

ACCADEMIA NAZIONALE DEI LINCEI

ANNO CDIII - 2006

CONTRIBUTI DEL
CENTRO LINCEO INTERDISCIPLINARE
«BENIAMINO SEGRE»

N. 114

CONVEGNO INTERNAZIONALE

organizzato in collaborazione con
ERPANET e la Fondazione Ezio Franceschini

ARCHIVI INFORMATICI PER IL PATRIMONIO CULTURALE

(Roma, 17-19 novembre 2003)

ESTRATTO



ROMA 2006
BARDI EDITORE
EDITORE COMMERCIALE

DINO BUZZETTI*

BIBLIOTECHE DIGITALI E OGGETTI DIGITALI COMPLESSI ESAUSTIVITÀ E FUNZIONALITÀ NELLA CONSERVAZIONE

1. Il titolo della mia comunicazione può essere fuorviante. Si tratta infatti di un titolo tecnico e molto concreto, mentre le considerazioni che svolgerò sono di natura piuttosto astratta e concettuale. La preoccupazione maggiore dalla quale queste considerazioni sono ispirate è quella di tener fede al titolo di questa prima sezione del convegno, *Qualità dell'archivio e conservazione dell'informazione*. Mi sono quindi proposto di valutare quali sono le condizioni per la conservazione dell'informazione nel processo di trasferimento dei documenti e delle conoscenze in ambiente digitale. Ho parlato di 'documenti e conoscenze' perché a mio avviso il problema non è solo quello della digitalizzazione, intesa come mera riproduzione o replica digitale del documento, ma è anche e soprattutto quello della rappresentazione adeguata del suo contenuto informativo. Per chiarire il problema, vorrei fare immediatamente un esempio, che traggo dalla *Gesamtausgabe* delle opere di Martin Heidegger (1973) (fig. 1).

Nel volume su *Kant e il problema della metafisica* è inserita la riproduzione di uno schizzo manoscritto di Heidegger sul contenuto del volume e, nella pagina accanto, una sua trascrizione diplomatica (fig. 2).

Ora, perché si è sentita la necessità di inserire, in un'edizione come questa – un'edizione filologicamente e criticamente accurata – la riproduzione del manoscritto? Non discuto, qui, della sua importanza filologica. Ciò su cui voglio richiamare l'attenzione è la presenza, affiancata, della riproduzione e della sua trascrizione diplomatica. Credo infatti che questo esempio metta in luce, in modo molto diretto, quali sono le due condizioni essenziali per la

* Dipartimento di Filosofia - Università degli Studi di Bologna - Via Zamboni, 38 - 40126 BOLOGNA.
buzzetti@philo.unibo.it

conservazione dell'informazione: l'esaustività e la funzionalità. La riproduzione del manoscritto ci richiama all'esaustività della conservazione. Tutta, o quasi, l'informazione è contenuta nella riproduzione del documento originale. A poco servirebbe però la riproduzione se non fossero al contempo forniti strumenti per la sua decifrazione e interpretazione. E a ciò serve la trascrizione diplomatica. Ho parlato di trascrizione diplomatica, o di tipo diplomatico; ma in realtà, a ben vedere, la sua funzione non è quella di riprodurre l'originale, quanto quella di analizzarne l'informazione: presentarla in una forma che ne consenta la lettura e l'interpretazione. La trascrizione è in funzione di un'operazione analitica eseguita sull'informazione contenuta nel documento. Si pensi ora alla trasposizione dell'esempio mostrato al caso della rappresentazione digitale. Nel caso mostrato, la trascrizione è al servizio di un'operazione mentale di lettura e di interpretazione, nel caso della rappresentazione digitale, la trascrizione dovrebbe servire piuttosto all'elaborazione automatica del contenuto informativo del documento. Una cosa però dovrebbe essere chiara e tenuta ben ferma: nell'edizione considerata, la riproduzione e la trascrizione sono presentate *assieme*; si è cioè avvertita la necessità di presentare entrambe le rappresentazioni. Senza la funzionalità, l'esaustività della riproduzione non serve a molto, o quanto meno serve poco all'elaborazione automatica del contenuto informativo, e la funzionalità della trascrizione, senza una riproduzione esaustiva, non ci restituisce l'integrità dell'originale e delle conoscenze.

2. Come si è cercato di rispondere a queste due esigenze, almeno nel caso dell'informazione testuale? Dal punto di vista strettamente informatico, il testo è una pura sequenza di caratteri codificati. Il testo è cioè un tipo di dato, ossia una forma particolare di rappresentazione dell'informazione. L'informazione è rappresentata come una pura sequenza di caratteri alfanumerici. Ma questa forma di rappresentazione è espressivamente troppo povera per restituire tutta l'informazione testuale contenuta in un documento stampato o manoscritto – si pensi alla pura sequenza dei caratteri, senza i segni diacritici e senza l'impaginazione, nell'esempio qui fornito. Per rispondere alla condizione di esaustività nella rappresentazione testuale, è stato introdotto il *markup* – mi sia permesso di usare quest'espressione perché il termine italiano «codifica» è ambiguo e può riferirsi tanto alla codifica binaria dei caratteri quanto alla codifica del testo, ossia all'inserimento di marcatori (*tags*) all'interno di una stringa di caratteri, per integrare la rappresentazione dell'in-

formazione testuale. La definizione di metodi adeguati per la rappresentazione e la codifica dell'informazione testuale ha lungamente impegnato gli studiosi ed è approdata alla pubblicazione delle *Guidelines* della Text Encoding Initiative (TEI) (Sperberg-McQueen e Burnard, 1994). Ho richiamato cose molto note per insistere su un punto. Lo strumento tecnico scelto a suo tempo dalla TEI per la definizione di linguaggi di *markup*, ovvero linguaggi di marcatura del testo, fu lo Standard Generalized Markup Language (SGML), sostituito in seguito dal suo derivato, l'eXtensible Markup Language (XML). Ora, è stato più volte osservato che lo SGML è un linguaggio per la rappresentazione dei dati, ma è privo di una semantica (Raymond *et al.*, 1996: 25-36; Renear *et al.*, 2003a: 303-305), ossia non definisce un insieme di operatori algebrici in grado di operare su quei dati. In linguaggio meno tecnico, questo significa che una rappresentazione SGML/XML è in grado di definire la struttura di un documento, ma non è in grado di definire operatori che ne elaborino il contenuto informativo. È quindi possibile affermare che la codifica SGML/XML soddisfa, se mai, il requisito dell'eshaustività, ma non risponde in modo soddisfacente all'esigenza di funzionalità, ossia di elaborazione del contenuto del documento. La difficoltà è avvertita dagli stessi promotori della Text Encoding Initiative, alcuni dei quali hanno promosso un progetto di ricerca per lo sviluppo di una semantica formale per il *markup* in grado di migliorare la funzionalità e l'interoperabilità della rappresentazione digitale dei documenti (Renear *et al.*, 2003b). Su questo progetto ritornerò con alcune considerazioni più oltre.

Ma che cosa si può dire, intanto, a proposito della funzionalità? L'idea di funzionalità è connessa con l'idea dell'operabilità e nel caso della rappresentazione digitale comporta la possibilità della sua elaborazione automatica. A questo proposito, occorre fare qualche precisazione. Una rappresentazione digitale in sé è, in quanto tale, elaborabile automaticamente, ma si tratta di stabilire, di volta in volta, qual è il tipo di informazione di cui si tratta, ossia qual è esattamente l'informazione associata al dato elaborato. Consideriamo le due forme principali di rappresentazione digitale dei documenti, il testo e l'immagine. L'immagine digitale può essere elaborata, ma ciò che si elabora è l'informazione visiva, la gradazione del colore, per esempio, ma non ciò che l'oggetto colorato rappresenta. Sono note le difficoltà di riconoscimento degli oggetti attraverso un'analisi automatica delle immagini digitali, o della lettura automatica della scrittura corsiva. Nel caso del testo, a sua volta, si elaborano pure e semplici stringhe di caratteri – una ricerca per stringhe, per esempio,

restituisce gli omografi, indipendentemente dal loro significato; oppure si elabora l'aspetto tipografico del testo, cioè la sua impaginazione. Dunque ciò che si elabora, in realtà, è la forma esteriore della rappresentazione, non l'informazione rappresentata. Per rendere più precisa la distinzione, possiamo ricorrere, nel caso del testo, alla nota distinzione tra *espressione* e *contenuto*. Nel caso di una rappresentazione digitale, dobbiamo tenere presente la distinzione tra strutture di dati e modelli di dati: le *strutture di dati* sono le forme in cui si articola la rappresentazione simbolica dell'informazione, i *modelli di dati*, invece, definiscono le operazioni che si possono eseguire sui valori assegnati ad una certa rappresentazione simbolica dell'informazione. Per richiamare un esempio più intuitivo, si può ricordare che le leggi dell'aritmetica sono indipendenti dai sistemi di numerazione e di notazione numerica, binario, decimale, arabo, romano e così via (fig. 3).

Numerazione		Notazione
decimale	$4 + 2 = 6$	araba
binaria	$100 + 10 = 110$	araba
decimale	IV più II è uguale a VI	romana

Fig. 3.

In sostanza, occorre considerare se ciò che si elabora è la *sintassi* della rappresentazione o piuttosto la sua *semantica*.

Perché insistere su queste ovvie distinzioni? Per mostrare, ancora una volta, che nel caso della rappresentazione digitale esaustività e funzionalità della rappresentazione sono requisiti distinti e non sempre realizzabili congiuntamente. L'esigenza di esaustività va a scapito, spesso, della funzionalità: la rappresentazione concreta di un oggetto particolare può non essere immediatamente elaborabile secondo regole astratte, dipendenti dal tipo dell'informazione rappresentata. Viceversa, l'esigenza della funzionalità va a scapito dell'esaustività: per eseguire operazioni sull'informazione rappresentata, occorre fornirne un modello; e un modello è per definizione astratto e necessariamente non esaustivo. La difficoltà sembra connessa con le proprietà essenziali stesse degli oggetti culturali: gli oggetti culturali sono tali non per la loro

natura fisica, ma per la loro funzione sociale e il loro valore simbolico. Nel caso di una rappresentazione funzionale, si parte dalle esigenze di operatività e di elaborazione dell'informazione e in base a ciò si definiscono il modello e le forme di rappresentazione del contenuto informativo. La rappresentazione dell'oggetto procede scegliendone caratteristiche definite in modo astratto e selettivo. Si pensi allo schema entità-relazione-attributo secondo cui sono costruiti i *database* relazionali. Un conto è quindi inserire informazione testuale in un *database*, un altro conto presentarla semplicemente sul Web. Nel primo caso l'informazione può essere recuperata, con le procedure previste dal linguaggio di interrogazione del *database*, secondo determinati criteri semantici; nel secondo caso, l'informazione può essere recuperata dal testo visualizzato solo mediante un'operazione mentale di lettura. I *browser* comunemente usati consentono al massimo semplici operazioni di ricerca per stringhe non interpretate.

