MACCHINE PER LEGGERE / Firenze 16.11.93

Tito ORLANDI

Integrazione dei sistemi di lettura intelligente e banche dati nel progetto Medioevo-Europa.

1. La «lettura ottica»

Fra i pacchetti applicativi che da alcuni anni hanno suscitato grandi speranze fra i cultori o fautori delle applicazioni dell'informatica alle discipline umanistiche, stanno in particolare rilievo i cosiddetti pacchetti di Optical Character Recognition (OCR). Queste speranze hanno addirittura generato una specie di mitologia, che riguarda l'efficacia e l'utilità dei sistemi OCR(/ICR). Credo che valga la pena di chiarire alcune aspetti sia tecnici sia teorici che li riguardano, per poi passare a descrivere la loro utilizzazione nell'ambito del progetto Medioevo Europa.[1]

Si tratta, come è noto, di sistemi in grado di esplorare un'immagine testuale digitalizzata, riconoscerne la disposizione in linee di caratteri alfabetici e segni diacritici, e produrre un file in cui il medesimo testo è rappresentato in base alla convenzione chiamata: ASCII.

Abbiamo parlato in un primo tempo di pacchetti applicativi, poi piú genericamente di sistemi; in effetti è opportuno chiarire di volta in volta la reale composizione di un O(I)CR, anche per quanto concerne il rapporto fra hardware e software. Col ter-

---

1. Cf. David D. BUSH, *The Complete Scanner Handbook for Desktop Puglishing. PC Edition*, Homewood 1990. David J. BIRNBAUM, *Optical Character Recognition & Non-Latin Alphabets*, «Bits & Bytes Review» 2 (1991) 6-7 p. 22-28. René van HORIK, *Optical Character Recognition and Historical Documents: Some Programs Reviewed*, «History and Computing», 4 (1992) 211-220. Inoltre gli atti del Congresso *** di prossima pubblicazione.

mine «sistema O/ICR» (o sigle analoghe) si designa comunemente tutto ciò che occorre per attuare un procedimento automatico all'origine del quale sta una pagina stampata, e alla fine del quale sta una serie di byte residenti su memoria magnetica.

Per ottenere questo procedimento, oltre ad un computer (diciamo cosí) normale, che funzioni come base di tutto il procedimento, occorre una macchina in grado di digitalizzare le immagini, ed un pacchetto applicativo in grado di interpretare le immagini, riconoscere i caratteri, e codificarli.

Un sistema come il noto Kurzweil,[2] p.es., mantiene strettamente uniti i due aspetti. In sostanza fornisce una macchina dotata insieme di hardware (scanner e processore RISC) e di software, che trasmette al computer «principale» direttamente il testo codificato. Altri sistemi, come l'altrettanto noto OmniPage,[3] sono puri pacchetti applicativi (software), che gestiscono immagini digitalizzate secondo sistemi di memorizzazione standard. Queste premesse tecniche erano secondo me indispensabili, per invitare a tener distinta l'immagine testuale dal testo e dalla codifica del testo.[4]

In un primo tempo il procedimento di O/ICR avveniva in modo relativamente rigido. Il pacchetto funzionava esclusivamente in base al riconoscimento della forma dei caratteri di un certo numero di polizze (font) ritenute di maggior uso; ovvero poteva essere istruito da un operatore in modo da riconoscere una polizza non prevista. Gli inconvenienti principali di questo sistema primitivo erano due, possiamo dire uno in entrata ed uno in uscita. L'inconveniente in entrata era costituito dal fatto che l'immagine digitalizzata che è all'origine del pro-

2. Attualmente Kurzweil K5200, vers. 6.1, della Xerox Imaging Systems, Cambridge MA (USA).

3. Omnipage Professional, vers. 2.0, della Caere Corporation, Los Gatos CA (USA).

4. E' altrettanto importante la distinzione fra pacchetti basati sulla *pattern recognition* e pacchetti basati sulla *feature extraction*, negli algoritmi di riconoscimento dei caratteri; ma non è rilevante per il presente argomento.

cedimento doveva essere perfetta. Ove occorressero imperfezioni nella stampa dell'originale, ovvero macchie o altro sulla carta, il sistema non era in grado di ricostruire la lettera in questione, ovvero la scambiava con un'altra lettera. L'inconveniente in uscita era costituito dal fatto che, quando pure il sistema fosse in grado (come spesso era) di riconoscere certe particolarità della polizza (corsivo, grassetto, maiuscoletto, etc.), tali particolarità non venivano registrate nel file prodotto, che conteneva dunque i caratteri alfabetici corrispondenti, senza alcuna discriminazione.

A questi e ad altri inconvenienti si è cercato di ovviare con pacchetti piú sofisticati che hanno preso il nome di Intelligent Character Recognition (ICR), in cui la qualifica di intelligente deriva da un certo parallelismo con i sistemi cosiddetti di Intelligenza Artificiale, cioè quei sistemi che basano i loro procedimenti anche su un patrimonio di conoscenze, che è in grado di crescere in modo automatico durante l'attuazione dei procedimenti stessi. In particolare, gli ICR sono stati dotati di dizionari di riscontro, in base ai quali riconoscere sequenze di lettere come lecite, e ricostruire al loro interno le eventuali singole lettere non riconosciute. In conseguenza poi di questa "esperienza" essi possono imparare le idiosincrasie dei font con cui sono confrontati.

Inoltre gli ICR sono stati messi in grado di analizzare meglio l'immagine digitalizzata, in modo da riconoscere alcune strutture dell'impaginazione dei testi, escludendo p.es. le figure, ovvero trattando a parte delle finestre estranee al testo principale; e sono stati forniti della capacità di inserire nel file prodotto elementi di codifica che avvertano dell'esistenza delle particolarità sopra menzionate.

