*Archiviazione digitale dei dati e adeguatezza della rappresentazione del testo**

Dino Buzzetti

1. La comparsa delle cosiddette 'comunità virtuali' nel già ricco e variegato panorama delle comunità scientifiche solleva problemi che investono non solo il campo della comunicazione o le forme della discussione, ma la qualità stessa della ricerca. Allo stato attuale, anche nell'ambito delle discipline umanistiche, considerare concreti progetti di lavoro interattivo e distribuito in rete sullo stesso oggetto di ricerca non significa entrare in un campo di sole possibilità future. La realizzazione di simili progetti dipende ormai soltanto dalla distribuzione delle risorse finanziarie. Non si tratta più di un problema tecnologico, ma di un problema prevalentemente politico e di organizzazione della ricerca. Per parlare di un caso concreto, è possibile fare riferimento a un progetto di digitalizzazione su larga scala per rendere accessibili in rete materiali manoscritti custoditi in archivi storici e biblioteche di conservazione.[1] A questo progetto partecipano, oltre alle istituzioni consorziate, singoli studiosi. È previ-

* Comunicazione presentata al convegno *Le comunità virtuali e i saperi umanistici: Una cultura per il nuovo millennio*, organizzato a Milano dal Centro Linguistico dello IULM nei giorni 26-28 novembre 1997.

[1] Cfr. *European Manuscript Server Initiative*, <URL: http://helmer.hit.uib.no/emsi/>. Per iniziative analoghe, cfr. il progetto *Digital Recording of Historical Archives (Digitale Erschließung von Archivbeständen)* dello Stadtarchiv Duderstadt, <URL: http://www. archive.geschichte.mpg.de/duderstadt/dud.htm>, e la biblioteca digitale *Literaturquellen zum deutschen, österreichischen und schweizerischen Privat- und Prozessrecht des 19. Jahrhunderts* del Max-Planck-Institut für europäische Rechtsgeschichte di Francoforte, <URL: http://www.mpier.uni-frankfurt.de/dlib/>.

sto che essi operino simultaneamente a distanza, nelle diverse sedi, lavorando in modo interattivo e distribuito sulle rappresentazioni digitali dei manoscritti. Essi potranno così contribuire fin dall'inizio, dall'avvio stesso della campagna di digitalizzazione, allo studio dei materiali, per organizzarli in insiemi coerenti di dati, per permetterne la ricerca in rete e per presentare in rete i primi risultati dell'esame critico dei documenti e dell'analisi dei testi.

La realizzazione di un progetto di questo genere presuppone che la ricerca faccia uso, immediatamente, della rappresentazione digitale dei testi e dei documenti e che i risultati stessi della ricerca siano prodotti in forma digitale. Sicché la comunità degli studiosi, nel suo insieme, non può non essere investita, oggi, dal problema dell'accreditamento e del riconoscimento della validità scientifica dell'edizione digitale delle fonti primarie, oltre che dal problema della valutazione e dell'accettabilità scientifica non tanto della *pubblicazione* in forma digitale, quanto della *produzione* stessa in forma digitale dei risultati della ricerca. L'uso scientifico di risorse digitali è dunque un fatto che si è reso ormai inevitabile anche nel campo delle discipline umanistiche. Ma non può essere d'altro canto solo la *presentazione* dei dati in forma digitale a garantire la validità e l'affidabilità della produzione scientifica. Merita quindi attenzione il problema della forma scientificamente valida dell'archiviazione digitale dei dati, in particolare dei dati testuali, e dei risultati della ricerca sui testi. Da quali requisiti le risorse digitali possono trarre i titoli della loro legittimità scientifica? Il problema può essere affrontato a diversi livelli, e conviene procedere con ordine.