3. L'esigenza di una presentazione funzionale dell'informazione è stata chiaramente avvertita nelle forme più avanzate di progettazione di architetture per biblioteche digitali e nella definizione di oggetto digitale complesso. Il contributo fondamentale all'avvio della discussione su questi argomenti è costituito dallo studio di Robert Kahn e Robert Wilensky per un'infrastruttura in grado di collegare un'ampia gamma di servizi di informazione digitale distribuiti, tra i quali vengono annoverate anche, se non soprattutto, le biblioteche digitali vere e proprie (cfr. Kahn e Wilensky, 1995). Ciò che voglio mettere in evidenza è che, a partire da questa impostazione, si è venuto affermando il principio che «l'architettura deve distinguere chiaramente, tra gli oggetti digitali nella forma in cui sono creati dal produttore (*originator*), gli oggetti digitali archiviati in una raccolta, e gli oggetti digitali nella forma in cui sono distribuiti agli utenti»¹. Nelle biblioteche digitali si è così giunti ad una sempre crescente diffusione di oggetti digitali complessi che rispondono a svariate esigenze, dalla semplice aggregazione di formati diversi ad una più diretta associazione di informazioni supplementari e di servizi dinamici ai *file* tradizionali (Nelson *et al.*, 2001). In molti casi, dunque, l'oggetto restituito all'utente non è una pagina Web statica e fissa una volta per tutte, ma il risultato di un'elaborazione. L'oggetto presentato è costruito a partire dall'informazione archiviata nel sistema e può essere definito a partire da par-

¹ «The architecture must distinguish carefully between digital objects as they are created by an originator, digital objects stored in a repository, and digital objects as disseminated to a user» (Arms, 1995).

ticolari esigenze funzionali. In questo processo, però, l'identità dell'oggetto digitale non coincide necessariamente con l'identità del documento che riproduce l'originale. Nel Kahn-Wilensky Framework, l'identità dell'oggetto digitale è fissata da un «manico», o «appiglio» (viene chiamato genericamente *handle*), che consiste in una denominazione certificata e unica per l'oggetto in questione, indipendente dal luogo in cui è archiviato. L'oggetto digitale è una struttura complessa che può contenere, come sua parte, la riproduzione dell'oggetto originale. L'articolo di Renear *et al.* (2003a) può essere preso ad esempio per mostrare come possono essere distinte le due cose, l'oggetto digitale complesso che ne costituisce l'archiviazione funzionale (fig. 4)² e la riproduzione del documento fisico originale (fig. 5)³.

La riproduzione dell'originale – come si può vedere – non è dotata delle funzionalità che consentono l'attivazione dei *link* delle voci bibliografiche citate.

Un esempio più pertinente alla conservazione del patrimonio culturale è costituito dalla raccolta dei *Codices Electronici Ecclesiae Coloniensis*⁴. In questo sistema le descrizioni dei codici sono prodotte dinamicamente ed è implementato un sistema di identificazione dei diversi tipi di riproduzione dei singoli fogli – si tratta di immagini digitali diverse, ciascuna con diverso valore di risoluzione. Lo schema di identificazione proposto, definito a partire dall'esigenza di mantenere il sistema tradizionale di citazione dei materiali originali, è stato dettagliatamente descritto e giustificato da Manfred Thaller (2001) in un articolo apparso su *D-Lib Magazine*. Quello che vorrei qui fare osservare è che di ogni singola pagina manoscritta vengono fornite diverse riproduzioni digitali, ciascuna rispondente a esigenze funzionali distinte: un'immagine ad altissima risoluzione, per esempio, consente la lettura e l'analisi, anche automatica, di certi particolari che le immagini a risoluzione inferiore non consentono. Si può quindi osservare, in conclusione, che i requisiti combinati dell'eshaustività e della funzionalità producono, per lo stesso oggetto culturale, rappresentazioni digitali distinte e di forma diversa, eventualmente aggregate in oggetti digitali complessi, e fanno sorgere delicati problemi di identità.

² <http://portal.acm.org/citation.cfm?id=827192>.

³ http://portal.acm.org/ft_gateway.cfm?id=827192&type=pdf&coll=GUIDE&dl=GUIDE&CFID=14057334&CFTOKEN=5744073.

⁴ Cfr. <http://www.ceec.uni-koeln.de>

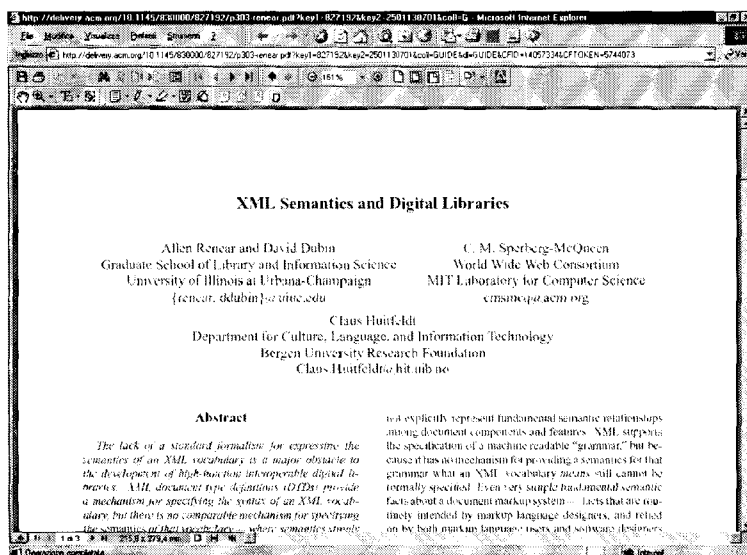


Fig. 5.

4. Mette conto sul tipo di identità definita di volta in volta per gli oggetti digitali, nella prospettiva che qui è stata assunta e che tende a valutare l'eshaustività e la funzionalità della rappresentazione digitale degli oggetti culturali. Nello schema Kahn-Wilensky è prevista una registrazione (*log*) delle modifiche apportate all'oggetto digitale e il riferimento è alla struttura funzionale dei dati archiviati, più che a una particolare forma di restituzione dell'oggetto presentato all'utente. Nel caso dei *Codices Electronici Ecclesiae Coloniensis*, l'identità è riferita alle diverse immagini digitali del documento originale. L'immagine mira all'eshaustività, che si realizza però a gradi diversi in relazione a diverse esigenze di lettura o di elaborazione funzionale. Ancora diverso è il riferimento scelto da Dubin e Renear – gli autori dell'articolo citato precedentemente – per stabilire l'identità di un documento costituito da una rappresentazione XML. In questo caso il documento digitale è definito come «un'espressione simbolica astratta che può essere ripetutamente esemplata in diversi *media*», anziché essere considerato come l'oggetto fisico concreto che ne costituisce un singolo esemplare (Renear e Dubin, 2003: 181). Per giungere a questa conclusione, gli autori muovono dall'osservazione che lo sviluppo del *markup* descrittivo ha permesso di identificare la struttura impli-

cita soggiacente al documento, struttura costituita da una «ordered hierarchy of content objects» (una «OHCO») (DeRose *et al.*, 1990: 6). Questa caratterizzazione del documento a partire dalla sua struttura astratta non è però sufficiente ad identificarlo in assenza di un metodo formale che ne specifichi la semantica. Di qui il progetto, già ricordato, per lo sviluppo di una semantica formale assegnabile al *markup* che fissa la struttura del documento (cfr. Renear *et al.*, 2003b). Gli autori assumono che il significato del *markup* sia costituito dalle inferenze che esso autorizza⁵ e sostengono che l'identità del documento è determinata dall'equivalenza logica delle inferenze che possono essere ricavate dai valori assegnati agli elementi della struttura astratta definita dalla codifica.

Ci troviamo quindi di fronte ad un ventaglio di definizioni diverse e diversamente valutabili. La definizione dello schema Kahn-Wilensky privilegia la struttura funzionale dell'oggetto digitale, la definizione dei CEEC si riferisce alla riproduzione visiva del documento originale e la definizione di Dubin e Renear fa riferimento alla struttura astratta di una rappresentazione XML. In tutti i casi, però, ci troviamo di fronte ad oggetti digitali complessi, costituiti da rappresentazioni multiple e di tipo diverso; di conseguenza, la definizione dell'identità dell'oggetto digitale fa riferimento ad un aspetto invariante, di volta in volta diversamente identificato, delle varie rappresentazioni. Nella prospettiva che qui ho assunto è possibile tuttavia affermare che, dal punto di vista dell'eshaustività, l'elemento invariante è determinato dall'espressione, o dalla forma della rappresentazione del documento originale; e che invece, dal punto di vista della funzionalità, l'elemento invariante è determinato dal contenuto informativo della rappresentazione e dal suo modello semantico. È anche possibile osservare che i due punti di vista ora distinti corrispondono ai diversi orientamenti di quelle che sono state chiamate rispettivamente la *document community* e la *database community*. Mentre nella *document community* è prevalsa la preoccupazione di assicurare la «permanenza» dei «documenti» nel trasferimento dei dati da un sistema all'altro, nella *database community* è prevalsa la preoccupazione di salvaguardare non tanto la «permanenza» della rappresentazione dei dati quanto la stabilità della «semantica delle applicazioni» che operano su quei dati (Raymond *et al.*, 1996: 4-5). Così, in un caso è stata privilegiata l'*invarianza della rappresentazione* dell'informazione,

⁵ «Le asserzioni 'autorizzate' (*licensed*) dalla rappresentazione serializzata XML [...] sono il significato di quella rappresentazione» (Renear e Dubin, 2003: 186).

mentre nell'altro è stata privilegiata l'*invarianza del contenuto* informativo. Si ottengono così diversi criteri di identità a seconda delle condizioni di invarianza fissate nell'elaborazione dei dati.

5. Questa discussione sull'identità permette di mettere in luce una caratteristica interessante degli oggetti digitali. Anch'essi, al pari dei loro corrispettivi materiali, le iscrizioni e le rappresentazioni visive, possono essere assimilati a quegli oggetti che Bruno Latour (1986: 1-40) ha efficacemente descritto come oggetti «mobili immutabili», mobili per la variabilità del supporto e della sua posizione, immutabili per l'invarianza della rappresentazione e del suo contenuto. Nel caso dei libri o di altri documenti stampati, ad esempio, «l'immutabilità è assicurata dalla stampa di molte copie identiche; la mobilità dal numero di copie, dalla carta e dai caratteri mobili» (Latour, 1986: 11). Ma come le iscrizioni e le immagini, anche gli oggetti digitali sono caratterizzati dalla compresenza dei due tratti, in apparenza contraddittori, della variabilità e dell'invarianza. E se questa caratteristica ambivalente è propria tanto di oggetti materiali quanto di oggetti digitali, l'opposizione tra «fisso» e «fluidico» non può essere assunta come criterio discriminante tra i documenti tradizionali di natura materiale e i nuovi documenti di natura digitale. Come ha fatto osservare David Levy (1994: 26), «tutti i documenti, indipendentemente dalla tecnologia, sono fissi e fluidici» ad un tempo. Ma l'opposizione non pare ridursi solo ad una questione di grado⁶, di «velocità di cambiamento», o di diverso «ritmo di fissità e fluidità» (Levy, 1994: 26): la variabilità del documento non dipende unicamente dalle modificazioni del suo supporto, o dalle trasformazioni della sua configurazione materiale. Ma non dipende nemmeno, esclusivamente, dalle «interazioni tra la fissità dei documenti e la flessibilità delle pratiche sociali» che essi promuovono e coordinano (Brown e Duguid, 1995): il grado di fissità e fluidità dei documenti non può essere soltanto «relativo agli scopi e alle pratiche» che ne modificano la percezione (Levy, 1994: 26-27). Nell'un caso, infatti, faremmo riferimento solo alla componente materiale e nell'altro solo alla componente semantica dei documenti, mentre la fissità e la fluidità degli oggetti «mobili immutabili», materiali o digitali che siano, dipendono in ultima analisi dalla loro natura di vettori di informazione e dal rapporto, necessariamente indeterminato, tra la loro componente fisica e la loro componente semantica.