Come si vede, di propriamente intelligente in questi sistemi non c'è niente; c'è tuttavia una crescente versatilità ed efficacia, che giustifica le speranze degli studiosi, che vorrebbero avere a disposizione una sempre maggiore quantità di testi in Machine

Readable Form, per poterli sottoporre a procedimenti di analisi linguistica o contenutistica di tipo automatico. È evidente come la prospettiva di compiere osservazioni, analisi, e sperimentazioni sui testi mediante procedimenti informatici sia entusiasmante. Ma quei procedimenti presuppongono che i relativi testi siano accessibili su supporto magnetico, cosa che equivale a riscriverli completamente, se non si trova un sistema che sostituisca il lavoro umano. Affidare questo compito ad una macchina sembra risolvere il grande problema preliminare.

In realtà ci si accorge presto che per utilizzare seriamente e scientificamente i testi su supporto magnetico occorre ben altro. Ci si accorge in sostanza quanto poco intelligente, meglio ancora, intelligibile, sia il prodotto offerto dagli ICR o anche da qualunque procedimento manuale che non vada oltre quanto offre un ICR; e quale sia la distanza fra la capacità di riconoscere le sequenze di caratteri che compongono un testo e la capacità di riprodurre su supporto magnetico il testo stesso, del quale i caratteri (glifi) sono soltanto uno degli elementi costitutivi.

2. La rappresentazione del testo

Occorre distinguere due problemi diversi che sorgono una volta che sia concluso il primo passaggio di OCR. Da un lato quello della banale correzione del testo, quando l'OCR ha sbagliato nel riconoscere un carattere; non è questo un problema su cui valga la pena di soffermare l'attenzione in questa sede.

È invece molto diverso il problema della validità del testo MRF rispetto a quello stampato. Quest'ultimo infatti è accessibile direttamente per mezzo dei sensi dalla persona umana; nel secondo caso il risultato è accessibile soltanto per mezzo di macchine. Per questo motivo il testo scritto presuppone sempre un lettore competente che sia in grado di interpretare non solo i segni in se stessi, ma molti altri fenomeni che contribuiscono a caricare i segni di ulteriori significati. Si tratta: della posizione

dei segni nell'impaginazione; di diversità nella grafia dei segni (spaziature, corsivi, legature, grassetti etc.), e altre particolarità che attengono alle varie sezioni di un testo (titoli, divisione di paragrafi etc.).

Il lavoro di codifica per usi automatizzati tende a limitarsi alla codifica di righe di testo intese come sequenze di lettere alfabetiche. Può essere conveniente invece distinguere due metodi di codifica: quello di *constatazione* e quello di *interpretazione*.[5]

La codifica di constatazione è rivolta al puro aspetto materiale del segno da codificare. La codifica di interpretazione, oltre a prendere in considerazione tale aspetto (caratteri alfabetici, etc., ovverossia i glifi)[6] che costituisce la base del testo, inserisce anche delle annotazioni (*tag*) che rendono possibile alla macchina di recepire quelle informazioni che al lettore vengono date da elementi diversi dai caratteri stessi. Tale interpretazione non si ferma all'aspetto individuale del singolo glifo, ma prende in considerazione la posizione, eventuali elementi come ornamenti etc., fino ad arrivare alla caratterizzazione delle sequenze

5. Sui procedimenti che chiamiamo qui di codifca, che talora vengono chiamati anche di modellizzazione (modeling), cf.: GIGLIOZZI, Giuseppe (ed.), *Studi di codifica e trattamento automatico di testi*, (Informatica e discipline umanistiche, 1), Roma, 1987, 234 p.: Bulzoni Editore (in particolare: ADAMO, Giovanni, *La codifica come rappresentazione. Trasmissione e trattamento dell'informazione nell'elaborazione automatica di dati in ambito umanistico,*, pp. 39-63; GIGLIOZZI, Giuseppe, *Codice, testo e interpretazione*, pp. 65-84; MORDENTI, Raul, *Appunti per una semiotica della trascrizione nella procedura ecdotica computazionale*, pp. 85-124.) ORLANDI, Tito, *Problemi di codifica e trattamento informatico in campo filologico*, in: [5273] G. Savoca (ed.), Lessicografia, filologia e critica, Firenze 1986, pp. 69-82.

6. Ritengo opportuna l'introduzione di questo termine, equivalente italiano dell'inglese *glyph*, recentemente riscoperto in ambiente informatico, per sottolineare il puro aspetto grafico del segno stampato (e per estensione anche manoscritto). Il termine «carattere» si applica infatti non solo all'aspetto grafico, ma anche a quello fisico del pezzetto di piombo che appunto imprimeva il glifo.

di caratteri del tipo «paragrafo», «titolo», etc.

Questo è quanto è reso possibile dall'uso di un sistema di mark-up, ovvero metalinguaggio, tipo lo SGML (Standard General Mark-up Language), che offre la possibilità di indicare, oltre le singole lettere, interpretazioni come: titolo, citazione, evidenziazione, etc.[7] È evidente che anche in questo modo parte dell'informazione, che un lettore umano ricava implicitamente dal testo, viene perduta; tuttavia è questo l'unico mezzo che permette di giungere ad analisi automatiche che almeno conservino alcune di quelle alterazioni di significato che altrimenti vengono completamente perdute.

Per questo motivo la codifica mediante SGML è stata scelta come base per la costituzione della banca dati, ad un tempo testuale e fattuale, di Medioevo Europa (cf. sotto). E per questo motivo ci sembra non inutile soffermarci brevemente sulle caratteristiche piú importanti dello SGML. Esso non costituisce propriamente un sistema di codifica, ma la *struttura sintattica di*

