Paolo Buonora

INDICIZZAZIONE AUTOMATICA E COLLEGAMENTO TRA BASI DI DATI

Prima di parlare di problemi di indicizzazione vorrei riprendere i temi discussi con più interesse nella mattinata. Ci si è soffermati con particolare attenzione sulle caratteristiche peculiari della documentazione che viene prodotta attualmente: tanto per la documentazione informatica quanto per la documentazione di tipo audio–visivo appare determinante la presenza di un mezzo meccanico. Il grosso cambiamento rispetto alla documentazione meno recente è che mentre lo scrivano del primo periodo dell'unità d'Italia impiegava un tempo considerevolmente lungo a produrre un documento, il ritmo di produzione dei documenti è oggi quasi vertiginoso. Giustamente, si diceva stamattina che non è assolutamente pensabile conservare tutto quello che viene prodotto, anche invocando l'aiuto del ritrovamento automatico dell'informazione: la mole stessa della produzione documentale non è assolutamente controllabile, nemmeno con strumenti automatici. Un esempio significativo è quello dell'amministrazione svedese, che ha stabilito la distruzione entro un determinato lasso di tempo di tutta la documentazione prodotta su supporti inadatti alla conservazione permanente e il contemporaneo trasferimento di parte dell'informazione su supporti più stabili.

Vedevamo stamattina l'esempio del nastro magnetico che perde con grande facilità le proprie capacità di conservare l'informazione, in un periodo di tempo che in definitiva non supera i dieci anni; è da aggiungere che in ogni caso anche le procedure più avanzate nella produzione di questi supporti, la loro conservazione a temperatura costante e controllata, il periodico riavvolgimento o ricopiatura su altri nastri ogni cinque o dieci anni, non basta a dare una risposta al problema: il ritmo di crescita della documentazione prodotta, e quindi di quella da conservare, renderà impossibile compiere tutte queste operazioni per tutti i nastri.

La risposta sembra venire dalla tecnologia dei video–dischi a lettura laser: il video–disco non è un supporto di natura magnetica, e quindi non è soggetto al deterioramento di cui si diceva: è un disco di metallo protetto da una

pellicola di materia plastica in cui le informazioni sono state tradotte in scalfitture microscopiche, le quali vengono « lette » da un raggio laser. L'informazione naturalmente è di tipo digitale, e il mezzo offre capacità di conservazione nel tempo superiori probabilmente a quelle del supporto cartaceo più tradizionale, che si può considerare più resistente di supporti usati adesso: le stampatrici dei computer non rispondono assolutamente per gli inchiostri ed il tipo di carta usata alle caratteristiche richieste per la conservazione permanente.

I video–dischi li abbiamo ormai sotto gli occhi nei negozi di musica: da tempo è stato messo in commercio il lettore laser, e in alta fedeltà si vendono comunemente i compact–disk. Questa tecnologia si sta dunque diffondendo a livello commerciale; tuttavia si deve considerare l'alto costo del trasferimento dell'informazione su un supporto del genere, perché l'incisione della documentazione su video–disco può essere giustificata dalla vasta diffusione di un prodotto di alta fedeltà, ma è cosa ben diversa quando si tratta di produrre un solo esemplare per la conservazione permanente.

Teoricamente il mezzo si presta tanto alla conservazione di informazione di tipo verbale — e in questo offre capacità di stoccaggio veramente considerevoli, vale a dire in uno di questi dischi si può immagazzinare l'equivalente di 500.000 pagine di libro — quanto alla conservazione di altri tipi di informazione: data l'elevata quantità di informazioni digitalizzate che può contenere, il disco–laser offre la possibilità di conservare suoni e immagini. In tal caso i costi cominciano a salire vertiginosamente, mentre il rapporto costo/rendimento è più ragionevole per la documentazione di tipo verbale.

Tutto questo per rispondere ai problemi che Ortoleva sollevava stamattina. C'è però un altro aspetto della questione: quando anche saremo riusciti a conservare per un periodo di tempo considerevole questa gran mole di documentazione, adeguandoci ai suoi ritmi di crescita, come faremo a ritrovare l'informazione? Con migliaia e migliaia di bobine magnetiche e di film conservati, chi sarà in grado di vederli tutti, per sfruttare le informazioni in esse contenute?