⁶ «Levy dice che tanto i documenti fisici quanto i documenti digitali hanno gradi diversi di fissità e fluidità» (Lagoze, 2000).

La fissità e la fluidità dei documenti possono così essere ricondotte a quella che è stata propriamente descritta da Jerome McGann come «condizione testuale», e alla sua natura palesemente «paradossale» (McGann, 2001: 175, trad. it., 2002: 194). Questa condizione può essere propria di ogni tipo di oggetto: in quanto vettore di informazione lo stesso «ordine materiale delle cose» può «essere considerato come un testo». Per il fisico, «gli incommensurabili della meccanica quantistica» sono testo (McGann, 1991: 201, nota 5), così come è testo, per lo storico dell'arte, l'opera dell'artista. Ma nessun testo, e nessun elemento del testo, «può essere considerato uguale a se stesso» (McGann, 2001: 175; trad. it. 2002: 194): «la variazione», infatti, è ciò che costituisce «la regola invariante della condizione testuale» (McGann, 1991: 185). È dunque la natura semiotica stessa degli oggetti digitali, ciò che li apparenta agli oggetti «mobili immutabili» di natura materiale; è l'instabilità del rapporto tra la materialità propria di ogni segno e l'informazione che esso veicola, ciò che conferisce loro quella caratteristica ambivalente di variabilità ed invarianza che li rende singolarmente simili al testo. Il richiamo al testo non deve tuttavia sembrare motivato solo dalle preoccupazioni dell'umanista. Con la diffusione sempre crescente del linguaggio XML, impostosi come linguaggio standard per l'interscambio e la gestione dei dati sul *www*, si assiste in effetti ad una sorta di testualizzazione di ogni tipo di dato, comunque strutturato. Il linguaggio XML, infatti, può essere considerato come un meccanismo di linearizzazione, un metodo di rappresentazione della conoscenza in forma seriale e lineare. Ogni tipo di dato può essere rappresentato come una stringa di caratteri, ossia ridotto a testo. Effettivamente, «i documenti XML sono semplicemente dei *file* di testo» (Williams *et al.*, 2000: 2) e qualche riflessione sulla natura del testo può essere utile proprio per affrontare il problema del rapporto tra esaustività e funzionalità nella rappresentazione digitale dei documenti. Il sapere umanistico, così legato all'uso del testo come forma tradizionalmente privilegiata di rappresentazione della conoscenza, può venire utilmente in soccorso all'impiego dell'informatica e dell'elaborazione dei dati per la conservazione del patrimonio culturale.

6. I due aspetti dell'identità, la variabilità e l'invarianza, diversamente definiti in relazione all'esaustività e alla funzionalità della rappresentazione digitale, si ritrovano dinamicamente ed essenzialmente coniugati nella natura stessa del testo. Si pensi ai fenomeni correlativi della sinonimia e della polisemia: in breve, ci possono essere modi diversi per esprimere la stessa cosa e

una stessa frase o espressione può essere intesa e interpretata in modi diversi. L'identità del testo dipende da un elemento invariante, ma l'invariante può essere costituito, di volta in volta, dall'espressione o dal contenuto. Fissando il contenuto si possono produrre varianti testuali e fissando l'espressione si possono produrre varianti interpretative. È noto che non solo la trasmissione, ma la composizione stessa del testo produce delle varianti, ed è altrettanto evidente che un testo ammette diverse interpretazioni. Il testo è un oggetto mobile e immutabile ad un tempo, mobile per la sua variabilità e immutabile per la sua invarianza. Ma il vincolo dell'identità del testo determina una forma di compensazione tra l'invarianza dell'espressione e la variabilità dell'interpretazione, ovvero tra l'invarianza del contenuto fissato da una certa interpretazione e la variabilità delle espressioni in grado di veicolarlo.

Questa caratteristica di oggetto «mobile immutabile» si esprime ordinariamente nel testo attraverso le notazioni diacritiche, o attraverso qualunque altra espressione che ne manifesti la cosiddetta «metalinguisticità riflessiva» (cfr. De Mauro, 1982: 93-94; 2002: 89, 91-93). È una proprietà essenziale del linguaggio naturale quella di poter descrivere se stesso. I segni diacritici e le espressioni autoriflessive del testo possiedono così il doppio carattere di rappresentazione, in quanto fanno parte dell'espressione del testo, e di rappresentazione della rappresentazione, in quanto descrivono l'espressione stessa del contenuto del testo. La mobilità del testo si manifesta nell'ambivalenza semantica delle espressioni autoriflessive e diacritiche. Esse possono essere considerate di volta in volta come espressioni appartenenti al linguaggio costitutivo del testo o come descrizioni esterne, metalinguistiche, della sua struttura. In termini formali, esse alternativamente mostrano⁷, oppure descrivono, la forma logica del testo, così come di volta in volta possono rappresentare un operatore funzionale, oppure il valore che ne dipende⁸. Intese come espressioni autoriflessive del linguaggio-oggetto, esse comportano forme di predicazione di ordine superiore.

⁷ Cfr. Wittgenstein (1964: 28-29): «ciò, che nel linguaggio esprime *sé*, *noi* non possiamo esprimere mediante il linguaggio»; secondo Wittgenstein, un'espressione del linguaggio-oggetto «non può rappresentare» la sua forma logica, ma la «mostra» o la «esibisce».

⁸ I segni diacritici possono essere considerati come funzionalmente analoghi ai segni (*marks*) impiegati nel *calcolo delle indicazioni* introdotto da George Spencer-Brown (1969). Nel calcolo di Spencer-Brown, ogni indicazione è espressa da un «segno (*token*)», o «nome», che può essere inteso in due modi, come un'operazione o un «atto di distinzione», oppure come il «valore» di quell'atto o indicazione (Varela, 1979: 110-111).

L'immutabilità del testo si manifesta invece nel suo carattere di invarianza. Secondo Cesare Segre (1985: 29) «il testo è l'*invariante*, la successione dei valori, rispetto alle variabili dei caratteri, della scrittura» (il corsivo è aggiunto), o delle espressioni che lo costituiscono e il vincolo dell'identità conferisce natura olistica alla sua rappresentazione strutturale. La struttura del testo può essere concepita, ancora con Segre (1985: 44), come «l'assieme delle relazioni latenti» tra le sue parti, ed è in effetti costituita dalla totalità invariante del testo che ne definisce le reciproche relazioni. L'ambivalenza delle espressioni diacritiche determina i «punti critici» (cfr. Caglioti, 1994: 34), o di instabilità dinamica, della struttura del testo e definisce le zone di cerniera tra diverse forme varianti ed effettivamente coesistenti dell'insieme totale, unico e invariante, delle sue possibili forme di organizzazione. Il vincolo dell'identità dell'insieme pone così una condizione di equivalenza tra la formulazione linguistica e la formulazione metalinguistica delle espressioni diacritiche, che possono essere reciprocamente convertite l'una nell'altra. L'ambivalenza delle espressioni diacritiche permette di esprimere il rapporto di compensazione tra la variabilità e l'invarianza, rispettivamente, dell'espressione e del contenuto del testo; il vincolo dell'identità permette invece di formularne la legge (cfr. Buzzetti, 2000). La variabilità delle parti e l'invarianza del tutto è una delle proprietà essenziali che caratterizzano le strutture olistiche e il testo in quanto oggetto «mobile immutabile» ne rappresenta un esempio tipico e particolare.

Una conseguenza significativa della mobilità del testo è costituita dall'interscambiabilità delle sue componenti, l'espressione e il contenuto, interscambiabilità che si esprime nell'ambivalenza delle espressioni diacritiche. Nel segno diacritico è possibile scorgere, contemporaneamente, un elemento materiale dell'espressione del testo e un elemento formale dell'articolazione strutturale del suo contenuto. Nel testo, l'elemento materiale veicola in forma non analizzata l'intero contenuto informativo del testo, a cui l'elemento formale assegna un modello e un'articolazione strutturale analitica. Così, l'elemento materiale assicura l'esaustività della rappresentazione e l'elemento formale ne determina la funzionalità. Le relazioni dinamiche tra i due aspetti della rappresentazione, l'esaustività e la funzionalità, dipendono dall'interazione reciproca tra la componente materiale e la componente semantica del testo. Come si è visto, la convertibilità e la compensazione dei due elementi, l'espressione e il contenuto, si manifestano nel ruolo ambivalente dei segni diacritici; ma è precisamente questo, nella rappresentazione *digitale* del testo, il ruolo svolto dal *markup*. La natura del *markup*, infatti, è essenzialmente diacritica e ambi-

valente: proprio come ogni altro segno diacritico, «il *markup* è ad un tempo *rappresentazione* e *rappresentazione della rappresentazione*» del testo (Buzzei, 1999: 153). Quindi, nella forma digitale di rappresentazione del testo, il temperamento delle due esigenze fondamentali, l'eshaustività e la funzionalità della rappresentazione, dipende in modo essenziale dalla funzione specifica svolta del *markup* nell'assegnare all'informazione testuale forme diverse di organizzazione strutturale.

Un modello funzionalmente adeguato dell'articolazione semiotica del testo richiede rappresentazioni distinte della struttura dell'espressione, della struttura del contenuto e della legge della loro dipendenza reciproca. Tali rappresentazioni distinte possono essere ottenute grazie all'ambivalenza diacritica del *markup* e possono essere riaggregate in un oggetto digitale complesso che riproduce l'organizzazione funzionale del modello semiotico del testo. A sua volta, il processo di serializzazione e linearizzazione di ogni tipo di rappresentazione digitale dell'informazione, che si è avviato e si sta progressivamente sviluppando sul Web con l'impiego sempre crescente di XML, dipende in ultima analisi dall'applicazione di una forma di *markup*, qual è in effetti l'uso del linguaggio XML. La rappresentazione in forma lineare o testuale di ogni tipo di dato permette così di assumere la rappresentazione del testo come modello esemplare e prototipo funzionale per la rappresentazione e la conservazione in forma digitale di ogni tipo di documento e di oggetto culturale. Nel processo di migrazione e trasposizione in ambiente digitale dell'intero patrimonio delle nostre conoscenze, il testo non solo mantiene, ma addirittura estende la sua prerogativa principale e si conferma forma privilegiata di rappresentazione della conoscenza.

La sfida per la rappresentazione in forma digitale di ogni tipo di informazione e per la sua conservazione adeguatamente esauriente e funzionale è dunque costituita dalla possibilità di rappresentare il testo come un oggetto digitale complesso e dalla capacità di riprodurre funzionalmente le forme di interazione tra la struttura dell'espressione e la struttura del contenuto dell'informazione testuale. La rappresentazione formale di entrambe le componenti strutturali e dei processi della loro mutua conversione e compensazione si pone come condizione di adeguatezza imprescindibile e fondamentale. Infatti, mentre le forme di rappresentazione tendenzialmente esaurienti tendono a restituire la forma esteriore dell'oggetto rappresentato, ossia a privilegiarne l'espressione, le forme di rappresentazione funzionalmente soddisfacenti operano sul contenuto informativo che esso trasmette e ne ripropongono una

«vista» o, se si vuole, un'interpretazione particolare. La risposta a questa sfida dipende dall'elaborazione di un modello dinamico e funzionale per la rappresentazione digitale del testo e quindi in ultima analisi da un uso appropriato del *markup*. Ma come si può applicare il *markup* per riprodurre il testo come un oggetto digitale complesso idoneo alla rappresentazione esaustiva e funzionale dell'informazione testuale?