7. BARNARD, David T. - HAYTER, Ron - KARABABA, Maria - LOGAN, George M. - McFADDEN, John, *SGML-Based Markup for Literary Texts: Two Problems and Some Solutions*, «Computers and the Humanities», 22 (1988) 4, 265-276. BRYAN, Martin, *SGML. An Author's Guide to the Standard Generalized Markup Language*, Wokingham (UK), 1988, xvii-364 p.: Addison-Wesley. ROBINSON, Peter M.W., *The Transcription of Primary Textual Sources using SGML*, (Office for Humanities Communication Publications), Oxford, 1993: Office for Humanities Communication. GOLDFARB, Charles F. - RUBINSKY, Yuri, *The SGML Handbook*, Oxford, 1992 (ristampa), xxiv-663 p.: Clarendon Press. VAN HERWIJNEN, Eric, *Practical SGML*, Dordrecht-Boston-London, 1990, xviii-307 p.: Kluwer Academic Publishers. SMITH, Joan M., *The Standard Generalized Markup Language (SGML) for Humanities Publishing*, «Literary and Linguistic Computing. Journal of the Association for Literary and Linguistic Computing», 2 (1987) 3, 171-175. COVER, Robin C. - DUNCAN, Nicholas - BARNARD, David T., *The Progress of SGML (Standard Generalized Markup Language): Extracts from a Comprehensive Bibliography*, «Literary and Linguistic Computing. Journal of the Association for Literary and Linguistic Computing», 6 (1991) 3, 197-209.

un linguaggio. Tale linguaggio è stato studiato particolarmente per poter descrivere in modo completo ed efficace un testo inteso come un insieme di elementi alfabetici, grafici, e semantici.

Tuttavia l'origine e lo sviluppo dello SGML non sono cosí chiari e univoci, da non generare qualche incertezza fra i possibili utilizzatori. Lo SGML nasce come strumento di unificazione di ordini di impaginazione di un testo (dunque descrizione del testo da parte del suo autore, non di un suo lettore o interprete), che superasse l'inconveniente dell'esistenza di molti linguaggi «diretti» di impaginazione, ciascuno con una sua sintassi, direttamente legati ai differenti programmi di impaginazione automatica (le famiglie dei troff, TeX, etc.). Esso era rivolto soprattutto ad istituzioni di tipo pragmatico, non scientifico (enti, uffici, editori), e veniva incontro a due esigenze (oltre a quella sopra indicata): la conservazione di informazioni contenute non nel testo in sé, ma nella sua disposizione (p.es. il fatto che il nome che si trova in alto a destra di una lettera sia quello del destinatario, e in basso quello dello scrivente); e la trasmissione di documenti da un gruppo di persone ad un altro, con l'indicazione inequivoca di dove trovare all'interno del testo i diversi tipi di informazione eventualmente cercata.

L'applicazione di un tale linguaggio e del conseguente sistema di segni (tag) nel campo delle discipline storiche comporta un rovesciamento simmetrico del punto di vista del codificatore. Invece di essere il produttore del testo che, insieme col contenuto, indica esplicitamente la struttura, è il lettore o fruitore del testo che ne deve compiere prima di tutto un'analisi, per ritrovare le unità strutturali che vengono quindi esplicitate, marcate, con le opportune segnalazioni. In particolare si deve tener presente che il lavoro di codifica, in questo caso, presenta insieme elementi di descrizione (della pagina etc.) e di interpretazione del testo, ambedue operazioni ora piú ora meno soggettive, che occorre compiere con piena consapevolezza, dichiarando accuratamente ed esplicitamente il metodo e gli scopi per cui

si è operato.

Lo SGML offre anche un'altra caratteristica importante. Per mezzo degli opportuni segni (tag) inseriti nel testo, è possibile delimitare sezioni di testo aventi particolari attributi, che per mezzo di un programma di analisi (parser) vengono poi riconosciute come elementi di informazione (record) di una banca dati diciamo cosí virtuale. La banca dati può diventare reale quando lo stesso parser provveda ad estrarre i dati dall'archivio testuale per formare un archivio dati strutturato.

Sul compromesso fra descrizione pura ed interpretazione dei testi d'interesse storico è basata la concezione teorica di Manfred Thaller, il noto pioniere dell'informatica applicata alle fonti storiche, e sviluppatore di un interessante sistema di analisi storica (Kleio)[8] di cui desideriamo citare un brano significativo, avvicinandoci cosí all'ambiente storico-letterario nel quale si colloca il progetto di Medioevo-Europa.

> «Processing of Historical sources is different from the processing of present day data by a number of reasons: on the most abstract level, this is the case because Historians, when they start their research, do not really «know» with absolute certainty, what their texts mean. Therefore, historical data should be administered in a way, which closely resem-

8. THALLER, Manfred, *Beyond Collecting: The Design and Implementation of CLIO, a DBMS for the Social-Historical Sciences,* in: R.F. Allen (ed.), The International Conference on Data Bases in the Humanities and Social Sciences 1983, Osprey (Fl) 1985, pp. 328-334. THALLER, Manfred, *Automation on Parnassus. CLIO - A Databank Oriented System for Historians,* «Historical Social Research. Historische Sozialforschung», (1980) 15, 40-65. THALLER, Manfred, *The Historical Workstation Project,* in: J. Smets (ed.), Histoire et Informatique, Montpellier 1992, pp. 251-260. THALLER, Manfred, *The Historical Workstation Project,* «Computers and the Humanities», 25 (1991) 2-3, 149-162. THALLER, Manfred, κλειω, *Ein fachspezifisches Datenbanksystem für die Historischen Wissenschaften, Version 2.1.1,* Göttingen, 1988: Max-Planck-Institut für Geschichte.

> bles the basic principles of a printed edition as used in the historical disciplines, particularly in the schools of medieval studies. The source itself, we said, could not possibly be wrong: if a name was spelled differently at two occassions this could have been an oversight of the scribe; it could be just as well, however, that this «scribal error» was just the only trace left of the existence of two individuals, separated by a minor difference in the spelling of their names. This being so, we claimed, genuinely «historical» data processing must keep the source as closely to the uncorrected original as possible. Now, six different orthographical representations of one word quite obviously tend to frustrate computer supported analysis; as does the use of currencies of unknown interpretation, complex references to calendar dates and the like. All these problems, however, occur also in printed editions: where in the best ones, from a historian's point of view, therefore two types of information are carefully kept: the literal transcription of a text and a complex environment of apparatuses and appendices, which contain the interpretations the editor has for the text he presents to the historian using the edition.
> This structure, we claimed, would have to be repeated in historical data processing.»[9]

3. Il progetto Medioevo Europa

Il Progetto strategico del Consiglio Nazionale delle Ricerche «Medioevo Europa» si configura come una grande iniziativa culturale a livello europeo relativa all'informazione scientifica sul Medioevo. Esso si propone di progettare, costituire, e quindi gestire una banca dati relativa alle fonti della storia medievale e

9. THALLER, Manfred, *Historical Information Science: Is There Such a Thing? New Comments on an Old Idea,* in: T. Orlandi (ed.), Discipline umanistiche e informatica, Roma 1993, pp. 52-53.

agli autori medievali, che abbiano scritto sia in latino che nelle altre lingue in uso durante tale periodo./*F Nella sua fase di piena realizzazione, la banca dati sarà consultabile per via telematica dagli studiosi.

