Initiative, grazie al quale è possibile tralasciare, o rinviare a fasi successive della codifica, gli elementi che non interessano immediatamente lo studioso che la compie.

4. Text Encoding Initiative (TEI)

Nel novembre 1987, la Association for Computers and the Humanities organizza un Convegno al Vassar College (Poughkeepsie, N. Y.), con l'intento

[18] Cfr. Sperberg-McQueen, Burnard 1994, 13. Si vedano anche: Alinei M.L. 1968, *Spogli elettronici dell'italiano delle origini e del Duecento. II: Forme. 1: Prose fiorentine*, The Hague, xlix-l; Lana 1994, 58 ss.

[19] Cfr. Sperberg-McQueen, Burnard 1994, 13, riportato anche in Lana 1994, 94 nota 35.

[20] Cfr. Sperberg-McQueen, Burnard 1995, 18.

di individuare un formato per lo scambio di testi in versione elettronica, raccogliendo la documentazione relativa ai più significativi schemi di codifica esistenti ed elaborando raccomandazioni per la codifica di nuovi materiali testuali[21]. L'istanza primaria del dibattito concerne l'individuazione delle caratteristiche da codificare e il modo con cui rappresentarle. Vengono sanciti alcuni princìpi fondamentali per lo sviluppo di una metodologia di codifica e di scambio dei dati linguistici e letterari, i cosiddetti «princìpi di Poughkeepsie»[22].

Si tratta di una sorta di atto di nascita o, se si vuole, del concepimento di un progetto internazionale di ricerca della durata di quattro anni, la TEI, che prende avvio nel mese di giugno 1988, sotto il patrocinio dell'Association for Computers and the Humanities (ACH), dell'Association for Computational Linguistics (ACL) e dell'Association for Literary and Linguistic Computing (ALLC), in parte finanziato dallo United States National Endowment for the Humanities (NEH), dalla Direzione XIII della Commissione della Comunità Europea, dalla Andrew W. Mellon Foundation e dal Social Science and Humanities Research Council canadese. Gli obiettivi principali del progetto possono essere così riassunti:

1. consentire lo scambio di dati testuali fra singoli studiosi, istituzioni di ricerca, sistemi diversi di elaborazione;
2. favorire il trattamento di dati secondo un formato indipendente da software e hardware utilizzati;
3. mettere a disposizione uno strumento che serva di guida nelle operazioni di codifica o di «cattura» di testi in *machine-readable form*.

La TEI, nata all'interno della comunità scientifica internazionale per soddisfare le necessità della ricerca, si rivolge al più ampio pubblico di utenti (primi fra tutti biblioteche, archivi, case editrici e centri di documentazione) con l'obiettivo di rendere efficace e consistente lo scambio di materiali su supporto elettronico[23], anche attraverso contatti e scambi con altre iniziative e progetti correlati[24].

Nel luglio 1990, il lavoro compiuto viene pubblicato in versione provvisoria con il titolo *Guidelines for the Encoding and Interchange of Machine-Readable Texts* (TEI P1). L'elaborazione dei commenti e delle osservazioni ricevute dà luogo a una seconda versione pubblicata nell'aprile del 1992 (TEI P2), nella quale confluiscono anche i contributi dei gruppi di lavoro costituiti

[21] Cfr.: Genet, Zampolli 1992, 87; Ide, Sperberg-McQueen 1995, 5.

[22] Cfr.: Ide, Sperberg-McQueen 1995, 6; Sperberg-McQueen, Burnard 1994, 10.

[23] Cfr.: Ide, Sperberg-McQueen 1995, 7; Sperberg-McQueen, Burnard 1995, 17-18.