7. Attraverso l'uso del *markup* è possibile rendere esplicito, nella rappresentazione digitale, il fenomeno della compensazione tra la variabilità del contenuto e l'invarianza dell'espressione ovvero, reciprocamente, tra l'invarianza del contenuto e la variabilità dell'espressione del testo. Lo *status* logico-linguistico del *markup* è ambivalente come quello di ogni altro segno diacritico attraverso cui si esprime la cosiddetta metalinguisticità riflessiva del linguaggio naturale. Si pensi, per fare un esempio, alla punteggiatura, che è stata legittimamente considerata come una vera e propria forma di *markup* (Coombs *et al.*, 1987: 934-935). Se l'editore di un testo medievale introduce una forma di punteggiatura secondo l'uso moderno, che tipo di operazione realizza? Rende semplicemente esplicito ciò che era implicitamente contenuto nel testo, o vi introduce una descrizione esterna, metalinguistica, della struttura sintattica che gli è propria? L'operazione può essere intesa in entrambi i modi, ma nei due casi cambia lo *status* logico rispettivamente assegnato al *markup*. Nel primo caso, il *markup* viene interpretato come un'espressione autoriflessiva del linguaggio-oggetto e in quanto tale esprime asserzioni sul linguaggio stesso, che possono essere espresse attraverso forme di predicazione di ordine superiore; nel secondo caso, il *markup* viene interpretato come una descrizione metalinguistica della struttura di volta in volta assegnata al testo. In entrambi i casi il *markup* permette di ottenere una rappresentazione strutturale del testo, ma di quale struttura si tratta? la struttura dell'espressione, o la struttura del contenuto? la struttura di dati che costituisce la rappresentazione del testo codificata in forma digitale, o la struttura dell'informazione che tale aggregato di dati permette di rappresentare? A seconda delle proprietà semantiche assegnate al *markup*, la totalità invariante delle relazioni strutturali costitutive del testo viene identificata o con l'una o con l'altra delle sue componenti fondamentali, l'espressione e il contenuto. Questa peculiare ambivalenza semantica è proprio ciò che permette di utilizzare il *markup* come meccanismo di conversione dell'informazione strutturale riferita all'espressione in informazione strutturale riferita al contenuto. Vediamo come.

A ben vedere, la natura del testo è fondamentalmente «ingiuntiva». Il testo può essere considerato come l'espressione materiale, o il *valore*, di un'operazione costitutiva di senso. L'espressione del testo, a sua volta, può essere intesa non come il valore o il risultato di tale operazione, ma come un *operatore* o come il simbolo dell'operazione inversa, cioè come un insieme di «istruzioni» per l'interpretazione del testo (McGann, 2001: 205; trad. it. 2002: 226-227). In tale processo, le determinazioni strutturali dell'espressione e del contenuto si manifestano esplicitamente attraverso i segni diacritici ovvero – nella rappresentazione del testo in forma digitale – attraverso il *markup*. Se assumiamo il *markup*, o le determinazioni strutturali dell'espressione, come il valore di un'operazione costitutiva di senso, fissiamo l'espressione, ma ne lasciamo indeterminato il contenuto. Di qui la possibilità di considerare il *markup*, o le determinazioni strutturali dell'espressione, come operatori o come regole per assegnare al testo un'interpretazione e determinare la struttura del suo contenuto. Se ora assumiamo queste nuove determinazioni strutturali del contenuto, o il *markup* che così si aggiunge al testo, come il valore di tale operazione o interpretazione, fissiamo il contenuto, ma ne lasciamo indeterminata l'espressione. Questo offre nuovamente la possibilità di considerare il nuovo *markup*, o le nuove determinazioni strutturali del contenuto, come operatori o come regole per assegnare all'espressione una nuova struttura, ovvero come la struttura profonda o il modello generativo della struttura superficiale del testo. E se assumiamo, a loro volta, queste determinazioni strutturali dell'espressione, o il *markup* che si aggiunge ulteriormente al testo, come il valore di quest'ultima operazione, ossia come una riformulazione dell'espressione del testo, siamo alla fine riportati al punto di partenza. Il processo può ricominciare e l'intero ciclo della compensazione tra la determinazione e l'indeterminazione strutturale dell'espressione e del contenuto del testo può essere nuovamente ripercorso.

Un modello di questo tipo può permettere di ottenere una rappresentazione del testo che ne riproduce la dinamica interna in modo esaustivo e funzionale. L'ambiguità diacritica del *markup* può essere opportunamente utilizzata per dare risposta adeguata ad entrambe le esigenze. A seconda delle proprietà semantiche associate al *markup*, la totalità invariante delle relazioni strutturali costitutive del testo viene identificata o con l'una o con l'altra delle sue componenti fondamentali, l'espressione e il contenuto. Se il *markup* viene inteso come espressione autoriflessiva appartenente al linguaggio-oggetto, fornisce indicazioni per l'interpretazione del contenuto del testo e per asse-

gnargli, di conseguenza, una struttura – e in tal caso la totalità invariante assunta come unità costitutiva dell'insieme delle relazioni strutturali latenti del testo è identificata col suo contenuto informativo. Se invece il *markup* viene inteso come espressione metalinguistica che descrive specifiche caratteristiche del testo, designa determinate stringhe di caratteri e assegna una determinata forma logica all'espressione del testo – e in tal caso la totalità invariante assunta come unità costitutiva dell'insieme delle relazioni strutturali latenti del testo è identificata con la sua espressione. Ora, la rappresentazione della struttura dell'espressione e le operazioni ad essa riferite soddisfano le esigenze di riproduzione fedele e integrale dell'oggetto rappresentato e mirano ad una restituzione esaustiva di tutta l'informazione trasmessa dalle sue caratteristiche materiali; a loro volta, la rappresentazione della struttura del contenuto informativo e le operazioni ad essa riferite soddisfano le esigenze di una restituzione analoga ed equivalente di tutti gli usi comunicativi dell'oggetto rappresentato e mirano ad una riproduzione funzionalmente adeguata del suo contenuto informativo. Dunque, nell'uso metalinguistico, il *markup* opera direttamente sull'espressione del testo ed esplicita gli aspetti che concorrono prevalentemente a soddisfare le esigenze di esaustività della sua rappresentazione; invece, nell'uso linguistico e autoriflessivo, il *markup* opera direttamente sul contenuto del testo ed esplicita gli aspetti che concorrono prevalentemente a soddisfare le esigenze di funzionalità della sua rappresentazione digitale.

Il modello dinamico della mobilità interna del testo sopra delineato prevede l'impiego di entrambe le forme di *markup* e permette di ottenere rappresentazioni strutturali distinte dell'espressione e del contenuto del testo. In tale modello, il *markup* funge da meccanismo di scambio tra i due tipi di rappresentazione e determina le forme della loro mutua dipendenza. Il *markup* permette così di articolare la rappresentazione del testo in un oggetto digitale complesso che ne rende possibile al tempo stesso sia la riproduzione esaustiva, sia la riproduzione funzionale. La capacità del *markup* di rappresentare distintamente le componenti strutturali dell'informazione testuale e di descrivere le loro mutue relazioni permette di fissare in aspetti di volta in volta diversi l'elemento invariante del testo e di stabilire le condizioni di identità dell'oggetto digitale complesso che ne costituisce la rappresentazione in funzione delle specifiche modalità d'impiego dell'informazione testuale.

8. Può l'analogia tra gli oggetti digitali complessi e la rappresentazione digitale del testo – un'analogia fondata sulla nozione di oggetto 'mobile im-

mutabile' e sulla sua caratteristica ambivalenza di variabilità ed invarianza – esserci d'aiuto per contemperare esaustività e funzionalità nella conservazione dell'informazione? Un primo aspetto del problema riguarda l'identità degli oggetti digitali complessi. A che cosa può condurre la consapevolezza che nessun elemento del testo può mai essere considerato uguale a se stesso? Al riconoscimento che nessuna componente dell'oggetto digitale può essere considerata elemento invariante esclusivo per fissarne l'identità e che nella complessità strutturale dell'oggetto digitale si riflette la complessità semiotica dell'oggetto culturale. L'identità degli oggetti digitali complessi non è univocamente definibile e di fatto si riscontra il ricorso ad una molteplicità di criteri di volta in volta diversi. Questo accade perché la determinazione dell'identità dipende caso per caso dalle modalità d'uso dell'informazione veicolata. Per stabilire l'identità del testo, occorre vincolare e rendere invariante uno dei suoi elementi mutanti, l'espressione o il contenuto, ed ammettere tutte le possibili variazioni dell'elemento lasciato di conseguenza indeterminato, considerando la totalità delle relazioni strutturali implicite tra le sue parti integranti. Allo stesso modo, l'identità degli oggetti digitali, definita da una condizione di invarianza fissata in relazione all'uso, comporta il riferimento all'insieme virtuale di tutte le possibili configurazioni degli altri elementi varianti. Ma le condizioni di invarianza imposte rispettivamente dai requisiti di esaustività e di funzionalità della rappresentazione digitale sono diverse: nel primo caso occorre conservare l'integrità dell'*espressione*, o presentazione dell'oggetto, in quanto vincolo invariante di tutte le possibili interpretazioni dell'informazione trasmessa; nel secondo caso occorre conservare l'integrità del *contenuto*, in quanto insieme di regole operative invarianti per la generazione di tutte le possibili viste, o presentazioni, dell'oggetto.

Connesso col problema dell'identità e dei criteri della sua determinazione è il problema delle modalità di conservazione dell'oggetto digitale. Quali caratteristiche dell'oggetto digitale occorre salvaguardare per soddisfare le esigenze di conservazione esaustiva e funzionale dell'oggetto culturale che esso rappresenta? La nozione museale della conservazione dell'oggetto materiale in quanto tale non vale per l'archiviazione conservativa dell'oggetto digitale. L'oggetto digitale è costitutivamente diverso da un oggetto concreto, e la sua conservazione non può fondarsi, per esempio, «sul principio che l'oggetto fisico che incorpora un diploma (*record*) non abbia subito nessun cambiamento tale da modificare in alcun modo il messaggio

che intendeva comunicare» (InterPARES Preservation Task Force, 2001). In campo archivistico, quindi, si è venuto affermando il principio che «la conservazione digitale non è un semplice processo di conservazione di oggetti fisici, ma un processo di conservazione della capacità di riprodurre gli oggetti»; infatti, «non si può provare di aver conservato un oggetto finché non lo si è ricostituito in una forma che ne permetta l'uso da parte degli esseri umani o dei programmi di un sistema di elaborazione automatica» (Thibodeau, 2002: 12). Ciò che interessa conservare, quindi, non è la 'persistenza' (*persistence*) dell'oggetto digitale nel formato originario, ma la sua 'operatività' (*performance*), ed è emersa l'esigenza di distinguere tra gli aspetti 'tecnici,' o 'fisici,' e gli aspetti 'concettuali,' o 'intellettuali,' dell'oggetto digitale (Hofman, 2002a: 17); una distinzione che peraltro stenta ad emergere chiaramente anche nel recente progetto dell'OCLC per lo sviluppo di un esauriente sistema (*comprehensive framework*) di metadati per la conservazione degli oggetti digitali (OCLC/RLG, 2002), dove pare, secondo Hans Hofman, «che l'oggetto digitale venga considerato solo come un'entità tecnica e che sia essa in quanto tale e non l'oggetto intellettuale ad esser posta al centro dell'attenzione» (Hofman, 2002a: 17). È vero invece che «gli oggetti digitali hanno sia aspetti intellettuali – il diploma o la pubblicazione che vengono visualizzati sullo schermo – sia aspetti tecnici – le componenti digitali (*file* di dati e *software*) che ne contengono la necessaria rappresentazione digitale» (Hofman, 2002b). Ma si può veramente identificare l'«oggetto intellettuale» con la sua presentazione sullo schermo? Di nuovo l'analogia col testo ci può essere utile per concepire un adeguato modello semiotico dell'oggetto digitale.