L'iniziativa è basata sulla cooperazione di tre centri di ricerca extrauniversitari italiani che, sorti in momenti diversi e con finalità anche in parte diverse, godono di una consolidata rinomanza internazionale in questo campo di studi: il Centro Italiano di Studi sull'Alto Medioevo (CISAM) di Spoleto, presieduto dal Prof. Ovidio Capitani; l'Istituto Storico Italiano per il Medio Evo (ISIME) di Roma, presieduto dal Prof. Girolamo Arnaldi; e il Dipartimento di Studi sul Medioevo e il Rinascimento (DISMER), con cui coopera la Società Internazionale per lo Studio del Medioevo Latino (SISMEL) di Firenze, presieduta dal Prof. Claudio Leonardi.

Il CISAM, fondato nel 1952, è noto soprattutto per le sue «settimane di studio», che sono arrivate nel 1992 alla quarantesima edizione e costituiscono un importante appuntamento annuale dei medievisti di tutto il mondo; a scadenza non fissa il CISAM organizza inoltre congressi in varie città d'Italia, dedicati allo studio dei diversi «medievi» locali. Fino ad oggi ne sono stati tenuti dodici, i cui atti sono stati pubblicati in sedici volumi. Il CISAM è anche editore di *Studi medievali*, il maggior periodico specializzato in questo settore.

L'ISIME, fondato nel 1883 con il compito precipuo di curare l'edizione delle fonti medievali italiane, dal 1956 ha dato vita al *Repertorium Fontium Historiae Medii Aevi*, dove vengono registrate in ordine alfabetico tutte le fonti storiche edite, riguardanti l'Europa medievale, fornendo notizie essenziali circa opere e autori e le relative indicazioni bibliografiche.

La SISMEL, attiva dal 1978, cura la redazione e la stampa di *Medioevo Latino*, Bollettino bibliografico della cultura europea

9. La presentazione generale del progetto rispecchia la relazione presentata al CNR, nella forma redatta dal dr. *** Orioli.

dal secolo VI al XIII, dove con periodicità annuale si dà notizia degli studi circa gli autori medievali che scrivono in latino e le loro opere edite e inedite. Ha inoltre intrapreso la compilazione della lista degli autori latini medievali, la *Clavis Auctorum Medii Aevi*.

Il Progetto prevede la costituzione di un gruppo di archivi su supporto elettronico, e quindi il loro aggiornamento e la loro diffusione telematica. Tali archivi conterranno informazioni relative alle fonti storiche e, in genere, alla produzione scientifico-letteraria del Medioevo europeo, a partire dall'anno 400 circa all'anno 1500, con un'ulteriore dilatazione fino al secolo XVII per i paesi dell'Europa centro-orientale. Inoltre prenderà in considerazione l'Alto Medioevo nei suoi diversi aspetti. La banca-dati informatica sarà localizzata presso la sede del CISAM di Spoleto, e nel contempo saranno automatizzate le attività di ricerca dell'ISIME, della SISMEL (che già opera con mezzi informatici), e dello stesso CISAM, nonché dei centri periferici di documentazione che operano in collaborazione con le suddette istituzioni.

Il Progetto prevede la costituzione di tre poli informatici presso la sede dei tre Enti (Spoleto, Roma e Firenze). Ciascun polo sarà dotato di opportune strumentazioni capaci di supportare le attività locali, peculiari di ciascun ente, e di collegarsi in via telematica con la banca-dati centralizzata allo scopo di inviare aggiornamenti, reperire informazioni e richiedere servizi sotto forma di ricerca e stampa di informazioni selezionate.

La banca-dati verrà pertanto a costituirsi attraverso l'acquisizione delle informazioni contenute nei volumi già editi o in corso di stampa delle «Settimane di studio» e dei congressi promossi dal CISAM, del Repertorium e di Medioeuo Latino; e servirà a:

- raccogliere man mano le informazioni selezionate dalle tre redazioni mettendole reciprocamente a loro disposizione;
- consentire l'aggiornamento permanente dei volumi già apparsi

del Repertorium anche in vista di eventuali riedizioni o di edizioni dei soli aggiornamenti;
- gestire i dati concernenti la compilazione della Clavis Auctorum Medii Aevi;

A sua volta la struttura informatica di cui verranno ad essere dotati i tre centri di ricerca è stata studiata al fine di sfruttare anche possibilità avanzate di edizione di testi, con la disponibilità di numerosi font di caratteri, per poter eliminare i ritardi e ridurre i costi derivanti dagli attuali numerosi passaggi tra centri di edizione e tipografie.

Si tratta dunque di un piano complesso,[10] che prevede l'integrazione di archivi e di procedure, secondo uno schema che si può riassumere nei seguenti termini:

Archivi: Coesistenza di archivi testuali (i testi delle «Settimane» ed altri Atti del CISAM, e lo stesso *Repertorium* dell'ISIME in forma testuale) con codifica SGML (cf. sopra), e di archivi di dati sotto di file organizzati secondo un modello relazionale (cf. sotto).

Procedure per l'acquisizione delle informazioni, utilizzando inizialmente le tecniche di O/ICR (cf. sopra), per ottenere il recupero dati da pubblicazioni bibliografiche, da repertori, e da saggi e monografie.

Procedure per la diffusione delle informazioni, utilizzando impaginatori automatici, eventualmente basati su parser SGML, cioè che siano guidati nell'impaginazione dalla codifica inserita nei file. In questo modo si potrà ottenere in via semi-automatica, dalle informazioni contenute nella banca dati, la stampa di bibliografie, repertori, e saggi.