[24] ISO, HyTime, EAGLES, ACL Data Collection Initiative, European Corpus Initiative, Network of European Research Corpora, Consortium for Lexical Research, Coalition for Networked Information, Center for Electronic Texts in the Humanities, Text Software Initiative: cfr. Ide, Sperberg-McQueen 1995, 14.

nei due anni precedenti. L'ultima versione, considerata non più provvisoria, ma comunque soggetta ad approfondimenti destinati a non alterarne l'impianto, è pubblicata nell'aprile 1994 con il titolo *Guidelines for Electronic Text Encoding and Interchange* (TEI P3): si tratta di due volumi, di complessive 1290 pagine, che raccolgono il lavoro più significativo finora compiuto in ambito internazionale a proposito della codifica informatica di testi. Le *Guidelines*, più che un vero e proprio standard, costituiscono un complesso di linee guida per la codifica di materiali linguistici e letterari e rappresentano una convenzione metodologica per descrivere la struttura fisica e logica e le caratteristiche di un'ampia gamma di dati testuali: testi in prosa e in versi, opere teatrali, dizionari e repertori terminologici, trascrizioni di testi parlati, edizioni critiche, fonti storiche, tabelle e grafici.

La maggiore garanzia di solidità dell'impianto complessivo è assicurata dall'adozione dello Standard Generalized Markup Language (SGML), proposto da Charles Goldfarb negli anni Settanta e perfezionato come standard ISO nel 1986. SGML è un metalinguaggio che consente di descrivere la struttura di un testo, anche mediante la rappresentazione di determinate caratteristiche testuali[25]. L'orientamento per la *codifica descrittiva* più che per quella procedurale, la classificazione e l'identificazione di un *tipo di documento*, e l'*indipendenza* da ogni sistema di elaborazione costituiscono i tratti distintivi di SGML. In particolare:

1. una codifica descrittiva consiste nel segnalare, mediante una denominazione predeterminata, la rappresentazione di una certa caratteristica di un testo (per esempio, una parola evidenziata all'interno del testo): appositi programmi di gestione, non direttamente legati a SGML ma dipendenti dalla macchina e dall'ambiente operativo effettivamente utilizzati, si fanno carico di tradurre quella segnalazione in un comando che consenta di ottenere la rappresentazione grafica richiesta.
2. Per quanto riguarda la definizione dei tipi di documento (*Document Type Definition*, DTD), SGML prevede che per ciascun documento debba essere individuata una struttura logico-formale che consenta di descriverne l'articolazione e quindi di attribuirne l'appartenenza a una determinata tipologia (si pensi a un'entrata di dizionario, a una citazione bibliografica, a un articolo scientifico, a un volume a stampa), all'interno della quale potranno, per esempio, essere predisposte verifiche automatiche della coerenza dei singoli documenti o essere messe in relazione parti omologhe di documenti diversi.
3. L'indipendenza dei dati da hardware e software è assicurata mediante un meccanismo di dichiarazione dei caratteri presenti in ciascun documento; tale meccanismo consente di adottare algoritmi di sostituzione automatica

[25] Cfr.: Burnard 1995, 44; Ide, Sperberg-McQueen 1995, 10.

delle sequenze di caratteri, per soddisfare le esigenze dell'ambiente operativo nel quale debbono essere condotte le elaborazioni.

Facendo proprio l'impianto dello SGML, la TEI ha elaborato una serie di Definizioni di tipo di documento (DTD, cfr. il precedente punto 2.): una *DTD principale* per la descrizione dei testi, e varie *DTD ausiliarie* per la codifica di meta-informazioni relative alle diverse tipologie di testi considerate. La DTD principale è definita come «un insieme di insiemi di *tag* (unità di marcatura)», che possono essere usati secondo le più svariate combinazioni, in maniera analoga a quello che avviene in uno schema complesso di una base di dati, da cui ciascun utente può costruire una *vista* personalizzata per rispondere alle esigenze delle proprie interrogazioni o elaborazioni dei dati[26].

Non si può negare l'impressione di una difficoltà complessiva nell'approccio allo SGML e alle *Guidelines* della TEI. Gli stessi Curatori, che pure dichiarano la necessità di un ulteriore lavoro di approfondimento e di verifica pratica dei criteri enunciati, mostrano di esserne consapevoli e sostengono, per contro, che l'interpretazione dei dati è compito istituzionale degli studiosi, troppo spesso fuorviati dall'uso di strumenti appariscenti e sofisticati, che per gran parte si limitano alla mera riproduzione grafica dei testi[27]. Vi è poi una fondata obiezione in merito alla difficoltà e ai margini di soggettività inerenti l'individuazione di una "struttura" del testo preso in esame: si tratta di un'operazione più agevole da compiere su un testo elaborato *ex novo*, e che comunque richiede una conoscenza assai profonda della realtà testuale da rappresentare.