2. Non più di qualche anno fa, Tim Berners-Lee, l'ideatore del World WideWeb, il sistema ipertestuale distribuito sull'intera rete che tutti conosciamo, ci assicurava che l'«idea originaria» era stata quella di offrire «un mezzo per comunicare attraverso la condivisione delle conoscenze». Ma al tempo stesso si rammaricava che l'auspicata trasformazione del Web in uno strumento «interattivo» per un effettivo lavoro ipertestuale di gruppo fosse ancora di là da venire. A suo parere, la ragione doveva essere trovata nella necessità di assicurare «qualità di contenuto e di produzione» per «l'insieme dell'informazione» accessibile in rete. E osservava che «l'alta qualità è ottenuta soltanto» con la separazione della prerogativa di produrre informazione, riservata a pochi, dalla possibilità di accedervi, aperta a tutti. Perciò, l'attuale struttura del Web permette a tutti di accedere ai do-

cumenti, ma riserva ai pochi che li hanno prodotti la facoltà di modificarli, o di collegarli ancora con altri documenti, o con parti di altri documenti.

Ciò significa, per un verso, che la comunità scientifica non accetta di fatto la pratica dell'uso euristico del Web per *produrre* risultati, ma solo l'uso del Web per *presentare* risultati prodotti altrimenti; e, per un altro verso, che la «qualità di contenuto e di produzione» è certo condizione necessaria, ma non è ancora condizione sufficiente a giustificare l'esigenza di ricorrere alla rappresentazione digitale nella produzione delle conoscenze. L'«idea originaria» di un Web «a cui ciascuno possa contribuire» e «in cui la persona ordinaria sia stata ri-abilitata come autore e pensatore, ossia come produttore (un *linker*) e non solo come fruitore di *link* (un *clicker*)», resta allo stato di pura aspirazione.[2] Si ricorre alla rappresentazione digitale soltanto per presentare il prodotto, non per produrre il risultato della ricerca. L'accreditamento della pratica di ricerca fondata sull'uso e sull'elaborazione di dati in forma digitale per produrre risultati in forma digitale non può dunque dipendere dal semplice controllo della produzione scientifica, una condizione che riguarda qualunque forma di pratica scientifica e non solo, specificamente, una pratica fondata sulla necessità dell'uso della rappresentazione digitale dei dati e delle conoscenze. Tale necessità non può che dipendere dalla natura e dalle caratteristiche intrinseche della rappresentazione digitale. Giova quindi riflettere sulle peculiarità dei formati digitali per ottenere forme adeguate di archiviazione dei dati in grado di garantire l'idoneità delle risorse digitali per fini di ricerca.

3. Soffermiamoci ancora un poco sul WorldWideWeb. Esso non consente, in quanto tale, nessuna rappresentazione ed elaborazione semantica dei materiali non solo testuali, ma multimediali in genere, che vi sono presentati. Il Web può essere infatti descritto solo come Traditional MultiMedia (TMM), una qualifica che può essere applicata ad un sistema in cui «il computer non ha alcuna comprensione di ciò che sta presentando», ma è concentrato soltanto sul «modo migliore di presentarlo». Infatti la codifica HTML usata per predisporre i materiali presentati sul Web, non va molto oltre la specificazione degli aspetti che riguar-

[2] T. Berners-Lee, *Foreword*, in P. Flynn, *The WorldWideWeb Handbook: An HTML guide for users, authors and publishers*, London-Boston, International Thompson Computer Press, 1995, pp. viii-ix.