---

10. La consulenza informatica al progetto è fornita dal Centro Interdipartimentale per l'Automazione nelle Discipline Umanistiche (CISADU) dell'Università di Roma La Sapienza, diretto da chi scrive.

Procedure per la diffusione delle informazioni su rete telematica, costituendo a Spoleto presso il CISAM il nodo di una rete internazionale.[11]

Le idee che guidano la realizzazione del piano sono:

- mantenimento del diverso carattere degli archivi, rispettando la concezione delle fonti originarie a stampa
- utilizzazione di procedimenti O/ICR per facilitare un primo passaggio nella memorizzazione dei dati
- utilizzazione del linguaggio di codifica SGML per la classificazione delle informazioni all'interno del testo, ed il suo successivo recupero
- utilizzazione di parser per il collegamento fra archivi testuali e archivi di dati
- adozione del modello relazionale per la struttura logica della banca-dati

Riteniamo di un certo interesse riassumere qui quanto è stato accettato della teoria relativa al modello relazionale nella formulazione adottata per Medioevo Europa e descriverlo in linguaggio di normale comprensione in ambiente umanistico.[12]

Secondo il mio punto di vista, la teoria relazionale, sebbene concepita per migliorare la gestione delle banche dati, non consiste tanto in una serie di accorgimenti tecnici, quanto in un modo teoricamente fondato di considerare la struttura dei dati, e prima ancora la struttura della realtà di cui i dati sono l'espressione o rappresentazione. Per ottenere questo risultato essa distingue i dati in: entità, che sono le sostanze primarie che vengono prese in considerazione; attributi, che sono analoghi a ciò che Aristotele chiamava «accidenti», cioè caratteristiche che es-

11. Attualmente non è ancora stato deciso a quale rete appoggiarsi, in quanto il problema concreto si porrà tra piú di un anno, ed il settore è in rapida crescita e diversificazione.

istono solo in quanto accompagnano le entità; e relazioni, che esprimono il procedimento aristotelico di «predicazione», mettendo in relazione le entità e di conseguenza i loro attributi.

Normalmente, quando si agisce nell'ambito delle banche dati, si tende a considerare entità i fenomeni del mondo reale: persone, oggetti, strutture, etc. Si parla quindi di banche o basi di dati in archeologia, dove si classificano reperti, o in storia dell'arte, per i manufatti artistici, o in storia, per gli eventi storici; e solo per estensione diciamo cosí metaforica si parla di banche o basi di dati testuali, intendendo raccolte di testi in Machine Readable Form, variamente indicizzati.

È possibile invece considerare come entità di una banca dati relazionale le singole parole di un testo, e come attributi (p.es.) la loro collocazione in un relazione alle altre parole dello stesso testo, oltre che le loro caratteristiche morfologiche, grammaticali e semantiche. È questo il principio in base al quale alcuni pacchetti di concordanze sono stati fatti per mezzo di un DBMS come il dBase. Ma soprattutto è questo il principio per cui la concezione relazionale (unita all'uso di un metalinguaggio come SGML) permette l'integrazione di una banca dati tradizionale e una banca dati testuale.

Infatti gli archivi di indici, che rimandano alle unità e sezioni di un testo, segnalate per mezzo del SGML, si trasformano

---

12. Sul modello relazionale cf. DATE, C. J., *An Introduction to Database Systems, Reading, 1983, Addison-Wesley. TIBERIO, Paolo (ed.), Basi di Dati. Stato dell'arte e prospettive, Collana AICA di Informatica, Milano, 1985, Masson Italia. ATZENI, Paolo - BATINI, Carlo - DE ANTONELLIS, Valeria, La teoria relazionale dei dati, Torino, 1985, Boringhieri. KORTH, Henry F. - SILBERSCHATZ, Abraham, Database System Concepts, New York, 1986, McGraw-Hill. SCHIAVETTI, Emilio, Data Base. Introduzione ai sistemi relazionali, Milano, 1989, Gruppo Editoriale Jackson. PAREDAENS, Jan - DE BRA, Paul - GYSSENS, Marc - GUCHT, Dirk van, The Structure of the Relational Database Model, EATCS Monographs on Theoretical Computer Science, Berlin - Heidelberg, 1989, 231 p. - 53 fig., Springer-Verlag. ELMASRI, Ramez - NAVATHE, Shamkant B., Fundamentals of Database Systems, Redwood City (CA), 1989, Benjamin-Cummings.*

facilmente in tabelle di relazione fra le unità e sezioni del testo e gli elementi di archivi di bibliografia o di repertorio. Con questo criterio viene dunque affrontato il problema dell'unificazione delle strutture degli archivi (in partenza cartacei) dei tre Istituti che partecipano al progetto, e vengono individuate le "tabelle" e le relazioni fra le tabelle.

Come abbiamo accennato sopra, il lavoro va diviso fra due problemi interconnessi ma che vanno affrontati separatamente. Da un lato occorre fare in modo che l'attività dei tre Istituti non subisca ritardi a causa dell'informatizzazione, ma anzi ne sia agevolata; dall'altro occorre fare in modo che le procedure organizzate per il lavoro «nuovo» (raccolta dei dati e loro inserzione negli archivi; stampa dei volumi) siano compatibili con quelle organizzate per ottenere la memorizzazione dei dati contenuti nei volumi stampati in precedenza.

Si è partiti da uno studio approfondito dei volumi editi al fine di definirne la struttura, individuando le tipologie, i criteri di ordinamento delle informazioni, le regole tipografiche. Questo studio permette di ottenere un file che contiene ad un tempo il testo del volume e la sua descrizione strutturale e logica, avvalendosi delle codifiche proprie del sistema SGML. Esso renderà possibile da un lato l'impaginazione per la stampa, dall'altro il riversamento dei dati in un *data base management system*.

Un problema diverso è rappresentato dal recupero del pregresso. L'acquisizione dei volumi editi tramite scanner costituirà sicuramente un'operazione difficile a causa della complessità tipografica dei volumi. Sarà inoltre necessario un intervento di notevole entità per inserire i *tag* che identifichino i campi di informazione per ottenere il recupero dei dati. Inoltre i dati stessi dovranno essere completati là dove si è scelta per la stampa una forma semplificata.