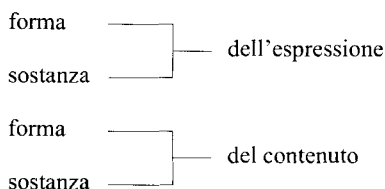
9. Certamente «un oggetto digitale» può essere definito, con Kenneth Thibodeau, come «un oggetto informativo, qualunque sia il tipo di informazione rappresentata o qualunque ne sia il formato, purché essa sia espressa in forma digitale» (Thibodeau, 2002: 6). Ma anche l'analisi che egli ne propone, ispirandosi alla comune distinzione tra 'dati,' 'informazione' e 'conoscenza,'⁹ non sembra del tutto adeguata da un punto di vista semiotico. Secondo Thibodeau:

⁹ FOLDOC, 2003: «Knowledge differs from data or information in that new knowledge may be created from existing knowledge using logical inference. If information is data plus meaning then knowledge is information plus processing».

ogni oggetto digitale è, insieme, un oggetto fisico, un oggetto logico e un oggetto concettuale, e le sue proprietà a ciascuno di questi tre livelli possono essere considerevolmente diverse. Un oggetto *fisico* è semplicemente un'iscrizione di segni su un supporto fisico; un oggetto *logico* è un oggetto riconosciuto ed elaborato da un programma; e l'oggetto *concettuale* è l'oggetto così com'è riconosciuto e compreso da una persona, o così com'è riconosciuto ed elaborato, in alcuni casi, da un programma capace di eseguire operazioni finanziarie (Thibodeau, 2002: 6).

Questa classificazione si fonda, in effetti, «sulla definizione di 'dati' (*data objects*), 'informazione' e 'conoscenza' più semplice possibile»; ma se questa definizione può essere giustificata, da un punto di vista ingegneristico, per «minimizzare lo sviluppo dell'infrastruttura» che serve all'amministrazione di oggetti digitali complessi, non pare sufficientemente articolata per affrontare i problemi della relazione tra l'eshaustività e la funzionalità della rappresentazione digitale. Secondo questa definizione minimale, 'dati' o «oggetti di dati (*data objects*) sono sequenze di bit (*bit streams*); 'informazione' sta per qualsiasi unità o elemento di dati etichettato (*tagged*); 'conoscenza' sta per qualsiasi relazione intercorrente tra due unità di informazione» (Moore, 2001: 30). Ma la distinzione tra sequenze di bit (oggetti fisici), *tagged attributes* o attributi assegnati da marcatori (oggetti logici), e relazioni tra attributi (oggetti concettuali), può esser causa di equivoci e generare difficoltà per una corretta analisi semiotica. Quando si parla di informazione logica o strutturale, a quale struttura ci si riferisce?

È precisamente a questo riguardo che il modello semiotico del testo e la distinzione tra *espressione* e *contenuto* possono ancora tornare utili. Secondo Hjelmslev, è necessario distinguere tra la *forma* e la *sostanza*, tanto dell'espressione quanto del contenuto. Ne risulta uno schema articolato nel modo seguente:



Secondo questo modello, le determinazioni strutturali dipendono dal «proiettarsi della forma sulla materia» (cfr. Hjelmslev, 1968: 52-65), ma nel

caso dell'espressione, la forma assegna struttura ad un oggetto fisico, nel caso del contenuto, invece, ad un oggetto concettuale. L'informazione di tipo 'logico,' o 'strutturale,' può quindi riguardare tanto gli aspetti sintattici dell'espressione, quanto gli aspetti semantici del contenuto. Di conseguenza, la definizione di oggetto logico proposta da Thibodeau risulta sostanzialmente ambigua. Infatti, da un lato si afferma che «le regole che si applicano al livello logico determinano il modo in cui l'informazione è codificata in bit e il modo in cui codifiche diverse sono tradotte in altri *formati*»; e dall'altro si dice che «una stringa di dati tutti conformi allo stesso *tipo di dato* costituisce un oggetto logico» (Thibodeau, 2002: 7, i corsivi sono aggiunti). Secondo questa definizione, quindi, l'oggetto logico può essere di volta in volta costituito da un semplice *formato*¹⁰, che consiste in un'aggregazione puramente sintattica, oppure da un *tipo di dato*, che consiste invece nell'associazione di un insieme di dati elementari, o di aggregati di dati, con gli operatori ad essi applicati (cfr. Lodi e Pacini, 1990: 16-17). Nel primo caso si tratterà di una semplice *struttura di dati*, nel secondo di un *modello di dati*¹¹. La nozione di aggregato, o struttura, di dati è sintattica e riguarda la forma dell'espressione, mentre la nozione di tipo, o modello, di dati è semantica e riguarda la forma del contenuto. La semantica di cui si tratta, però, è una semantica astratta, o un formalismo¹², applicabile a svariati tipi di oggetti che ne rispettino le specificazioni formali, tra i quali possono essere compresi anche gli 'oggetti di dati,' ossia i dati elementari o gli aggregati di dati non interpretati. Tale semantica astratta può essere intesa come una semantica concreta applicata agli oggetti reali, materiali o concettuali, ai quali i dati si riferiscono, solo se ne costituisce un modello strutturale adeguato, ossia se le relazioni tra gli oggetti di dati e le relazioni tra gli oggetti reali sono le stesse. Ma questo

¹⁰ Un *formato* può essere definito come la «sintassi» secondo cui una certa «rappresentazione» dell'informazione è stata «codificata» (Joloboff, 1989: 75-76).

¹¹ Questa distinzione terminologica tra 'struttura di dati' e 'modello di dati' può essere opportuna per evitare l'uso ambiguo del termine 'struttura di dati': «In accordo con l'uso comune, e quando il contesto non è ambiguo, il termine 'struttura di dati' sarà usato sia per indicare un singolo aggregato di dati sia per indicare la struttura di dati propriamente intesa», cioè un «insieme di aggregati di dati con i suoi operatori» (Lodi e Pacini, 1990: 17). Per un uso dei termini conforme alla distinzione qui proposta, cfr. Raymond *et al.*, 1993.

¹² Un *formalismo* può essere descritto come ciò che fornisce un'«interpretazione» ad una «rappresentazione» dell'informazione, ossia come ciò che assegna un certo «significato» alla «sequenza dei valori digitali», o bit, da cui è costituita (Joloboff, 1989: 75-76).

non vale in generale per gli elementi sintatticamente definiti del testo, da un lato, e le cose alle quali il testo si riferisce, dall'altro. Secondo l'antico adagio stoico, la parola 'cane' non morde e anche Thibodeau ammette che «lo stesso contenuto concettuale può essere rappresentato con codifiche molto diverse e [che] la struttura concettuale può essere notevolmente diversa dalla struttura dell'oggetto logico» (Thibodeau, 2002: 8).

L'esempio addotto da Thibodeau per arrivare a questa conclusione ci riporta all'affermazione di Hofman da cui siamo partiti, secondo cui l'oggetto intellettuale sarebbe costituito dal documento visualizzato sullo schermo (cfr. Hofman, 2002*b*). Thibodeau si riferisce infatti alla possibilità di visualizzare «il *contenuto* di un documento», che «può essere codificato in forma digitale come immagine della pagina» (Thibodeau, 2002: 8; corsivo aggiunto), per esempio nel formato Adobe PDF, oppure nel «formato dell'elaboratore di testo» orientato ai caratteri (Thibodeau, 2002: 11) con cui è stato composto, per esempio Microsoft Word. Il formato PDF visualizza il documento nella stessa forma in cui è stato realizzato con l'elaboratore di testo, ma in questo caso l'oggetto invariante di cui si danno due rappresentazioni digitali diverse è l'*espressione*, non il *contenuto* del testo, come invece afferma Thibodeau. Il documento presentato sullo schermo non può essere identificato con l'oggetto concettuale, ma solo con un suo aspetto particolare. L'immagine digitale di un manoscritto può essere considerata una rappresentazione più esaustiva di una sua trascrizione digitale, ma la trascrizione è certamente molto più funzionale dell'immagine per quanto riguarda l'elaborazione del contenuto del testo. La distinzione tra espressione e contenuto è dunque rilevante ai fini della conservazione dell'informazione, se è vero che essa deve «permettere la ricostituzione e l'interpretazione della struttura e del contenuto dei dati archiviati» (Ludäscher *et al.*, 2001*a*: 54). Lo si riconosce implicitamente quando si afferma che i criteri della conservazione «dipendono dall'uso che si intende fare dell'oggetto», ma che «non è possibile prescrivere o predire tutti gli usi» che ne saranno fatti (Thibodeau, 2002: 14). E a seconda dell'uso può essere necessario restituire una rappresentazione visiva dell'espressione o una rappresentazione funzionale del contenuto del documento.

Non è però possibile tenere adeguatamente conto della distinzione tra struttura dell'espressione e struttura del contenuto, se si associano in modo biunivoco la «struttura» e il «contenuto» dei dati archiviati rispettivamente con l'«informazione» e la «conoscenza» circa i dati archiviati (Ludäscher *et al.*, 2001*a*: 54). La classificazione in 'dati', 'informazione' e 'conoscenza' per-

mette senz'altro di distinguere le componenti digitali dalle componenti concettuali dell'oggetto digitale. La definizione di componente digitale è chiara:

le componenti digitali di un oggetto sono gli oggetti logici e fisici che sono necessari a ricostituire l'oggetto digitale;

ma questa definizione non permette di cogliere in modo chiaro la relazione tra componenti digitali e componenti concettuali. Ciò risulta evidente quando si afferma, per esempio, che le componenti digitali «non sono necessariamente limitate agli oggetti che contengono il contenuto di un documento» e che «possono contenere dati necessari alla struttura o presentazione dell'oggetto concettuale» (Thibodeau, 2002: 12); oppure che

il contenuto e la struttura di un oggetto concettuale debbono essere contenuti in qualche modo nell'oggetto logico o negli oggetti logici che rappresentano quell'oggetto in forma digitale (Thibodeau, 2002: 8).