Sono tuttavia da notare anche i vantaggi considerevoli insiti nella codifica descrittiva proposta dalle *Guidelines*: la possibilità di ottenere una verifica automatica della congruità di ciascun documento con il modello di struttura predefinito, e la possibilità di elaborare i dati con un livello di precisione e di affinamento maggiore di quello che consentirebbe la loro semplice trascrizione. D'altra parte, l'accoglienza e il favore crescente riservati all'iniziativa inducono a prevederne una diffusione sempre più capillare[28], soprattutto con la progettazione e la diffusione di software conforme ai criteri SGML. È opportuno segnalare inoltre che recentemente si è reso disponibile un nuovo strumento di introduzione alla TEI, forse più agevole e diretto. L'annata 29 (1995) di «Computers and the Humanities» dedica infatti i primi tre fascicoli, per un totale di 231 pagine, al lavoro svolto dalla TEI, con l'intento di divulgare quanto non poteva essere incluso nei due già ponderosi volumi delle *Guidelines*: spiegazioni, esemplificazioni e soprattutto gli aspetti più significativi della discussione condotta in seno alle diverse Commissioni nelle

[26] Cfr. Sperberg-McQueen, Burnard 1995, 27.

[27] Cfr. Burnard 1995, 49-50.

[28] Cfr.: Genet, Zampolli 1992, 7, 87-88; Lana 1994, 60, 66-67; Lancashire 1991, 503-507, 547; Marcos Marín 1994, 89-109; Scolari 1995, 116-132.

quali si è articolata l'attività della TEI. Si è voluto, insomma, dare spazio ai problemi dibattuti, all'esperienza di lavoro e al contesto complessivo all'interno del quale sono state prodotte le norme proposte[29].

5. Expert Advisory Group on Language Engineering Standards (EAGLES)

EAGLES è un progetto intrapreso nel febbraio 1993 nell'ambito del programma Ricerca e ingegneria linguistica della Direzione XIII della Commissione dell'Unione Europea, che ne assicura il finanziamento, con l'obiettivo di predisporre standard per l'elaborazione e lo scambio di dati testuali nelle seguenti aree: *corpora* testuali, lessici computazionali, formalismi grammaticali, lingua parlata. Vi sono coinvolti più di trenta centri di ricerca, organizzazioni industriali, associazioni professionali e reti, articolate in oltre cento unità. Il coordinamento è affidato al Consorzio Pisa Ricerche. È prevista la cooperazione anche con organismi internazionali. Il presupposto è costituito dalla riutilizzazione dei materiali linguistici provenienti dall'attività di ricerca o industriale, nel quadro dello sviluppo della comunicazione tra le differenti aree linguistiche europee. Al termine della seconda fase di attività, entro l'anno corrente, è annunciato un *EAGLES Handbook* che si propone come primo tentativo di stabilire linee-guida, raccomandazioni e standard nei settori interessati all'ingegneria linguistica.

Per quanto attiene più specificamente alla codifica dei testi, è allo studio un *Corpus Encoding Standard* destinato all'interscambio di dati provenienti da *corpora* testuali, che verrà formulato come applicazione dello SGML conformemente ai criteri della TEI. Lo schema proposto dalla TEI non è tuttavia ancora sufficientemente verificato su grandi *corpora* linguistici (in particolare multilingui) e, non rientrando negli obiettivi della TEI, non fornisce raccomandazioni su quali siano gli elementi da codificare in un *corpus*.

Lo schema di codifica previsto da EAGLES considera tre livelli di standardizzazione interrelati: ciascun livello implica l'adozione dei criteri stabiliti per il livello precedente e influenza le operazioni di codifica del livello successivo[30]. I dati sono passibili di codifica a livello di metalinguaggio, sintattico e semantico.

1. Lo standard di codifica a livello di metalinguaggio ha il compito di definire il formalismo delle regole sintattiche e degli schemi di marcatura dei documenti, senza fornire specificazioni in merito al *markup* stesso (per esempio, i nomi dei *tag*). SGML è uno standard a questo livello: la sintassi concreta di riferimento definisce le forme di *tag*, il set di caratteri, le regole di denominazione, le parole riservate, le caratteristiche ammesse (per esempio, l'omissione dei *tag* di chiusura).