In conclusione: con Medioevo Europa si cerca di compiere un'operazione tecnologica che sia strettamente connessa con

una riflessione sulle conseguenze metodologiche per le discipline interessate, e sui problemi generali di codifica e recupero dell'informazione che toccano tutte le applicazioni dell'informatica in ambito umanistico.

---------- Codifica generale:

1379@CHASE, Michael D., *Data Transfers Between Incompatible Operating Systems*, «Computers and the Humanities», 22 (1988) 2, 153-156. 1422@GANESHSUNDARAM, P.C., *Processing of Japanese Kanji on a Microcomputer*, «Computers and the Humanities», 21 (1987) 3, 157-167. 1453@DUNCAN, S. - MUKAII, T. - KUNO, S., *A Computer Graphics System for Non-Alphabetic Orthographies*, «Computer Studies in the Humanities and Verbal Behavior», 2 (1969) 3, 113-132. 1676@HUNTER, Lawrence W., *A Data Representation Code for Text Processing Systems*, «International Journal of Computer and Information Science», 1 (1972) 1, 29-42. 1812@GIGLIOZZI, Giuseppe (ed.), *Studi di codifica e trattamento automatico di testi*, (Informatica e discipline umanistiche, 1), Roma, 1987, 234 p.: Bulzoni Editore. 1813@ORLANDI, Tito, *Informatica umanistica. Riflessioni storiche e metodologiche, con due esempi*, in: [1812] G. Gigliozzi (ed.), Studi di codifica e trattamento automatico di testi, Roma 1987, pp. 1-38. 1814@ADAMO, Giovanni, *La codifica come rappresentazione. Trasmissione e trattamento dell'informazione nell'elaborazione automatica di dati in ambito umanistico*, in: [1812] G. Gigliozzi (ed.), Studi di codifica e trattamento automatico di testi, Roma 1987, pp. 39-63. 1815@GIGLIOZZI, Giuseppe, *Codice, testo e interpretazione*, in: [1812] G. Gigliozzi (ed.), Studi di codifica e trattamento automatico di testi, Roma 1987, pp. 65-84. 1816@MORDENTI, Raul, *Appunti per una semiotica della trascrizione nella procedura ec-*

dotica computazionale, in: [1812] G. Gigliozzi (ed.), Studi di codifica e trattamento automatico di testi, Roma 1987, pp. 85-124. 1817@BUIARELLI, Paolo - TIRADRITTI, Francesco, *Codifica della scrittura geroglifica finalizzata all'analisi testuale,* in: [1812] G. Gigliozzi (ed.), Studi di codifica e trattamento automatico di testi, Roma 1987, pp. 125-144. 1818@GIGLIOZZI, Giuseppe - GIULIANI, Sandra - SENSINI, Paolo, *SEB - Sistema Esperto per l'analisi di Brani. Per un'analisi automatica di fiabe,* in: [1812] G. Gigliozzi (ed.), Studi di codifica e trattamento automatico di testi, Roma 1987, pp. 145-217. 1817@BUIARELLI, Paolo - TIRADRITTI, Francesco, *Codifica della scrittura geroglifica finalizzata all'analisi testuale,* in: [1812] G. Gigliozzi (ed.), Studi di codifica e trattamento automatico di testi, Roma 1987, pp. 125-144. 1935@KRISHNAMOORTY, S.G. - ISAAC, J.R. - BHAVSAR, V.C., *A Microprocessor-Based Multilingual Terminal for a Computerized Information Handling System,* «Computers and the Humanities», 14 (1980) 2, 91-104. 1988@SPERBERG-McQUEEN, C. Michael (ed.) - BURNARD, Lou (ed.), *Guidelines for the Encoding and Interchange of Machine-Readable Texts. Draft: Version 1.1, October 1990 Document Number: TEI P1,* Chicago-Oxford, 1990 (2nd printing, June 1991), xx-290 p.: Text Encoding Initiative. 3053@ORLANDI, Tito, *Problemi di codifica e trattamento informatico in campo filologico,* in: [5273] G. Savoca (ed.), Lessicografia, filologia e critica, Firenze 1986, pp. 69-82. 5105@BRODIE, Michael L. (ed.) - MYLOPOULOS, John (ed.) - SCHMIDT, Joachim W. (ed.), *On Conceptual Modelling. Perspectives from Artificial Intelligence, Databases, and Programming Languages,* (Topics in Information Systems), New York - Heidelberg - Tokyo - Berlin, 1984, xi-510 p.: Springer-Verlag. 5106@CROSS, M. (ed.) et Al., *Modelling and Simulation in Practice,* London, 1979: Pentech-Press. 5137@JARDINE, C.J. - MacFARLANE, Alan D.J., *Computer Input of Historical Records for Multi-Source Record*