dano la formattazione e la visualizzazione dei dati.[3] È vero che un *file* HTML può essere generato all'istante da uno *script* o da un programma eseguito nel momento in cui viene attivato il *link* che vi fa riferimento, ma l'elaborazione viene compiuta da un sistema esterno al Web, che ne «riformatta» semplicemente i risultati «come un *file* HTML da inviare all'utente».[4] Non è quindi possibile qualificare il Web come Intelligent MultiMedia (IMM), cioè come un sistema in grado di elaborare ed integrare informazione di tipo diverso (testuale, parlata, sonora, visiva), ove con ciò si intenda la rappresentazione digitale dell'informazione di vario tipo e la capacità di generarla e di interpretarla *semanticamente*. Nel Web, l'aspetto più direttamente semantizzabile è costituito dalla struttura dell'insieme dei *link*, che può essere concepita come un grafo, ai cui nodi e archi siano stati associati particolari attributi o descrizioni della natura dei contenuti collegati e del tipo di relazione che li collega. Questo consente di ottenere una forma di rappresentazione delle conoscenze accessibili nel Web elaborabile automaticamente ed è in questo senso che si è effettivamente orientato lo sviluppo di alcune tecniche di *information retrieval* (IR) applicate alle strutture ipertestuali.[5] Ma tali applicazioni non possono investire più di tanto la rappresentazione e l'elaborazione dei materiali primari. Una volta selezionati e localizzati, questi vengono in pratica solo presentati, senza che il loro contenuto possa venire ulteriormente elaborato. Il risultato dell'operazione di IR serve sostanzialmente per orientare e dirigere la navigazione e la visualizzazione. Chi elabora i materiali primari non è il computer, ma continua ad essere l'occhio del lettore che li ha rintracciati. E l'occhio elabora attraverso la forma percettiva, una forma, quella presentata dal Web, che assomiglia ancora molto, per i materiali testuali, a quella tradizionale del libro.

4. A questo proposito mette conto richiamare le considerazioni dedicate da Ivan Illich, in un acutissimo saggio,[6] al ruolo decisivo svolto

[3] Aalborg University, *International Master of Science Programme in Intelligent MultiMedia*, s.d., p. 2.

[4] Flynn, *The WorldWideWeb Handbook* cit., p. 203.

[5] Cfr. M. Agosti and A. Smeaton (eds.), *Information Retrieval and Hypertext*, Norwell, Mass.-Dordrecht, Kluwer, 1996.

[6] I. Illich, *Nella vigna del testo: Per una etologia della lettura* (1993), trad. it. di A. Serra e D. Barbone, Milano, Cortina, 1994.

dall'introduzione di nuovi criteri di disposizione della pagina nella nascita e nello sviluppo della nostra concezione del testo come «oggetto visibile» ma al tempo stesso «intangibile»(119), come «esteriorizzazione» e riflesso visivo «di una struttura di pensiero»(108). Come Illich illustra con abbondanza di riferimenti, verso la metà del XII secolo il libro cessa di essere traccia, o «registrazione della parola», della parola detta e ascoltata, e diviene la «registrazione del pensiero»(99), o lo «specchio del concetto»(97). E la trasformazione si compie attraverso l'introduzione di nuovi «dispositivi di ordinamento della pagina». Così, il testo trascritto si piega «all'immagine mentale della sua struttura»(101) e «la nuova impaginazione» (109) diviene il modo di «proiettare sullo spazio vuoto della pagina modelli di 'sapere' organizzati e quantificati mentalmente»(101); nella nuova disposizione della pagina, l'uso delle rubriche e dei titoli riflette «la volontà di usare l'articolazione visiva come mezzo di interpretazione» (109). Ed è proprio «l'affermarsi di questa architettura visiva che rende sempre più necessario, quando si legge, che si abbia il libro sotto gli occhi» (103). Non basta più ascoltare la lettura di un altro. Pietro Lombardo si preoccupa che la pagina venga organizzata in maniera tale che «chi legge non abbia bisogno di sfogliare molti volumi, ma possa trovare rapidamente e senza fatica ciò che lo interessa»,[7] come in una specie di ipertesto visivo. Che cosa significa allora sottrarre il testo all'occhio, rendendone indifferente la presentazione, o anche solo modificare la forma della sua percezione visiva? Quale nuova concezione del testo si affaccia con questa trasformazione? Infatti è proprio alle «nuove convenzioni grafiche» adottate dagli scribi che si deve «la conversione del libro da indicatore della *natura* a indicatore della *mente*»(124). Da «simbolo della realtà» il libro diventa «simbolo del pensiero»(125) e con la «rivoluzione scribale»(120) del XII secolo nasce «il testo in quanto oggetto» libresco(119), «materializzazione» visiva «dell'astrazione» mentale(120), che le tecniche meccaniche introdotte intorno al 1460 «reificano sotto forma di stampato»(120).