Ma sta forse in quest'ultima affermazione di Thibodeau, in cui si accenna al problema della rappresentazione, la chiave del problema. L'oggetto digitale, in quanto vettore di informazione, è una rappresentazione, ossia un segno dotato di significato. Ma il significato del segno, in questo caso l'oggetto concettuale, non può essere identificato immediatamente con l'oggetto reale che esso rappresenta, o con la sua visualizzazione sullo schermo. Il significato di un segno non è l'oggetto a cui esso si riferisce, ma la *relazione* tra il segno e l'oggetto rappresentato. In termini semiotici, si può dire che l'«oggetto concettuale» costituisce il *contenuto* rappresentativo dell'oggetto digitale e che le sue componenti digitali ne costituiscono l'*espressione*; a proposito dell'espressione, poi, dobbiamo distinguere la *materia*, ovvero l'«oggetto fisico», e la *forma*, ovvero l'«oggetto logico»; e a proposito della forma dell'espressione si possono distinguere, a loro volta, la «struttura superficiale», cioè il «formato», e la «struttura profonda», cioè il «tipo di dato», ossia quella proprietà formale che può ricondurre l'espressione «entro l'ambito di una teoria semantica» (Davidson, 1970: 202 e 210). Ho fatto ricorso alle precisazioni di un filosofo del linguaggio come Donald Davidson, per insistere nell'analogia col modello semiotico del testo e per mostrarne ancora una volta la produttività. Il riferimento alla struttura profonda può essere molto istruttivo perché l'oggetto logico può essere considerato come il modello generativo della riproduzione del documento. È dunque molto

importante stabilire se l'oggetto logico rappresenta la struttura dell'espressione o la struttura del contenuto dell'oggetto rappresentato, perché da ciò dipendono le caratteristiche di esaustività e di funzionalità della rappresentazione. Pertanto, se non si tiene adeguatamente conto di questa distinzione, si possono verificare seri inconvenienti.

10. Nella conservazione degli oggetti digitali, vengono giudicati insufficienti sia i procedimenti di emulazione, che cercano di preservare la capacità di elaborare i dati nel formato originario, sia i procedimenti di migrazione, nei quali si modifica il formato per rendere i dati elaborabili con programmi più aggiornati. Si sta così pensando alla conservazione non solo dei dati e degli oggetti logici, ma anche alla conservazione degli oggetti concettuali, attraverso la progettazione di apposite infrastrutture chiamate «archivi persistenti». La conservazione di «oggetti digitali persistenti» (Moore *et al.*, 2000) si fonda su metodi che, pur avendo molti aspetti in comune con quelli comunemente adottati, non operano esclusivamente sulle proprietà «logiche» dei tipi di dati, ma anche sulle «proprietà concettuali» degli oggetti conservati (Thibodeau, 2002: 26-27). In questo processo svolge un ruolo centrale il formato di scambio (*interchange format*) tra gli oggetti digitali. Nella migrazione dei dati, due sistemi diversi debbono «essere in grado di esportare informazione nel formato di scambio e di importare oggetti in questo formato da altri sistemi» (Thibodeau, 2002: 25). Al fine di ottenere rappresentazioni «ripristinabili», o «generiche», ossia «indipendenti dalle infrastrutture» tecniche, utili alla conservazione archivistica di lungo periodo degli oggetti digitali (Ludäscher *et al.*, 2001*b*), il progetto di archivio persistente che è stato sviluppato al San Diego Supercomputer Center (SDSC) e che costituisce il «nucleo» centrale del progetto di Electronic Records Archives (ERA) promosso dalla National Archives and Records Administration (NARA) degli USA (Thibodeau, 2001), prevede «un'infrastruttura archivistica completamente basata su XML» (Ludäscher *et al.*, 2001*a*: 55). Poiché XML può fungere da «meccanismo di scambio per dati strutturati» (W3C, 1999) e semi-strutturati e poiché «un archivio persistente deve mantenere piattaforme e formati che cambiano nel tempo», si è pensato che XML potesse offrire «una buona base per lo scambio di dati e l'interoperabilità» tra sistemi diversi (Ludäscher *et al.*, 2001*a*: 56). Così XML «si è imposto come uno dei principali veicoli per lo scambio di informazione digitale tra diverse piattaforme», benché in origine «sia stato progettato per il *markup*

interno dei documenti» (Thibodeau, 2002: 25). La rappresentazione XML può quindi fungere, in quanto formato di scambio, da modello generativo di ricostituzioni diverse dell'oggetto digitale e viene da chiedersi se non sia proprio per la sua natura sostanzialmente testuale che XML è risultato adatto a svolgere questa funzione. Una rappresentazione più funzionale del testo può quindi giovare immensamente al conseguimento della piena funzionalità nella conservazione digitale dell'informazione. Ma vediamo come.

Quali sono le conseguenze della conversione dei dati in formato XML? XML offre un meccanismo di «linearizzazione» o testualizzazione dei dati, ma che tipo di testualizzazione consente? Che tipo di «modello per la rappresentazione e l'elaborazione del testo» (Sperberg-McQueen e Huitfeldt, 1999: 30) permette di ottenere? Il progetto di 'archivio persistente' proposto dallo SDSC muove dall'osservazione che «XML, considerato come un modello di dati, corrisponde ad alberi ordinati ed etichettati, cioè ad un modello di dati semistutturato» (Ludäscher *et al.*, 2001a: 58). Ora, questo modello coincide con quello su cui si fonda la definizione del testo come «gerarchia ordinata di oggetti di contenuto», ossia col cosiddetto modello OHCO (Ordered Hierarchy of Content Objects) (DeRose *et al.*, 1990: 6; cfr. *supra*). Questa definizione del testo si chiarisce facilmente se si considera che un 'oggetto di contenuto' è costituito da una porzione di documento, che contiene o è contenuta in altri 'oggetti di contenuto,' o porzioni di documento, che sono così connesse strutturalmente tra loro secondo una 'gerarchia' di relazioni di inclusione, i cui elementi sono 'ordinati' in successione nella sequenza lineare di caratteri codificati che costituisce materialmente il documento. La struttura così descritta «è *gerarchica* perché questi oggetti sono 'annidati' l'uno nell'altro» ed «è *ordinata* perché tra gli oggetti c'è una relazione lineare», ossia, dati due oggetti qualunque all'interno di un testo, «uno viene prima dell'altro» (Renear *et al.*, 1996: 265). Il modello del progetto di 'archivio persistente' dello SDSC e il modello OHCO sono equivalenti, perché «nella terminologia della teoria dei grafi le gerarchie ordinate sono 'alberi radicati ordinati'» (*ibid.*: 278, nota 4).

La «tesi OHCO» (*ibid.*: 263), però, assume indebitamente che gli oggetti di cui parla, che sono oggetti di dati, cioè porzioni della sequenza di caratteri codificati che costituisce materialmente la rappresentazione digitale del testo, siano parti di ciò che esso rappresenta, confondendo così la struttura dell'espressione con la struttura del contenuto del testo: gli «oggetti di contenuto», si sostiene, sono chiamati in questo modo «perché, in un certo

senso, sono basati sul *significato* e sullo scopo della comunicazione» (*ibid.*: 265; il corsivo è aggiunto). L'oggetto logico, costituito dalla struttura gerarchica ad albero, viene così inavvertitamente confuso con l'oggetto intellettuale:

ciò che si intende sostenere è che in qualche accezione rilevante del termine (forse in quanto *oggetti intellettuali*) 'libro,' 'testo,' o 'documento' *sono* gerarchie ordinate di oggetti di contenuto (*ibid.*: 265).

Il risultato dell'indebita assimilazione delle due diverse rappresentazioni strutturali è una concezione del testo chiaramente inadeguata e la «debolezza» della 'tesi OHCO' e della nozione di testualità che ne deriva è stata severamente criticata da più parti¹³. Gli stessi proponenti, di fronte al sorgere di seri problemi che non dipendono, a quanto pare, da semplici «questioni tecniche», ma da qualche «più fondamentale carenza concettuale» – primo fra tutti il problema delle cosiddette strutture gerarchiche sovrapposte (*overlapping hierarchies*) (cfr. Barnard *et al.*, 1988: 265-276) – sono stati indotti ad «affinare» la loro concezione del testo e della codifica e a «fare qualche passo» verso una sostanziale revisione della loro tesi. Peraltro, anche i responsabili del progetto di 'archivio persistente' dello SDSC osservano che XML permette, in quanto tale, «un approccio puramente strutturale», ossia sintattico, e ravvisano la necessità di «estender[ne]» il modello «per includervi più informazione semantica» (Ludäscher *et al.*, 2001a: 58).

Tuttavia il problema non riguarda solo la confusione tra il piano dell'espressione e il piano del contenuto, ma riguarda, più in generale, l'adeguatezza e la sufficienza della rappresentazione strutturale ottenuta con XML al fine di garantire esaustività e funzionalità nella conservazione digitale dei documenti. Una rappresentazione strutturale puramente lineare, infatti, non pare completamente adeguata perché «sia la ricostruzione filologica, sia l'interpretazione letteraria della produzione testuale richiedono *rappresentazioni strutturali non lineari* del testo» (Buzzetti, 2000). Ora è pur vero che, data una struttura non lineare di dati, è possibile ricavarne una rappresentazione

¹³ McGann (2001: 141; trad. it., 2002: 156). Per la segnalazione e la valutazione delle critiche mosse alla 'tesi OHCO,' implicitamente accolta dalle *Guidelines* della Text Encoding Initiative (cfr. *supra*, Sperberg-McQueen e Burnard, 1994), si vedano: Unsworth, 2002; Vanhoutte, 2002; Eggert *et al.*, 2002.

lineare XML¹⁴; ma non è vero il contrario, cioè non è possibile, data una rappresentazione lineare XML, definire una struttura non lineare a partire dal suo ordine interno, ossia solamente in funzione delle posizioni dei marcatori nella sequenza lineare dei caratteri¹⁵. In generale, una struttura non lineare può essere espressa mediante un formalismo che opera con una sintassi lineare, ma non è possibile esprimere una struttura non lineare in funzione delle relazioni sintattiche lineari delle sue espressioni simboliche. Più banalmente, la parola ‘palla’ non è rotonda, ossia la struttura dell’espressione e la struttura del contenuto sono tra loro indipendenti. Dunque, in sostanza, una rappresentazione XML non è sufficiente in linea di principio alla ricostruzione di tutte le relazioni strutturali non lineari eventualmente rilevanti all’analisi del suo contenuto concettuale.

La difficoltà è chiaramente avvertita dai ricercatori del progetto di ‘archivio persistente’ del SDSC. Essi osservano che «il metodo corrente per ‘fissare il significato’ di un formato per l’archiviazione e lo scambio di dati è quello di fornire una DTD XML» – una Document Type Definition che specifica i requisiti strutturali della rappresentazione XML dei documenti; si può infatti constatare che «molte organizzazioni e comunità di studio definiscono lo standard del loro ‘linguaggio comunitario’ mediante DTD» (Ludäscher, 2001a: 61) – le *Guidelines* della Text Encoding Initiative, per esempio, non propongono altro che una DTD modulare XML per la codifica dei testi letterari¹⁶. Resta tuttavia il fatto, a fronte di tutto ciò, che una DTD permette di verificare «solo *vincoli strutturali*» di natura sintattica (Ludäscher, 2001a: 61). Quali soluzioni, dunque, sono praticabili per rendere le rappresentazioni digitali soddisfacenti anche sul piano semantico e per permettere un’elaborazione più funzionale del contenuto dei documenti? Il progetto dello SDSC considera l’intera gamma delle soluzioni attualmente disponibili per catturare

¹⁴ Raymond *et al.* (1996: 6): «è possibile esprimere qualsiasi struttura di dati come una sequenza di *markup* e di dati conforme a SGML»; e lo stesso si può dire del suo derivato XML.

¹⁵ Raymond *et al.* (1993: 9): una struttura non lineare «non è riducibile a una descrizione funzionale delle sottocomponenti» di un sistema lineare, perché all’interno di un sistema lineare, come la sintassi XML, non è possibile «esprimere una struttura che non sia un sottoinsieme delle posizioni dei caratteri nel testo».

¹⁶ «Le *Guidelines* della TEI sono espresse come una DTD modulare espandibile, che riguarda quasi tutti i tipi di testo elettronico» e «i moduli che compongono la DTD della TEI possono essere configurati in modo tale da essere usati come una DTD SGML o una DTD XML»; www.tei-c.org/Guidelines2/index.html. Sulle finalità della TEI, cfr. *supra*, nota 1.