Linkage, M.W. Flinn (ed.), Proceedings of the Seventh International Economic History Congress, Edinburgh 1979, pp. 71-78. 5395@MOOERS, Calvin N., *Codes and Coding*, Allen Kent - Harold Lancour (eds.), Encyclopedia of Library and Information Science, New York, Marcel Dekker, 1971, vol. 5, pp. 251-260. 5396@SMITH, John B., *Encoding Literary Texts: Some Considerations*, «Association for Literary and Linguistic Computing Bulletin», 4 (1976) 3, 190-198. 5398@ZARRI, Gian Piero, *L'enregistrement informatique de documents non-littéraires rédigés dans une langue ancienne: observations et suggestions*, «Association for Literary and Linguistic Computing Bulletin», 4 (1976) 2, 115-124. 5444@MACKENZIE, Charles E., *Coded Character Sets. History and Development*, (The Systems Programming Series), Reading (Mass.), 1980, xxi-513 p.: Addison-Wesley. 5487@BURNARD, Lou, *La standardisation des textes électroniques*, in: [5486] L. Fossier (ed.), Le médiéviste et l'ordinateur, Paris 1989, pp. 87-88. 5520@BOOT, Martin, *Linguistic Data Structure, Reducing Encoding by Hand, and Programming Languages*, in: [5506] A. Jones - R.F. Churchhouse (eds.), The Computer in Literary and Linguistic Studies, Cardiff 1976, pp. 255-270. 5554@SMITH, Joan M., *Language Representation, the Standard Way*, in: [5537] J. Hamesse - A. Zampolli (eds.), Computers in Literary and Linguistic Computing, Paris-Genève 1985, pp. 355-369. 5655@TOLLENAERE, Félicien de, *Encoding Techniques in Dutch Historical Lexicography*, «Computers and the Humanities», 6 (1972) 3, 147-152. 5781@SU, K.L., *The Creation of a Set of Alphabets for the Chinese Language*, in: [5780] J.L. Mitchell (ed.), Computers in the Humanities, Edinburgh 1974, pp. 27-38. 5784@JORY, Edward John, *New Approaches to Epigraphic Problems in Roman History*, in: [5780] J.L. Mitchell (ed.), Computers in the Humanities, Edinburgh 1974, pp. 184-190. 5802@GRAY, Max - LONDON, Keith R., *Comment normaliser les documents informatiques. Traduit de l'américain*

par G. Duon et C. Ellast, (L'informathèque, 17), Paris, 1973, xii-163 p.: Entreprise Moderne d'Édition. 5806@STAHL, Gérold, *Moins de traitement sémantique et plus de preédition en traduction assistée par ordinateur*, in: [5469] C. Charpentier - J. David (eds.), La recherche française par ordinateur en langue et littérature, Genève-Paris 1985, pp. 133-137. 5827@KAY, Martin, *Standards for Encoding Data in a Natural Language*, «Computers and the Humanities», 1 (1967) 5, 170-177. 7066@VAN LINT, Jacobus Hendricus, *Introduction to Coding Theory*, New York, 1982, ix-171 p.: Springer-Verlag. 7301@GOPNIK, Irwin (ed.) - GOPNIK, Myrna (ed.), *From Models to Modules. Studies in Cognitive Science from the McGill Workshops*, (Theoretical Issues in Cognitive Science), Norwood (NJ), 1986, viii-295 p.: Ablex. 7302@HAMMING, Richard W., *Coding and Information Theory*, Englewood Cliffs (NJ), 1986 (seconda edizione): Prentice-Hall. 7341@TOGNETTI, G., *Criteri per la trascrizione di testi medievali latini e italiani*, (Quaderni della Rassegna degli Archivi di Stato), Roma, 1982, 66 p.: . 7526@BALSAMO, L. - AVELLINI, L. - QUAQUARELLI, L., *Incunaboli Bolognesi (1471-1500). Procedure informatiche per l'analisi dei caratteri. Attività sperimentale 1990-1991*, Bologna, 1991: . 7561@GREENSTEIN, Daniel I. (ed.), *Modelling Historical Data: Toward a Standard for Encoding and Exchanging Machine-Readable Texts*, (Halbgraue Reihe zur Historischen Fachinformatik - Series A: Historische Quellenkunden, 11), St. Katharinen, 1991, xi-223 p.: Max-Plank-Institut für Geschichte in Kommission bei Scripta Mercaturae Verlag. 9225@SMITH, Joan M., *Transmitting Text: A Standard Way of Communicating Characters. Part 1 and 2*, «Association for Literary and Linguistic Computing Bulletin», 11 (1983) 2 e 3, 31-38; 63-67. 9279@SPERBERG-McQUEEN, C. Michael, *Text in the Electronic Age: Textual Study and Text Encoding, with Examples from Medieval Texts*, «Literary and Linguistic Computing. Jour-

nal of the Association for Literary and Linguistic Computing», 6 (1991) 1, 34-46. 9280@CLEWS, John, *Language Automation Worldwide: The Development of Character Set Standards*, Harrogate (UK), 1988, 165 p.: SESAME Computer Projects. 9333@BARNARD, David T. - FRASER, Cheryl A. - LOGAN, George M., *Generalized Markup for Literary Texts*, «Literary and Linguistic Computing. Journal of the Association for Literary and Linguistic Computing», 3 (1988) 1, 26-31. 9621@THALLER, Manfred, *A Draft Proposal for the Coding of Machine Readable Sources*, «Historical Social Research. Historische Sozialforschung», (1986) 40, 3-46. 9624@THALLER, Manfred, *A Draft Proposal for a Format Exchange Program*, in: [9376] J.-Ph. Genet (ed.), Standardisation et échange des bases de données historiques, Paris 1988, pp. 329-375. 9777@MARSDEN, Alan (ed.) - POPLE, Anthony (ed.), *Computer Representations and Models in Music*, London, 1991, 300 p.: Academic Press.

sgml:

1383@BARNARD, David T. - HAYTER, Ron - KARABABA, Maria - LOGAN, George M. - McFADDEN, John, *SGML-Based Markup for Literary Texts: Two Problems and Some Solutions*, «Computers and the Humanities», 22 (1988) 4, 265-276. 1395@BRYAN, Martin, *SGML. An Author's Guide to the Standard Generalized Markup Language*, Wokingham (UK), 1988, xvii-364 p.: Addison-Wesley. 4912@ROBINSON, Peter M.W., *The Transcription of Primary Textual Sources using SGML*, (Office for Humanities Communication Publications), Oxford, 1993: Office for Humanities Communication. 4955@GOLDFARB, Charles F. - RUBINSKY, Yuri, *The SGML Handbook*, Oxford, 1992 (ristampa), xxiv-663 p.: Clarendon Press. 4994@VAN HERWIJNEN, Eric, *Practical SGML*, Dordrecht-Boston-London, 1990, xviii-307 p.: Kluwer Academic

Publishers. 9343@SMITH, Joan M., *The Standard Generalized Markup Language (SGML) for Humanities Publishing*, «Literary and Linguistic Computing. Journal of the Association for Literary and Linguistic Computing», 2 (1987) 3, 171-175. 9490@COVER, Robin C. - DUNCAN, Nicholas - BARNARD, David T., *The Progress of SGML (Standard Generalized Markup Language): Extracts from a Comprehensive Bibliography*, «Literary and Linguistic Computing. Journal of the Association for Literary and Linguistic Computing», 6 (1991) 3, 197-209.