Il problema a me pare anche più grosso di quello puramente tecnico della conservazione: indubbiamente, se noi guardiamo a situazioni in cui gli archivi automatici e audio–visivi si sono sviluppati notevolmente, vediamo che ci si trova a livelli critici.

Vorrei citare il caso dei National Archives in USA, ove è stata stabilita un'indicizzazione dei versamenti basata su livelli gerarchici di produzione della documentazione, vale a dire un'indicizzazione automatica in base alla

quale i documenti possono essere ritrovati seguendo la via « dipartimento di Stato–divisione–ufficio–competenza », anche scendendo ad un buon livello di differenziazione: gli archivisti americani, che pure dispongono di un bilancio considerevole, non sono assolutamente in grado di procedere in questo caso al tradizionale lavoro dell'archivista di riordinare il fondo ed inventariarne il contenuto, ricorrendo ad una classificazione predefinita, scegliendo i termini da indicizzare documento per documento o anche serie per serie, e creando una serie di riferimenti incrociati tra gli stessi; se dovessero dedicarsi ad un'indicizzazione di questo genere, tra sessanta anni avrebbero finito di lavorare sulla documentazione di cui sono attualmente in possesso, ed avrebbero naturalmente da iniziare il lavoro per la documentazione prodotta nel frattempo con un ritmo di crescita esponenziale.

Pertanto, riguardo a tale valanga di documentazione, il problema del ritrovamento dell'informazione è forse quello che abbiamo più urgentemente di fronte.

Ci sono diversi metodi per creare dei sistemi di *information retrival*, e diverse esperienze sono state compiute sul piano applicativo. Credo che ieri si sia parlato abbastanza del sistema MISTRAL, che è usato dall'archivio storico dell'Ansaldo, ed è un *software* prodotto dalla consociata francese della Honeywell: il sistema è usato in alcune basi di dati degli Archives Nationales di Parigi.

Si tratta di un pacchetto di programmi che offre diverse possibilità. Anzitutto si può effettuare la ricerca detta « booleana », vale a dire l'incrocio tra diverse parole chiave: posso cercare le unità documentarie che contengono una certa parola, quelle che contengono diverse parole associate, quelle che contengono una certa parola ma non ne contengono un'altra e così via, allargando o restringendo il campo d'indagine. Vi è poi la possibilità di condurre una ricerca sui cosiddetti *thesauri*: un *thesaurus* è semplicemente l'insieme delle parole–chiave o termini–indice usati per l'insieme di dati, e può avere struttura diversa. Facciamo l'esempio di una delle basi di dati francesi citate, LEONORE, che riguarda le persone insignite della Legion d'onore; si tratta di dati essenzialmente biografici, ricavati da un fondo di circa 218.000 fascicoli, con documentazione di tipo molto omogeneo sulle persone interessate da onorificenze, pensioni di guerra e così via; il *thesaurus* che corrisponde a questa base di dati ha una struttura alfabetica o cronologica: si può cercare il nome Martin e chiedere la lista dei termini vicini alfabeticamente (Martinez, Marty, ecc.). Questo è un livello di aiuto alla ricerca a mio avviso abbastanza basso: è possibile ricorrere ad espedienti di tipo « enigmistico », troncando le

parole o aggiungendo altri prefissi, ma tale metodo è sostanzialmente limitato per applicazioni di carattere più concettuale.

Gli archivisti francesi hanno impiegato un *thesaurus* di diverso tipo nella costituzione di una base di dati tratta dal fondo delle Belle arti, nella quale trattando di affari amministrativi relativi alle varie opere d'arte (vendita, commissione, assegnamento a un museo) dovevano poi arrivare a descrivere l'opera d'arte stessa: per far questo hanno costruito un *thesaurus* gerarchizzato, in cui si parte da un primo livello di cinque rami (scena, personaggio, simbolismo, soggetto diverso dall'uomo, motivo decorativo) e si arriva per ramificazioni successive a soggetti più specifici. Pe fare un'altro esempio, l'archivio dell'Ansaldo ha creato un *thesaurus* gerarchico per i termini tecnici di carattere navale.