«conoscenza», cioè «informazione a livello concettuale», e le enumera in questo modo:

Le rappresentazioni possibili per esprimere tale conoscenza variano molto; si passa dai formalismi relativi ai *database*, come i diagrammi (E)ER, lo Schema XML, o i diagrammi di classe UML, ai formalismi relativi all'intelligenza artificiale e alla rappresentazione della conoscenza (AI/KR), come i programmi logici, le reti semantiche, le ontologie formali, le logiche descrittive, fino ad arrivare alle loro varianti più recenti applicate al Web, come lo Schema RDF e il linguaggio DAML(+OIL), portati a traino dalla barca del Semantic Web (*ibid.*: 57).

Di fronte al «pauroso» problema (*ibid.*: 57) di operare una scelta tra tanti formalismi possibili, la via seguita dal progetto dello SDSC per dotare le rappresentazioni XML di «estensioni semantiche» è quella di «considerarle come linguaggi di vincolo (*constraint languages*)» e di ricavarne «specificazioni eseguibili di vincoli semantici» (*ibid.*: 55). Ricorrendo alla nozione di linguaggio di vincolo, infatti, è possibile fornire «una prospettiva e una base unificante per confrontare formalismi» diversi «per espressività e complessità». Al fine poi di assicurare soluzioni indipendenti dall'infrastruttura impiegata, viene privilegiato, rispetto all'adozione di uno schema XML quale linguaggio di vincolo, un «formalismo universale generico» (*ibid.*: 61), «fondato su una base logica 'standard'», come «il calcolo dei predicati del primo ordine» o sue «estensioni 'classiche'»; un formalismo di questo tipo, infatti, permette di «esprimere le caratteristiche di altri formalismi in modo più potente mediante *specificazioni eseguibili*» (*ibid.*: 57).

11. Può essere questa la via per assicurare adeguata funzionalità alla conservazione digitale dell'informazione? Mi limito, per concludere, a due sole osservazioni. In primo luogo, l'efficacia e la funzionalità dei vincoli semantici dipendono dall'informazione espressa dalla rappresentazione XML su cui essi operano. Si ritorna quindi al problema del tipo di informazione strutturale espressa dalla rappresentazione XML dei documenti. Se ad esempio si ritiene, con David Blair, «che il recupero dei documenti (*document retrieval*) sia il recupero del loro *contenuto intellettuale*» è evidente che la rappresentazione XML può essere funzionale a questo tipo di operazione solo se riesce ad esprimere adeguatamente il contenuto concettuale degli oggetti digitali presi in esame. Infatti,

il recupero dei documenti dipende criticamente dal modo in cui i documenti sono rappresentati in un particolare sistema; questo sistema di rappresentazione consiste in una sorta di 'linguaggio' [nel nostro caso XML], in cui si possono descrivere il contenuto o il contesto del documento e in cui si possono esprimere le richieste di chi li ricerca; di conseguenza, le proprietà di questo 'linguaggio [di rappresentazione] dei documenti' possono influire sull'efficacia di tali descrizioni e di tali richieste (Blair, 2002: 275-276).

Così, se XML è impiegato solo come linguaggio per il *markup* interno dei documenti, restano i limiti, già ricordati, della sua utilizzazione come sistema adeguato di codifica del testo. Se invece XML è impiegato come formato di scambio e la rappresentazione XML è ottenuta da un altro tipo di rappresentazione strutturale, la sua adeguatezza dipende dall'adeguatezza funzionale ed esaustiva della rappresentazione d'origine. La funzionalità e l'esaustività dei vincoli semantici e delle operazioni che essi rendono possibili sulla rappresentazione XML non possono ovviamente andare oltre i limiti dell'informazione trasmessa dalla rappresentazione precedente. La testualizzazione di questa rappresentazione, ottenuta mediante l'impiego di XML come formato di scambio, non incrementa l'operatività del suo modello di dati. Nel progetto del Semantic Web, per esempio, XML funge da «meccanismo di scambio» tra rappresentazioni sintattiche espresse con schemi XML e rappresentazioni semantiche espresse con schemi RDF (cfr. *supra*, W3C, 1999) e il trasferimento dell'informazione dalle une alle altre opera sulla struttura assegnata dal *markup*, che rende la rappresentazione XML «opaca» ad un'analisi ulteriore del contenuto testuale (Buzzetti, 2003: 192).

La seconda osservazione riguarda la possibilità di esprimere con un linguaggio logico i vincoli semantici operanti sulla rappresentazione XML. L'uso di un linguaggio logico come «linguaggio di rappresentazione dei documenti» (Blair, 2002: 277), ripropone il problema della funzione del *markup* come espressione della forma logica dei testi. Se il *markup* può essere considerato come un mezzo per esprimere la forma logica del testo, la sua funzione può essere pienamente compresa solo movendo da considerazioni di filosofia del linguaggio. Infatti se è vero, come ha sostenuto David Blair, che «ogni teoria della rappresentazione dei documenti [...] deve fondarsi su una chiara teoria del linguaggio e del significato» delle espressioni linguistiche (Blair, 1990: vii-viii), è evidente che è proprio «la filosofia del linguaggio» ciò che può fornire «il necessario fondamento teorico» (Blair, 1992: 203) per valutare l'adeguatezza di ogni forma di rappresentazione digitale dei documenti.

Ora, il *markup* assegna una struttura alla sequenza di caratteri che rappresenta il testo e assegnare una struttura equivale ad assegnare una forma logica. A proposito della forma logica, la filosofia del linguaggio ci dice, con Davidson, che «se riscriviamo gli enunciati in qualche forma standardizzata, l'inferenza risulta semplificata e meccanizzata», e sappiamo inoltre che la validità delle inferenze dipende esclusivamente dalla forma sintattica degli enunciati e non dal loro contenuto semantico. Dunque la forma logica assegna una struttura all'espressione del testo. È però altrettanto vero, come abbiamo visto, che secondo Davidson «dare forma logica ad un enunciato significa» anche «ri-condurlo nell'ambito di una teoria semantica» (Davidson, 1970: 203 e 210; cfr. *supra*). Ed è proprio la coincidenza tra categorie sintattiche e categorie semantiche ciò che sta alla base dei programmi di formalizzazione promossi dai teorici dell'intelligenza artificiale per garantire la possibilità di elaborazione automatica delle conoscenze. Nel caso della rappresentazione digitale del testo, però, è possibile elaborare automaticamente la sequenza dei caratteri senza dovere necessariamente assegnare ad essa una struttura logica. Si può elaborare automaticamente il testo, o meglio la sua *espressione*, conservando l'«indeterminazione della rappresentazione del [suo] *contenuto*» (Blair, 2002: 286).

Ciò rappresenta a ben vedere un grande vantaggio, perché permette di evitare le secche di certe concezioni troppo rigide dell'intelligenza artificiale, ed è proprio a questo proposito che si può rendere più funzionale l'operazione di codifica e l'introduzione del *markup*. È stato opportunamente sostenuto, come già si è visto (cfr. *supra*, nota 5), che «descrivere il significato del *markup*» (Sperberg-Mc Queen *et al.*, 2000: 231) equivale a «stabilire quali inferenze esso autorizza (*licenses*)» (*ibid.*: 217) e anche l'espressione *inference licence* «non è nuova per un filosofo del linguaggio» (Buzzetti, 2003: 189). Nella filosofia del linguaggio, essa sta ad indicare una *regola* di inferenza espressa «ad un diverso e più sofisticato livello di discorso» di quello a cui appartengono le asserzioni alle quali si riferisce (Ryle, 1963: 116). La metalinguisticità riflessiva del *markup* (cfr. De Mauro, 1982: 93-94; 2002: 89, 91-93) può così essere sfruttata per riformulare la rappresentazione di una proprietà strutturale dell'espressione del testo, che il *markup* «esibisce in forma diacritica e autoriflessiva», in «un'asserzione esterna, o metalinguistica, espressa in un linguaggio formale di rappresentazione della conoscenza». Oppure, al contrario, l'inserimento del *markup* può essere utilizzato per «assegnare» all'espressione del testo un «modello formale» che permette di elaborarne funzionalmente

il contenuto (Buzzetti, 2003: 189-190). Questo tipo di conversione reciproca delle rappresentazioni strutturali permette di associare una rappresentazione semantica ad una rappresentazione sintattica, o viceversa, e di ottenere così una rappresentazione del testo che ne comprende entrambe le componenti e le mantiene distinte. In questo modo è possibile assumere come termine invariante e oggetto di elaborazione diretta ora l'espressione, ora il contenuto del testo e lasciare rispettivamente indeterminata l'altra componente, conservando del tutto integro il carattere virtualmente illimitato che è proprio sia della semiosi, sia dell'interpretazione dell'informazione testuale.

In sostanza, l'ambivalenza diacritica del *markup* permette di rappresentare il testo come un oggetto digitale complesso, che riesce a soddisfare al tempo stesso le esigenze di esaustività e di funzionalità della rappresentazione. Attraverso forme opportune di *markup* è infatti possibile mettere direttamente in rapporto la rappresentazione esaustiva dell'informazione testuale, che mira ad una resa fedele ed esauriente di tutte le proprietà dell'espressione del testo, con una rappresentazione funzionale del suo contenuto, che può a sua volta garantirne l'operatività ai fini dell'analisi e dell'elaborazione delle conoscenze. La sfida per la rappresentazione digitale esaustiva e funzionale di qualsiasi oggetto culturale pare così rappresentata dalla possibilità di estendere metodi di rappresentazione del testo che ne rispettino la complessa articolazione semiotica a tutte le forme di testualizzazione dei dati che si ottengono usando come formato di scambio un linguaggio di *markup*, qual è in ultima analisi XML. La cultura umanistica del testo può così recare ancora un contributo sostanziale all'intero processo di trasferimento e conservazione del patrimonio culturale in ambiente digitale.

BIBLIOGRAFIA

- ARMS W.Y., 1995. *Key concepts in the architecture of the digital library*. D-Lib Magazine, July. <http://www.dlib.org/dlib/July95/07arms.html>
- BARNARD D.T., HAYTER R., KARABABA M., LOGAN G., MCFADDEN J., 1988. *SGML-Based markup for literary texts: two problems and some solutions*. Computers and the Humanities, 22(4): 265-276.
- BLAIR D.C., 1990. *Language and representation in information retrieval*. Elsevier, Amsterdam.
- BLAIR D.C., 1992. *Information retrieval and the philosophy of language*. The Computer Journal, 35(3): 200-207.
- BLAIR D.C., 2002. *The challenge of commercial document retrieval, Part I: Major issues, and a framework based on search exhaustivity, determinacy of representation and document collection size*. Information Processing and Management, 38(2): 273-291.