Thaller:

THALLER, Manfred, *Auf dem Weg zu einem Standard für maschinenlesbare Quellen*, in: [5303] P. Becker et Al. (eds.), Datennetze für die historischen Wissenschaften?, Graz 1987, pp. 228-238. THALLER, Manfred, *Secundum manus. Zur Datenverarbeitung mehrschichtiger Editionen*, R. Härtel - G. Cerwinka - W. Höflechner - O. Pickl - H. Wiesflecker, Geschichte und ihre Quellen. Festschrift für Friedrich Hausmann zum 70. Geburtstag, Graz, Akademische Druck- und Verlagsanstalt, 1987, pp. 629-638. THALLER, Manfred, *Historical Information Science: Is There Such a Thing? New Comments on an Old Idea*, in: [3643] T. Orlandi (ed.), Discipline umanistiche e informatica, Roma 1993, pp. 51-86. THALLER, Manfred, *Warum brauchen die Geschichtswissenschaften fachspezifische datentechnische Lösungen? Das Beispiel kontextsensitiver Datenbanken*, in: [5918] M. Thaller - A. Müller (eds.), Computer in den Geisteswissenschaften, Frankfurt-New York 1989, pp. 237-264. THALLER, Manfred, *Can We Afford to Use the Computer: Can We Afford Not to Use It?*, in: [5212] H. Millet (ed.), Informatique et Prosopographie, Paris 1985, pp. 339-352. THALLER, Manfred, *Recycling the Drudgery. On the Integration of Software Supporting Secondary Analysis of Ma-*

chine-Readable Texts in a DBMS, in: [1529] L. Cignoni - C. Peters (eds.), Computers in Literary and Linguistic Research, Pisa 1984, pp. 253-268. THALLER, Manfred, *Methods and Techniques of Historical Computation,* in: [1430] P. Denley - D. Hopkin (eds.), History and Computing, Manchester 1987, pp. 147-156. THALLER, Manfred, *The Need for a Theory of Historical Computing,* in: [1010] P. Denley - S. Fogelvik - C. Harvey (eds.), History and Computing II, Manchester, Manchester Univerity Press, 1989, pp. 2-11. THALLER, Manfred, *Data Bases and Expert Systems as Complementary Tools for Historical Research,* in: [4856] R. Metz - E. Van Cauwenberghe - R. Van Der Voort (eds.), Historical Information Systems, Leuven 1990, pp. 21-31. JARAUSCH, Konrad H. - ARMINGER, Gerhard - THALLER, Manfred, *Quantitative Methoden in der Geschichtswissenschaft. Eine Einführung in die Forschung, Datenverarbeitung und Statistik,* Darmstadt, 1985, x-211 p.: Wissenschaftliche Buchgesellschaft. THALLER, Manfred, *Numerische Datenverarbeitung für Historiker. Eine praxisorientierte Einführung in die quantitative Arbeitsmethode und in SPSS (Statistical Package for the Social Sciences),* (Materialien zur Historischen Sozialwissenschaft, 1), Wien-Köln, 1982: Hermann Böhlaus Nachf. THALLER, Manfred, *Beyond Collecting: The Design and Implementation of CLIO, a DBMS for the Social-Historical Sciences,* in: [5328] R.F. Allen (ed.), The International Conference on Data Bases in the Humanities and Social Sciences 1983, Osprey (Fl) 1985, pp. 328-334. THALLER, Manfred, *Automation on Parnassus. CLIO - A Databank Oriented System for Historians,* «Historical Social Research. Historische Sozialforschung», (1980) 15, 40-65. THALLER, Manfred (ed.) - MÜLLER, Albert (ed.), *Computer in den Geisteswissenschaften. Konzepte und Berichte,* (Ludwig-Boltzmann-Institut für Historische Sozialwissenschaft - Studien zur Historischen Sozialwissenschaft, 7), Frankfurt-New York, 1989, 336 p.: Campus Verlag. THALLER, Manfred, *The His-*

torical Workstation Project, in: [5001] J. Smets (ed.), Histoire et Informatique, Montpellier 1992, pp. 251-260. THALLER, Manfred, *The Historical Workstation Project,* «Computers and the Humanities», 25 (1991) 2-3, 149-162. THALLER, Manfred (ed.), *Datenbanken und Datenverwaltungssysteme als Werkzeuge historischer Forschung,* (Historisch-Sozialwissenschaftlichen Forschungen, 20), St. Katharinen, 1986: Scripta Mercaturae. THALLER, Manfred, *A Draft Proposal for the Coding of Machine Readable Sources,* «Historical Social Research. Historische Sozialforschung», (1986) 40, 3-46. THALLER, Manfred, *The Daily Life of the Middle Ages, Editions of Sources and Data Processing,* «Medium Aevum Quotidianum», 10 (1987), 6-29. THALLER, Manfred, *|gr| kleiw #, Ein fachspezifisches Datenbanksystem für die Historischen Wissenschaften, Version 2.1.1,* Göttingen, 1988: Max-Planck-Institut für Geschichte. THALLER, Manfred, *A Draft Proposal for a Format Exchange Program,* in: [9376] J.-Ph. Genet (ed.), Standardisation et échange des bases de données historiques, Paris 1988, pp. 329-375. THALLER, Manfred, *Possiamo permetterci di usare il computer? Possiamo permetterci di non usarlo?,* «Quaderni storici» [Bologna], 20 (1985) 3 - NS n. 60, 871-889. KLENK, Ursula (ed.) - SCHERBER, Peter (ed.) - THALLER, Manfred (ed.), *Computerlinguistik und philologische Datenverarbeitung. Beiträge der Jahrestagung der Gesellschaft für Linguistische Datenverarbeitung e.V. 1986 in Göttingen,* (Linguistische Datenverarbeitung, 7), Hildesheim, 1987, viii-193 p.: Georg Olms.