Anche questo tipo di *thesaurus* presenta dei grossi inconvenienti, che riguardano le difficoltà di applicazione di una classificazione gerarchica a soggetti non predefiniti: nella classificazione di Dewey, organizzata in scatole sempre di dieci elementi, io devo sapere per arrivare alla « storia di Mondovì » che devo partire da « storia » ma non da « storia universale » bensì « storia d'Italia », scendere a « storia del Piemonte » per arrivare infine alla « storia di Mondovì ». Il problema è che posso trovare riferimenti che mi interessano non solo seguendo quell'unica strada, per arrivare alla fogliolina che è sull'ultimo dei rami, ma anche in rami diversi (p. es. « storia economica » o « storia religiosa »): tuttavia non ho relazioni che mi mettano direttamente in collegamento questi soggetti « distanti » e situati a vari livelli della gerarchia.

Ora, se vogliamo cercare situazioni e problemi più simili a quelli degli archivi italiani, è opportuno guardare ad un'ambito europeo: in USA gli archivisti hanno a che fare prevalentemente col problema degli archivi contemporanei oppure vi sono università abbastanza ricche da creare basi dati *ad hoc* e compiere ricerche finalizzate a un determinato argomento. Le esperienze più interessanti in materia di indicizzazione automatica mi sono sembrate pertanto quelle dei colleghi inglesi.

Al Public Record Office di Londra sono state sperimentate applicazioni di sistemi creati per la documentazione bibliografica, tendenti essenzialmente alla generazione automatica di indici: nel sistema PRECIS ad esempio una *index-entry* è costituita da una stringa di termini che possono essere permutati sintatticamente, generando in modo automatico altre *index-entries*. Inizialmente si sono avuti degli insuccessi, in quanto effettivamente l'informazione bibliografica presenta caratteristiche diverse da quella archivistica; tuttavia

si è giunti alla definizione di un tipo di *thesaurus* molto più « flessibile » di quelli di cui parlavo prima.

Ai fondi descritti nella *Current Guide* degli archivi del Public Record Office è stato applicato un sistema di automazione (PROSPEC) che inizialmente (1972–1974) comportava semplicemente la produzione automatizzata di edizioni aggiornate del testo e degli indici, e successivamente si è sperimentato con successo un sistema di indicizzazione basato sulla classificazione già adottata per la struttura della *Current Guide*. Si trattava di una classificazione per categorie generali, semplice e facilmente padroneggiabile da un punto di vista logico, una sorta di metalinguaggio esterno rispetto all'indice vero e proprio, cioè senza riferimento alla serie archivistiche: sotto queste comuni denominazioni (*catch–words*) si faceva riferimento a una serie di « entrate » (*entries*), o termini–indice: sotto « commercio » si poteva quindi trovare « industria », « produzione », « commercio di guerra », « commercio alimentare » e così via; scorrendo poi l'indice formato da questi termini si trovavano i rinvii alla documentazione, ma anche riferimenti ad altri termini–indice o ad altre categorie della classificazione generale esterna all'indice.

Detto per inciso i termini della classificazione erano scelti arbitrariamente mentre i termini dell'indice erano ricavati dal linguaggio naturale dei documenti; era stata inoltre abbandonata l'idea di generare automaticamente *index–entries* permutando l'ordine dei termini: il risultato era sempre una sovrabbondanza di voci, poco funzionale al ritrovamento dell'informazione.

Facendo riferimento ad una classificazione esterna era dunque possibile entrare e uscire dall'indice in qualsiasi punto senza compiere percorsi gerarchici; perfezionando all'interno dell'indice il sistema di riferimento o in alto, cioè a termini più generali, o in basso a voci di indice più dettagliate, o in parallelo a voci messe sullo stesso livello, si è arrivati alla struttura logica di *thesaurus* che viene chiamata in informatica « grafo », nella quale non vi è più gerarchia nemmeno tra una classificazione generale e l'insieme dei termini–indice. In un *thesaurus* di questo tipo è possibile arrivare al fondo o alla serie accedendo immediatamente all'indice se già si dispone di un termine preciso: sotto « debito pubblico » trovo già i riferimenti a determinate serie archivistiche, senza dover seguire un percorso logico « finanza – ... – debito pubblico » (ciò è importante, perché sotto « finanza » potrei trovare molte sotto–categorie, che rimandano a loro volta ad ulteriori sottopartizioni, e per leggerle tutte sarei obbligato a scorrere diverse pagine di indici); se viceversa devo arrivare al termine–chiave per passaggi successivi attraverso l'indice, posso arrivarci attraverso più percorsi: da finanza a debito pubblico

posso trovare un riferimento incrociato che passa attraverso due termini (« finanza pubblica – finanza governativa »), oppure per un passaggio solo (« debiti »).