In modo del tutto analogo, con la rappresentazione digitale del testo nasce oggi una nuova forma di testualità, che tuttavia è ancora alla ricerca del suo peculiare tipo di canonicità. Di qui l'importanza del modo e delle forme nuove in cui si rimodella la percezione visiva dei materiali presentati sul Web. E se si tratta dell'edizione di un testo antico? È in grado,

[7] *Patrologia latina*, vol. 192, p. 522.

tale rimodellazione, di rispecchiare fedelmente l'organizzazione concettuale disegnata sull'originale dall'architetura visiva dell'impaginazione? Riesce a cogliere, tale riproposizione visiva, tutti gli elementi del paratesto che non ne possono essere espunti, pena l'impoverimento, se non la compromissione stessa della natura del testo? Sono, queste, domande che non possono essere eluse, sebbene restino ancora sul piano della visualizzazione e della riconfigurazione percettiva del testo. Sicché, anche sul piano della pura presentazione il Web rischia di ridursi a rappresentazione impoverita della fonte, a meno che non venga consapevolmente affrontato il problema dell'adeguata restituzione dell'informazione testuale a seguito della migrazione sul nuovo *medium*. Non resta dunque che affrontare direttamente il problema della natura e della qualità della rappresentazione digitale.

5. Restando al testo, dovrebbe essere il *markup* ciò che garantisce l'adeguatezza della rappresentazione. Ora, il *markup* può essere considerato come una specie di trascrizione diplomatica ad uso del computer, che deve essere avvertito, con marche esplicite, dell'informazione non rappresentabile mediante la semplice stringa dei caratteri codificati. L'uso del *markup* pone però un altro problema. Un linguaggio di *markup* è solo un linguaggio di rappresentazione e non è un modello di dati. Il *markup* è solo l'espressione esplicita della struttura di dati che *rappresenta* il testo, anziché la formalizzazione del suo modello astratto. Sicché occorre preliminarmente chiedersi: quali strutture di dati realizzano il modello analitico dello studioso? e, quale tipo di elaborazione ne realizza il procedimento interpretativo? Ma procediamo per gradi e cerchiamo di chiarire meglio la natura del *markup*.

È noto che HTML, il linguaggio di codifica adottato dal WorlWideWeb è solo una forma impoverita dei linguaggi di *markup* ricavabili da SGML.[8] È noto anche che per la codifica dei testi letterari sono state emanate le *Guidelines* della Text Encoding Initiative (TEI), che rappresentano la proposta più organica, avanzata dalla comunità degli studiosi, per un modello di rappresentazione digitale dell'informazione te-

[8] Cfr. ISO 8879:1986, Information processing – Text and office systems – Standard Generalized Markup Language (SGML). Si tratta dello standard ISO per sistemi di *markup* descrittivo. Il *markup* descrittivo può essere definito come l'uso di nomi mnemonici per i diversi elementi della rappresentazione del testo.

stuale fondato su una codifica conforme a SGML.[9] Sono anche noti, e già usati diffusamente, molti strumenti sviluppati specificamente per il trattamento di *file* SGML, sia per la realizzazione a sé stante dell'edizione elettronica (p.es., DynaText), sia per la sua presentazione in rete (p. es., DynaWeb, o Panorama). Meno diffusa è tuttavia la consapevolezza della discussione sulla portata e i limiti dell'uso di SGML come linguaggio di rappresentazione dei dati testuali. Ed è solo su questo piano che può essere affrontata la valutazione dell'adeguatezza scientifica della rappresentazione digitale del testo. Intendo riferirmi alla definizione del testo, o più precisamente della natura della sua struttura formale, proposta dai promotori stessi della codifica descrittiva del testo su base SGML. Alla domanda *What is Text, Really?* Steven De Rose e gli altri coautori dell'omonimo saggio rispondevano definendo l'ontologia formale della sua rappresentazione in formato SGML: il testo è OHCO (Ordered Hierarchy of Content Objects), ossia una gerarchia ordinata di oggetti dotati di contenuto, e ha la struttura di un grafo che ne organizza i contenuti come oggetti astratti, ordinati in modo sequenziale e lineare, e rappresentati da segmentazioni successive e contigue del discorso.[10] Su questa base si è sviluppata la «pratica» di codifica orientata dalle *Guidelines* della TEI, ma la «markup theory» che se ne evince mette già in luce per bocca dello stesso Allen Renear, uno dei sostenitori della definizione del testo come OHCO, le prime difficoltà dell'assimilazione della struttura logica del testo al modello formale della gerarchia ordinata.[11] Infatti tale tesi non tiene conto del fatto che «talvolta lo stesso documento si conforma a diverse strutture sovrapposte», le quali «non possono essere inserite nella stessa struttura gerarchica», di modo che «la sua elaborazione varia a seconda della particolare struttura considerata».[12] Né può te-