[29] Ide N., Véronis J. 1995, *Introduction*, «Computers and the Humanities», 29, 1, 3.
[30] EAGLES 1994, 6-7, 11-12.

2. A livello sintattico, il meccanismo delle definizioni del tipo di documento (DTD) consente all'utente di definire i nomi di *tag* e i modelli di documento che specificano le relazioni tra i *tag*, e quindi di utilizzare uno standard sintattico di codifica, la cui coerenza può essere verificata mediante algoritmi di *parsing*.
3. La codifica a livello semantico introduce un margine di soggettività più elevato, determinato dalla difficoltà di verificare la correttezza nell'applicazione di un dato *tag* da parte dell'utente.

6. Unicode

Su un piano completamente diverso, sia dal punto di vista metodologico sia da quello dei risultati, si situa l'iniziativa Unicode, che pure sembra destinata a suscitare notevole interesse, perché si propone di affrontare e risolvere gli annosi problemi connessi con i cosiddetti codici standard per la rappresentazione dei caratteri alfanumerici. L'Unicode character encoding standard (o più semplicemente Unicode standard) nasce infatti come codice per la rappresentazione degli alfabeti internazionali e si propone di codificare tutti i caratteri utilizzati per la comunicazione scritta, di epoca moderna e storica[31].

Il progetto si avvia nel 1988, quando un gruppo di professionisti dell'informazione si trova a concordare sul fatto che nessuna metodologia di codifica dei caratteri in uso nel settore dell'elaborazione multilingue possiede l'immediatezza e la semplicità del codice ASCII. Essi propongono allora di adottare un'architettura di codifica a 16 bit che consenta di ottenere un numero di codici univoci sufficiente a rappresentare i caratteri e i simboli tecnici di uso comune e che, allo stesso tempo, faciliti la progettazione di un nuovo codice efficiente e flessibile[32].

Nel gennaio 1991 il Consorzio Unicode si trasforma nella Unicode Inc., organizzazione senza fine di lucro, che raccoglie le maggiori ditte e istituzioni attive nel settore dell'informatica a livello mondiale, con l'obiettivo comune di contribuire alla progettazione, all'implementazione, al mantenimento e alla promozione dello standard Unicode.

7. La sfida della codifica

In conclusione, desidero sottolineare che tutti gli aspetti fin qui trattati costituiscono soltanto una parte dei problemi che lo studioso umanista si trova ad affrontare nella delicata fase della trasmissione di informazione te-

[31] The Unicode Consortium 1991-1992. La versione 1.0 rappresenta un notevole contributo al lavoro di definizione dello Standard ISO/IEC 10646-1: 1993, che ne ha determinato il superamento.

[32] ISO/IEC 10646-1: 1993.

stuale, nel passaggio dall'edizione a stampa a quella informatica. Per una trattazione più ampia e articolata segnalo il saggio recentemente pubblicato da Tito Orlandi, *Alla base dell'analisi dei testi: il problema della codifica*[33]. Dalla sua lettura si è indotti a osservare come la codifica costituisca per lo studioso umanista una sfida antica quanto la scrittura: l'informatica costituisce un'occasione per riproporla alla nostra considerazione. Occorre infine constatare come si avverta ancora la necessità di un approfondimento teorico complessivo in grado di trascendere o meglio di assicurare un adeguato fondamento anche alle implicazioni pratiche connesse con la circolazione e lo scambio dei dati testuali in edizione informatica.

Giovanni Adamo
Lessico Intellettuale Europeo - CNR

BIBLIOGRAFIA

Testi e informatica

Bender T.K. 1976, *Literary texts in electronic storage: The editorial potential*, «Computers and the Humanities», 10, 4, 193-200

Wittig S. 1978, *The computer and the concept of text*, «Computers and the Humanities», 11, 4, 211-215

La situazione italiana

Adamo G. 1994, *Bibliografia di Informatica umanistica*, Roma, Bulzoni

Bartoletti Colombo A.M. (ed.) 1977-1979, *Legum Iustiniani Imperatoris Vocabularium. Novellae. Pars latina*, Milano, Cisalpino - La Goliardica, 11 voll.