Non essendovi una struttura predeterminata che limiti il numero delle *index-entries*, nel corso di questo lavoro di indicizzazione si possono creare continuamente nuovi riferimenti, nuove relazioni, o evidenziare relazioni fra termini già presenti nell'indice.

Questa « flessibilità » sembra oggi il requisito fondamentale per tutte le basi di dati che vengono prodotte. Si tenga presente innanzitutto che il controllo del linguaggio impiegato è un punto fondamentale nella creazione di una base di dati: dalle esperienze citate, come anche quelle personali nella redazione della *Guida generale degli Archivi di Stato italiani*, emerge la raccomandazione che le entrate all'indice siano o stese da una sola mano, o per lo meno controllate da una redazione centrale.

Riguardo poi alle metodologie di progetto delle basi di dati è stato teorizzato ampiamente sulla distinzione tra sistemi di ritrovamento dell'informazione e sistemi di gestione di basi di dati, ma la distinzione più importante è forse questa: le basi di dati sono sistemi in cui il produttore è diverso dall'utente, e le valenze che uno stesso termine può assumere da un punto di vista semantico o l'utilizzo che se ne può fare variano di molto rispetto a ciascuno degli utenti possibili della base di dati; negli archivi ad esempio ci troviamo ad avere documentazione utilizzata abitualmente da storici, architetti, avvocati: ciascuna di queste utenze può dare alla stessa parola un significato profondamente diverso, e allo stesso modo la struttura della base di dati deve rispettare diversi « schemi-utente »; ma su questo argomento fornirà ulteriori approfondimenti il dott. Lo Sardo nel prossimo intervento.

Questo aspetto « multiforme » dell'informazione ci introduce ad una questione di una certa attualità. Lo sviluppo tecnologico dei piccoli computer offre oggi la possibilità non solo di un'utenza, ma di una produzione « disseminata » di basi di dati; ciò apre molte possibilità, anzitutto in quanto avvicina le basi di dati agli utenti che le producono, ma anche perché crea uno scenario del tutto diverso da quello prima tradizionale: agli inizi dell'« era informatica » si organizzavano faraonici lavori di raccolta e schedatura di dati, compiuti non *on-line* ma in *batch-processing*, portando ad un elaboratore centrale di grandi dimensioni il pacco di schede perforate o il nastro che veniva quindi " *inputtato* " dal sistema. Oggi abbiamo viceversa la possibilità di una produzione diretta e disseminata di informazioni tramite *software* e *hardware* diversi e della creazione di sistemi informatici comuni mediante « traduttori »:

come in un convegno internazionale nel quale ciascuno continua a parlare la sua lingua, ciascun ente può produrre la sua parte di dati e metterla poi in comune con altri seguendo certi standard, che devono necessariamente essere stabiliti a livello locale, nazionale, internazionale; la fruibilità comune dei dati si può acquistare successivamente, tramite l'impiego di programmi di traduzione da un *software* all'altro.

La definizione e l'impiego di questi standard dovrebbe impegnare tutti coloro che operano nel campo della documentazione, ma sicuramente in maggior misura chi lavora in organismi statali ed ha più relazioni di carattere internazionale, perché solo tramite questi standard l'informazione può « scorrere » nel mercato dell'informazione da un punto all'altro della rete che si viene a creare; molti istituti di ricerca, pubblici e privati, memorizzano oggi dati per propri fini: se nel farlo seguiranno degli standard sarà possibile in tempi molto prossimi vendere questi dati, indipendentemente dagli utilizzi fatti internamente per le proprie ricerche, su un mercato internazionale della informazione che si sta sviluppando in maniera impetuosa. Molte ricerche documentarie si possono ormai svolgere con facilità dalla propria scrivania tramite terminale; le basi di dati che noi produciamo, che tutti possono produrre a qualsiasi livello grazie allo sviluppo tecnologico dei micro e mini-computer, debbono ormai guardare in questa direzione, cioè nella direzione di un mercato dell'informazione in sviluppo e non solo di ricerche settoriali, parcellizzate e specialistiche.