[9] Cfr. C. M. Sperberg-McQueen and L. Burnard (eds.), *Guidelines for Text Encoding and Interchange*, Chicago and Oxford, Text Encoding Initiative, 1994.

[10] Cfr. S. J. DeRose, D. D. Durand, E. Mylonas, A. H. Renear, *What is Text, Really?*, «Journal of Computing in Higher Education», I, 2, 1990, pp. 3-26.

[11] Cfr. A. Renear, E. Mylonas, D. Durand, *Refining our Notion of What Text Really Is: The problem of overlapping hierarchies*, in N. Ide (ed.), *Research in Humanities Computing*, vol. 4 (Selected Papers from the ALLC/ACH Conference, Christ Church, Oxford, April 1992), Oxford, Clarendon Press, 1996, pp. 263-280.

[12] D. Barnard, R. Hayter, M. Karababa, G. Logan e J. Mc Fadden, *SGML-Based Markup for Literary Texts: Two Problems and Some Solutions*, «Computers and the Humanities», 12, 1988, p. 266.

nere conto di relazioni strutturali di tipo non gerarchico, che pure si possono presentare nell'analisi del testo, poiché «le strutture logiche rilevanti di un certo testo non sono senza eccezione di tipo gerarchico».[13] Si può quindi affermare, come ho cercato di argomentare altrove, che

> assumendo l'SGML come base per la definizione delle norme e dei linguaggi di codifica, la TEI assume implicitamente questo modello strutturale come modello fondamentale di rappresentazione del testo e vincola di conseguenza la forma di rappresentazione del testo ad una struttura di dati affatto particolare, che non permette di riunire in una singola rappresentazione coerente modelli strutturali diversi e alla quale non pare così applicabile un modello computazionale pienamente adeguato alle necessità dell'analisi e dell'interpretazione del testo.[14]

Più che ribadire questa convinzione, qui importa semplicemente indicare i problemi che occorre affrontare per potere proporre su basi teoricamente adeguate una forma di rappresentazione digitale del testo come forma scientificamente valida a tutti gli effetti. La difficoltà fondamentale sembra consistere nel fatto che i sistemi di elaborazione dei documenti basati sul tipo di *markup* realizzato con SGML sono sostanzialmente «definiti nei termini di quella struttura di dati sequenziale che si presume sia il 'testo'».[15] Ma com'è stato convincentemente sostenuto, «il *markup* non è un modello di dati, bensì un tipo di rappresentazione».[16] Sicché il modello del testo che viene rappresentato con un certo documento finisce per essere ricavato proprio dal modello della sua rappresentazione. Ma è da un modello astratto del testo che deve essere ricavata la struttura di dati che lo rappresenta e non viceversa; non ci si può riferire

[13] RENEAR, et al., *Refining our Notion of What Text Really Is* cit., p. 279, nota 13.

[14] D. BUZZETTI, *Il testo 'fluido': Sull'uso dell'informatica nella critica e nell'analisi del testo*, in *Filosofia & informatica*, Atti del primo incontro italiano sulle applicazioni informatiche e multimediali nelle discipline filosofiche (Convegno Nazionale della Società Filosofica Italiana: Roma, 23-24 novembre 1995), a cura di Luciano Floridi, Torino, Paravia, 1996, p. 91.