Bartoletti Colombo A.M. (ed.) 1986-1989, *Legum Iustiniani Imperatoris Vocabularium. Novellae. Pars graeca*, Milano, Cisalpino - La Goliardica, 7 voll.

Fondazione IBM Italia 1992, *Calcolatori e scienze umane*, a cura di Marcello Morelli, Milano, Etas Libri

Gallino L. (ed.) 1991, *Informatica e scienze umane. Lo stato dell'arte*, Milano, Franco Angeli

Gigliozzi G. 1993, *Letteratura modelli e computer*, Roma, EUROMA-La Goliardica

Losano M.G. 1985-1986, *Informatica per le scienze sociali. Corso di informatica giuridica*, Torino, Einaudi, 3 voll.

Mordenti R. (ed.) 1990, *Bartolomeo Cerretani, «Dialogo della mutatione di Firenze». Edizione critica secondo l'apografo magliabechiano*, Roma, Edizioni di Storia e Letteratura

Orlandi T. 1990, *Informatica Umanistica*, Roma, La Nuova Italia Scientifica

Orlandi T. (ed.) 1993, *Discipline umanistiche e informatica. Il problema dell'integrazione (Roma, 8 ottobre 1991)*, Roma, Accademia Nazionale dei Lincei

Picchi E. 1989, *DBT. Data Base Testuale*, Pisa, Istituto di Linguistica Computazionale - CNR

Savoca G. (ed.) 1986, *Lessicografia, filologia e critica. Atti del Convegno Internazionale di Studi (Catania-Siracusa, 26-28 aprile 1985)*, Firenze, Leo S. Olschki

[33] Orlandi 1995 (cfr. Bibliografia: *La codifica dei dati testuali*).

SAVOCA G. (ed.) 1989, *Per la lingua di Montale. Atti dell'incontro di studio (Firenze, 26 novembre 1987), con appendice di liste alla concordanza montaliana*, Firenze, Leo S. Olschki

SPINOSA G. (ed.) 1990, with the collaboration of L. Farina, *Special Double Issue: Humanities Computing in Italy*, «Computers and the Humanities», 24, 5-6, 337-493

STOPPELLI P., PICCHI E. (eds.) 1993, *LIZ - Letteratura Italiana Zanichelli. CD-ROM dei testi della Letteratura italiana*, Bologna, Zanichelli

La codifica dei dati testuali

ADAMO G. 1987, *La codifica come rappresentazione. Trasmissione e trattamento dell'informazione nell'elaborazione automatica di dati in ambito umanistico*, in G. GIGLIOZZI (ed.), *Studi di codifica e trattamento automatico di testi*, Roma, Bulzoni, 39-63

ADAMO G. 1992, *Analisi informatica di testi: problemi e prospettive*, in *Calcolatori e scienze umane*, Milano, Etas Libri, 350-365

BURNARD L. 1995, *What is SGML and how does it help?*, «Computers and the Humanities», 29, 1, 41-50

CIOTTI F. 1994, *Il testo elettronico: memorizzazione, codifica ed edizione*, in C. LEONARDI, M. MORELLI, F. SANTI (eds.), *Macchine per leggere. Tradizioni e nuove tecnologie per comprendere i testi*, Spoleto, Centro Italiano di Studi sull'Alto Medioevo, 213-230

GIGLIOZZI G. (ed.) 1987, *Studi di codifica e trattamento automatico di testi*, Roma, Bulzoni

LANA M. 1994, *L'uso del computer nell'analisi dei testi*, Milano, Franco Angeli

MORDENTI R. 1987, *Appunti per una semiotica della trascrizione nella procedura ecdotica computazionale*, in G. GIGLIOZZI (ed.), *Studi di codifica e trattamento automatico di testi*, Roma, Bulzoni, 85-124

ORLANDI T. 1986, *Problemi di codifica e trattamento informatico in campo filologico*, in G. SAVOCA (ed.), *Lessicografia, filologia e critica*, Firenze, Olschki, 69-82

ORLANDI T. 1995, *Alla base dell'analisi dei testi: il problema della codifica*, in G. DEGLI ANTONI et Al., *Scrivere comunicare apprendere con le nuove tecnologie*, a cura di Mario Ricciardi, Torino, Bollati Boringhieri, 69-86