- BROWN J.S., DUGUID P., 1995. *The social life of documents*. Release 1.0, EDventure Holdings Inc., New York, October: 1-18.
<http://www2.parc.com/ops/members/brown/papers/sociallife.html>
- BUZZETTI D., 1999. *Rappresentazione digitale e modello del testo*. In: *Il ruolo del modello nella scienza e nel sapere*. Atti del Convegno (Roma, 27-28 ottobre 1998). Contributi del Centro Linceo Interdisciplinare «Beniamino Segre», 100, Accademia Nazionale dei Lincei, Roma: 127-161.
- BUZZETTI D., 2000. *Ambiguità diacritica e Markup: note sull'edizione critica digitale*. In: S. ALBONICO (ed.), *Soluzioni informatiche e telematiche per la filologia*. Atti del Seminario di studi (Pavia, 30-31 marzo 2000), Pubblicazioni telematiche, n. 1, Università degli Studi di Pavia, Pavia.
http://dobc.unipv.it/dipslamm/Pubblicazioni-telematiche/dino_buz-zetti.htm
- BUZZETTI D., 2003. *Codifica del testo e intelligenza artificiale*. Schede Umanistiche, n.s., 17(1): 171-197.
- CAGLIOTI G., 1994. *Simmetrie infrante nella scienza e nell'arte*. 2a ed., Città Studi Edizioni, Milano.
- COOMBS J.H., RENEAR A.H., DE ROSE S.J., 1987. *Markup systems and the future of scholarly text processing*. Communications of the ACM, 30(1): 933-947.
- DAVIDSON D., 1970. *Action and reaction*. Inquiry, 13: 140-148; ristampato in Id., *Essays on Action and Events*, Clarendon Press, Oxford: 137-146; e in Id., *Azioni ed Eventi*, trad. it. di E. Picardi, Il Mulino, Bologna: 202-213.
- DE MAURO T., 1982. *Minisemantica dei linguaggi non verbali e delle lingue*. Laterza, Bari.
- DE MAURO T., 2002. *Prima lezione sul linguaggio*. Laterza, Bari.
- DE ROSE S. J., DURAND D., MYLONAS E., RENEAR A., 1990. *What is text, really?* Journal of Computing in Higher Education, 2(1): 3-26.
- EGGERT P., BARWELL G., BERRIE PH., TIFFIN CH., 2002. *Authenticated electronic editions project*, <http://idun.itsc.adfa.edu.au/ASEC/JITM/Charlottesville20020220PRE.pdf>
- FOLDDOC, 2003. *Free on-line dictionary of computing*, edited by Denis Howe.
<http://www.nightflight.com/foldoc/>
- HEIDEGGER M., 1973. *Kant und das Problem der Metaphysik*. Klostermann, Frankfurt am Main.
- HJELMSLEV L., 1968. *I fondamenti della teoria del linguaggio* (1943, 1961²). Trad. it., Einaudi, Torino.
- HOFMAN H., 2002a. *Some comments on preservation metadata and the OAIS model*. DigiCULT. Info, Newsletter, 2: 15-20.
http://www.digicult.info/downloads/digicult_info2.pdf
- HOFMAN H., 2002b. *Can bits and bytes be authentic? Preserving the authenticity of digital objects*. Proceedings annual IFLA conference (Glasgow, 23 August 2002).
http://daedalus.lib.gla.ac.uk/archive/00000039/01/hofman_glasgow02.pdf
- INTERPARES PRESERVATION TASK FORCE, 2001. *How to preserve authentic electronic records*. In: *The long-term preservation of authentic electronic records: findings of the InterPARES project*. Appendix 6, 3.
http://www.interpares.org/book/interpares_book_o_app06.pdf
- JOLOBOFF V., 1989. *Document representations: concepts and models*. In: J. ANDRÉ, R. FURUTA, V. QUINT (eds.), *Structured documents*. Cambridge University Press, Cambridge: 75-105.
- KAHN R., WILENSKY R., 1995. *A framework for distributed digital object services*. Corporation for National Research Initiatives. May.
<http://www.cnri.reston.va.us/home/estr/arch/k-w.html>
- LAGOZE C., 2000. *Business unusual: how "event-awareness" may breathe life into the catalog?* In: *Proceedings of the bicentennial conference on bibliographic control for the new millennium: Confronting the challenges of networked resources and the Web* (Washington, D.C., November 15-17, 2000). Library of Congress, Cataloging Distribution Service, Washington, D.C.
<http://www.cs.cornell.edu/lagoze/papers/lago-zelc.pdf>

- LATOUR B., 1986. *Visualization and cognition: thinking with eyes and hands*. Knowledge and Society: Studies in the sociology of culture past and present, 6: 1-40.
- LEVY D.M., 1994. *Fixed or fluid? Document stability and new media*. In: *Proceedings of the 1994 ACM european conference on hypermedia technology (ECTH '94)*. ACM Press, New York: 24-31.
- LODI E., PACINI G., 1990. *Introduzione alle strutture di dati*. Bollati Boringhieri, Torino.
- LUDÄSCHER B., MARCIANO R., MOORE R., 2001a. *Preservation of digital data with self-validating, self-instantiating knowledge-based archives*. ACM SIGMOD Record, 30(3): 54-63.
<http://www.acm.org/sigmod/record/issues/0109/SPECIAL/ludaescher8.pdf>
- LUDÄSCHER B., MARCIANO R., MOORE R., 2001b. *Towards self-validating knowledge-based archives*. In: K. ABERER, L. LIU (eds.), *Eleventh International Workshop on Research Issues in Data Engineering (RIDE)*. IEEE Computer Society Press, Los Alamitos, Calif: 9-16.
<http://www.sdsc.edu/~ludaesch/Paper/ride01.html>
- MCGANN J.J., 1991. *The textual condition*. Princeton University Press, Princeton N.J.
- MCGANN J.J., 2001. *Radiant textuality: literature after the world wide web*. Palgrave, New York (trad. it., *La letteratura dopo il World Wide Web*. Bononia University Press, Bologna 2002).
- MOORE R., BARU C., RAJASEKAR A., LUDAESCHER B., MARCIANO R., WAN M., SCHROEDER W., GUPTA A., 2000. *Collection-based persistent digital archives, Part 1*. D-Lib Magazine, 6(3); *Part 2*. D-Lib Magazine, 6(4).
<http://www.dlib.org/dlib/march00/moore/03moore-pt1.html>
<http://www.dlib.org/dlib/april00/moore/04moore-pt2.html>
- MOORE R.W., 2001. *Knowledge-based grids*. In: *Large Scale Storage in the Web*. 18th IEEE Symposium on Mass Storage Systems, Online Proceedings: 29-40.
<http://storageconference.org/2001/2001CD/03moore.pdf>
- NELSON M.L., ARGUE B., EFRON M., DENN S., PATTUELLI M.C., 2001. *A survey of complex object technologies for digital libraries*.
<http://techreports.larc.nasa.gov/ltrs/PDF/2001/tm/NASA-2001-tm211426.pdf>
- OCLC/RLG, 2002. *A metadata framework to support the preservation of digital objects*. A report by The OCLC/RLG Working Group on Preservation Metadata, June, Dublin, Ohio.
http://www.oclc.org/research/projects/pmwg/pm_framework.pdf
- RAYMOND D.R., TOMPA F.W., WOOD D., 1993. *Markup reconsidered*. First International Workshop on Principles of Document Processing (Washington, DC, October 22-23, 1992), TR 356, Computer Science Department, University of Western Ontario, London, Canada, 1 May.
<http://www.csd.uwo.ca/tech-reports/356/tr356.ps>
<http://xml.coverpages.org/raymmark.ps>
- RAYMOND D., TOMPA F., WOOD D., 1996. *From data representation to data model: meta-semantic issues in the evolution of SGML*. Computer Standards and Interfaces, 18(1): 25-36.
<http://db.uwaterloo.ca/~fwtom-pa/papers/sgml.ps>
- RENEAR A., DUBIN D., 2003. *Towards identity conditions for digital objects*. In: S. SUTTON J. GREENBERG, J. TENNIS (eds.), *Proceedings of the 2003 Dublin Core Conference* (Seattle, WA, September 28 - October 2, 2003).
http://www.siderean.com/dc2003/503_Paper71.pdf
- RENEAR A., MYLONAS E., DURAND D., 1996. *Refining our notion of what text really is: the problem of overlapping hierarchies*. In: S. HOCKEY, N. IDE (eds.), *Research in humanities computing 4*. Selected Papers from the ALLC-ACH Conference (Christ Church, Oxford, April 1992). Oxford University Press, Oxford: 263-280.
<http://www.stg.brown.edu/resources/stg/monographs/ohco.html>

- RENEAR A., DUBIN D., SPERBERG-McQUEEN C.M., HUITFELDT C., 2003a. *XML Semantics and digital libraries*. In: C.C. MARSHALL, G. HENRY, L. DELCAMBRE (eds.), *Proceedings of the IEEE-CS/ACM joint conference on Digital Libraries*. IEEE Computer Society Press, Los Alamitos, Calif.: 303-305.
<http://portal.acm.org/citation.cfm?id=827192>
- RENEAR A., DUBIN D., SPERBERG-McQUEEN C.M., HUITFELDT C., 2003b. *Towards a semantics for XML markup*. In: R. FURUTA, J.I. MALETIC, E. MUNSON (eds.), *Proceedings of the 2002 ACM symposium on document engineering*. ACM Press, New York: 119-126.
<http://portal.acm.org/citation.cfm?id=585081>
- RYLE G., 1963. *The concept of mind*. 2nd ed., Penguin Books, Harmondsworth.
- SEGRE C., 1985. *Avviamento all'analisi del testo letterario*. Einaudi, Torino.
- SPENCER-BROWN G., 1969. *Laws of form*. Allen and Unwin, London.
- SPERBERG-McQUEEN C.M., BURNARD L. (eds.), 1994. *Guidelines for text encoding and interchange*. Text Encoding Initiative, Chicago - Oxford.
- SPERBERG-McQUEEN C.M., HUITFELDT C., 1999. *Concurrent document hierarchies in MECS and SGML*. *Literary and Linguistic Computing*, 14(1): 29-42.
- SPERBERG-McQUEEN C.M., HUITFELDT C., RENEAR A., 2000. *Meaning and interpretation of markup*. In: *Markup languages: theory & practice*, 2(3): 215-234.
<http://www.w3.org/People/cmsmcq/2000/mim.html>
- THALLER M., 2001. *From the digitized to the digital library*. *D-Lib Magazine*, 7(2).
<http://www.dlib.org/dlib/february01/thaller/02thaller.html>
- THIBODEAU K., 2001. *Building the archives of the future: advances in preserving electronic records at the national archives and records administration*. *D-Lib Magazine*, 7(2).
<http://www.dlib.org/dlib/february01/thibodeau/02thibodeau.html>
- THIBODEAU K., 2002. *Overview of technological approaches to digital preservation and challenges in coming years*. In: *The state of digital preservation: an international perspective*. Conference Proceedings, Council on Library and Information Resources, Washington, D.C.: 4-31.
<http://www.clir.org/pubs/reports/pub107/pub107.pdf>
- UNSWORTH J., 2002. *What is humanities computing and what is not?* *Jahrbuch für Computerphilologie*, 4: 71-84.
<http://computerphilologie.uni-muenchen.de/jg02/unsworth.html>
- VANHOUTTE E., 2002. *Texts and transcriptions: mapping scribal complexities onto a line of text*.
<http://www.kantl.be/ctb/seminar/01-02/seminar-vanhoutte.htm#n3>
- VARELA F.J., 1979. *Principles of biological autonomy*. North-Holland, New York.
- W3C, 1999. *The Cambridge communiqué*. www.w3.org/TR/1999/NOTE-schema-arch-19991007
- WILLIAMS K. et al., 2000. *Professional XML databases*. Wrox Press, Birmingham.
- WITTGENSTEIN L., 1964. *Tractatus logico-philosophicus*, 4.121. In: *Tractatus logico-philosophicus e Quaderni 1914-1916* (trad. it. di A.G. Conte). Einaudi, Torino.