[15] D. R. RAYMOND, F. W. TOMPA and D. WOOD, *Markup Reconsidered*, paper presented at the First International Workshop on Principles of Document Processing, Washington DC, October 22-23, 1992 <URL: http://www.csd.uwo.ca/staff/drraymon/.papers/markup.ps>, p. 16.

[16] Ibid.

alle proprietà del tipo di rappresentazione sintattica della struttura di dati, fornita da un linguaggio di *markup*, per ricavare il modello astratto che tale struttura dovrebbe rappresentare. Piuttosto, occorre prendere le mosse, euristicamente, dalle procedure analitiche e interpretative che vengono applicate allo studio dei testi dalla critica testuale e letteraria. Mi limiterò ad un esempio che mi è familiare e che ci permette di riflettere sul concetto stesso di edizione digitale.

6. L'esempio qui proposto è tipico di una forma di testualità affatto difforme da quella che si è imposta con l'introduzione della stampa e ha portato all'instaurazione di un'idea di canonicità testuale legata alla natura fissa, immobile, «monumentale» quasi, del libro stampato.[17] Molti testi medievali di produzione universitaria nascono direttamente dalla pratica di insegnamento e sono il risultato delle *reportationes* degli studenti, che annotavano il contenuto delle lezioni ascoltate dalla viva voce dal maestro. Si tratta di oggetti testuali instabili, mutevoli e 'fluidi', come testimonia l'estrema variabilità della tradizione manoscritta e delle stesse prime edizioni a stampa. Come può essere realizzata l'edizione di tali materiali? Un'edizione può essere concepita come la rappresentazione delle ipotesi critiche dell'editore sulla natura del testo. Nel caso in esame, la forma di rappresentazione digitale si impone come l'unica praticabile, poiché occorre in ogni caso partire da una rappresentazione completa dell'intera tradizione testuale. E ciò può essere efficacemente realizzato solo con un *database*. La rappresentazione in forma di *database* si presenta infatti come la più idonea a riprodurre la natura 'fluida' del testo. Il testo non può essere pensato, in questo caso, che come «somma logica»[18] delle diverse rappresentazioni prodottesi nel corso della sua stessa trasmissione. Di qui la necessità di una rappresentazione strutturale non lineare, da cui possano essere dinamicamente estratte diverse ricostruzioni ipotetiche del testo, ciascuna pensata come una realizzazione particolare del modello complesso che ne rappresenta, in modo organico, l'intera

[17] K. D. Uitti, *Old French Manuscripts, the Modern Book and the Electronic Image*, in ACH-ALLC'93 Joint International Conference (16-19 June 1993), *Conference Abstracts*, Georgetown University, Washington, D.C., 1993, p. 157.

[18] M. Thaller, *Historical Information Science: Is There Such a Thing? New Comments on an Old Idea*, in *Discipline umanistiche e informatica*, a cura di T. Orlandi, Roma, Accademia Nazionale dei Lincei, 1993, p. 64.

tradizione. L'editore può proporre dunque, di volta in volta, una rappresentazione sequenziale particolare, ricavata dal complesso modello non lineare che rappresenta la totalità delle diverse stratificazioni del testo. A sua volta l'interprete, movendo da una singola rappresentazione sequenziale, può assegnarle strutture diverse a seconda delle necessità interpretative di volta in volta individuate. Dall'unica rappresentazione complessa, somma logica delle diverse letture interpretative, può essere ricavato il modello che ne compendia le diverse ricostruzioni ermeneutiche. Di qui la necessità di individuare un modello di dati adeguato e funzionale a queste diverse operazioni di rappresentazione e di elaborazione dell'informazione testuale. Un'edizione digitale si contraddistingue così, rispetto ad altre forme di edizione non digitale, per le sue effettive proprietà operative. La realizzazione dell'edizione viene così a dipendere dalla scelta del modello computazionale che le rende possibili e una forma di archiviazione digitale dei dati che soddisfi legittimi requisiti di adeguatezza scientifica può essere pensata solo alla luce di queste funzionalità operative. La modalità dell'archiviazione, insomma, non è altro che lo specchio fedele della qualità della ricerca.