SPERBERG-MCQUEEN C.M. 1991, *Text in the electronic age. Textual study and text encoding, with examples from medieval texts*, «Literary and Linguistic Computing», 6, 1, 34-46

Text Encoding Initiative (TEI)

BRYAN M. 1988, SGML. *An Author's Guide to the Standard Generalized Markup Language*, Wokingham, Addison-Wesley

BURNARD L. 1995, *What is SGML and how does it help?*, «Computers and the Humanities», 29, 1, 41-50

GENET J.-PH., ZAMPOLLI A. 1992, *Computers and the Humanities*, Aldershot, Dartmouth

GOLDFARB CH.F. 1990, *The SGML Handbook*, Oxford, Clarendon Press

HERWIJNEN E. VAN 1994², *Practical SGML*, Boston, Kluwer

IDE N.M., VÉRONIS M. (eds.) 1995, *The Text Encoding Initiative: Background and Context*, «Computers and the Humanities», 29, 1, 1-98; 2, 99-179; 3, 181-231

IDE N.M., SPERBERG-MCQUEEN C.M. 1995, *The TEI: History, goals and future*, «Computers and the Humanities», 29, 1, 5-15

ISO 8879: 1986, *Information processing - Text and office systems - Standard Generalized Markup Language (SGML)*, Genève, ISO

ISO/TR 9573: 1988, *Information processing - SGML support facilities - Techniques for using SGML*, Genève, ISO

LANA M. 1994, *L'uso del computer nell'analisi dei testi*, Milano, Franco Angeli

LANCASHIRE I. 1991, *The Humanities Computing Yearbook 1989-90. A Comprehensive Guide to Software and Other Resources*, Oxford Clarendon Press

MARCOS MARÍN F.A. 1994, *Informática y humanidades*, Madrid, Editorial Gredos

SCOLARI A. 1995, *Gli standard OSI per le biblioteche. Dalla biblioteca-catalogo alla biblioteca-nodo di rete*, Milano, Editrice Bibliografica

SPERBERG-MCQUEEN C.M., BURNARD L. (eds.) 1990, *Guidelines for the Encoding and Interchange of Machine-Readable Texts (TEI P1)*, Chicago-Oxford, Text Encoding Initiative

SPERBERG-MCQUEEN C.M., BURNARD L. (eds.) 1992, *Guidelines for the Encoding and Interchange of Machine-Readable Texts (TEI P2)*, Chicago-Oxford, Text Encoding Initiative

SPERBERG-MCQUEEN C.M., BURNARD L. (eds.) 1994, *Guidelines for Electronic Text Encoding and Interchange (TEI P3)*, Chicago-Oxford, Text Encoding Initiative, 2 voll.

SPERBERG-MCQUEEN C.M., BURNARD L. 1995, *The design of the TEI encoding scheme*, «Computers and the Humanities», 29, 1, 17-39

Expert Advisory Group on Language Engineering Standards (EAGLES)

CALZOLARI N., MCNAUGHT J. 1994, *EAGLES Interim Report. Editors' Introduction* (EAG-EB-IR-2), Pisa, ILC-CNR.

EAGLES 1994, *Corpus Encoding. Draft - Work in progress* (EAG-CSG/IR-T2.1), Pisa, ILC-CNR.

Unicode

ISO/IEC 10646-1: 1993, *Information Technology - Universal Multiple-Octet Coded Character Set (UCS) - Part 1: Architecture and Basic Multilingual Plane*, Genève, ISO.

THE UNICODE CONSORTIUM 1991-1992, *The Unicode Standard. Worldwide Character Encoding. Version 1.0*, Reading (Mass.), Addison-Wesley, 2 voll.

ABSTRACT

The paper investigates the dynamic and multidimensional valency assumed by the representation of texts in machine readable form. Indeed, computer methodologies release each text from the static bonds of paper printing, and make it possible to read a text in new ways. In particular, the paper points out: 1. the need for a metalanguage to describe the information elements of a text, enabling computer processing for the purpose of linguistic and literary analysis, and for lexical, conceptual and terminological documentation; 2. the function of the international encoding standards for texts and corpora in machine readable form (hardware- and software - independent); 3. the proposals formulated on the basis of the Standard Generalized Markup Language (SGML, ISO 8879).