BIBLIOGRAFIA

ARCHIVI AUTOMATICI – NUOVE TECNOLOGIE: C.M. DOLLAR, *Computers, the National Archives and Researches*, in « Prologue », 8, n. 1 (Spring 1976) pp. 29–34; ID., *Appraising Machine-Readable Records*, in « The American Archivist », 41, n. 4 (October 1978) pp. 423–430; L. BELL, *The Archival Implications of Machine-Readable Records*, in « Archivum », XXVI (1979) pp. 85–92; M.H. FISHBEIN, *Guidelines for Administering Machine-Readable Records*, Committee on automation, ICA, 1980; M. ROPER, *Les archives et les nouvelles techniques de l'informatique*, in « Revue de l'UNESCO pour la science de l'information, la bibliothéconomie et l'archivistique », IV, n. 2 (avril–juin 1982) pp. 115–122; ID., *Advanced Technical Media: the Conservation and Storage of Audiovisual and Machine-readable Records*, in « Journal of the Society of Archivists », 7, n. 2 (October 1982); J.M. GRIFFITHS, *Les tendences principales dans la tecnologie de l'information*, in « Revue de l'UNESCO pour la science de l'information, la bibliothéconomie et l'archivistique », IV, n. 4 (octobre–décembre 1982) pp. 250–259.

INDICIZZAZIONE AUTOMATICA: A. CALMES, *Practical Realities of Computer-based Finding Aids: The NARS A-1 Experience*, in « The American Archivist », 42, n. 2 (April 1979) pp. 167–177; DIRECTION DES ARCHIVES DE FRANCE, *Information historique et gestion documentaire: les bases des données des archives*, in « Note d'information », n. 13 (octobre 1979); H. L'HUILLIER, *L'application PRIAM à la cité des Archives Contemporaines de Fontainebleau, ibid.*, n. 14 (1980); R.H. LYTLE, *Intellectual Access to Archives: Provenance and Content Indexing Methods of Subject Retrieval*, in « The American Archivist », 43, n. 1 (Winter 1980) pp. 63–75; ID., *Intellectual Access to Archives: II. Report of an Experiment Comparing Provenance and Content Indexing Methodsof Subject Retrieval, ibid.*, 43, n. 2 (Spring 1980) pp. 191–207; A. ARAD, *Enregistrement et Indexage automatique: les archives de l'Etat d'Israel*, in « Revue de l'UNESCO pour la science de l'information, la bibliothéconomie et l'archivistique », II, n. 2 (avril–juin 1980) pp. 127–137; A. ARAD, M.E. OLSEN, *An Introduction toArchival Automation*, Committee on automation, ICA, 1981; C.D. CHALMERS, *Computer indexing in the Public Record Office*, in « Journal of the Society of Archivist », 6, n. 7 (Avril 1981) pp. 399–413; C.D. CHALMERS, J.B. POST, *A flexible system for the cumulative general index, ibid.*, 6, n. 8 (October 1981) pp. 482–492; P. BUONORA, *L'informatica negli archivi francesi*, in « Rassegna degli Archivi di Stato », XLIII (1983)pp. 152–267; ID., *Informatica e archivi: esperienze internazionali*, in *Le fonti documentarie (Atti del VII Corso di archivistica, Loreto 24–28 ott. 1983)* Ancona 1984.

RETI DI COMUNICAZIONE E BASI DI DATI: T. LAZZARI, *Telematica e basi di dati nei servizi bibliotecari*, Roma 1982; UNESCO, *L'informatique, facteur vital de developpement*, Paris 1982; D. BEARMAN, *Toward National Information Systems for Archives and Manuscript Repositories*, in « The American Archivist », 45, n. 1 (Winter 1982) pp. 53–56; R.K. KESNER, *Microcomputer Application in Archives: toward an international information retrieval network*, in « ADPA », IV, nn. 1–2, pp. 58–65.