

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)



ROMA 2006  
BARDI EDITORE  
EDITORE COMMERCIALE

*Si ringrazia la «Associazione Amici della Accademia dei Lincei»  
per la collaborazione offerta all'edizione del presente volume*

FINITO DI STAMPARE NEL MESE DI DICEMBRE 2006

Azienda Grafica Eredi dott. G. Bardi S.r.l. – 00186 Roma, Piazza delle Cinque Lune, 113

*Azienda con Sistema Qualità certificato da BVQI*

COMITATO ORDINATORE

TULLIO GREGORY  
MARIA GUERCIO  
CLAUDIO LEONARDI  
TITO ORLANDI

*Il Direttore del Centro Linceo Interdisciplinare «Beniamino Segre», Prof. Maurizio Brunori, apre il Convegno con le seguenti parole:*

Dear friends and colleagues, ladies and gentlemen,

Professor Giovanni Conso, President of the Accademia Nazionale dei Lincei, is sorry to be unable to attend this opening ceremony, and asked me, as the Director of the Centro Linceo Interdisciplinare, to welcome all of you to this international workshop.

Before opening the meeting, I cannot refrain from addressing our attention to the tragedy that has struck our country and the whole Western World with the terroristic attack in Nassiriya, where nineteen italian citizens were killed. I ask you kindly to stand up for a minute of silence, in memory of these victims. Thank you very much.

As a biochemist, I have no specific competence in the topic of this important workshop on the repositories of cultural heritage. My pleasure and duty is to express a vote of thanks to the people and the institutions involved in making the workshop a reality.

The workshop was conceived as a collaboration between the Centro Linceo Interdisciplinare, Erpanet, and the Fondazione Ezio Franceschini. The Centro Linceo Interdisciplinare was founded in 1977 by Prof. Beniamino Segre, the eminent mathematician who was also President of the Accademia dei Lincei. Its mission is to foster interdisciplinary research by launching scientific projects, by supporting scientific meetings and workshops, and by providing a limited number of fellowships; it so happens that one of today's speakers, D.ssa Paola Moscati, was one of the first people to be awarded a fellowship of the Centro Linceo.

This Meeting was brought to our attention by Prof. Tito Orlandi, Linceo and member of the Scientific Board of the Centro Linceo, to whom we are very grateful. Moreover, I would like to express the appreciation of the Accademia to the other members of the Comitato Ordinatore, Prof. Tullio Gregory, Linceo and Professor of History of Philosophy at the University La Sapienza; Professor Claudio Leonardi, Linceo and President of the Fondazione Ezio Franceschini; and Dottoressa Maria Guercio, archivist at the University of Urbino and one of the directors of Erpanet, who will chair this morning's session.

On behalf of the Accademia, I wish all of you a very interesting and productive meeting, and hope that you will have enough time to visit Rome, possibly starting with the splendid Villa Farnesina, a property of the Accademia, decorated with a number of beautiful frescoes, and restored a few years ago.

I would like finally to thank all the invited speakers for their learned contributions, as well as the Secretary of the Centro Linceo Interdisciplinare, Signora Anna Anastasi, who took care of the organization.

MARIA GUERCIO\*

DEPOSITI DIGITALI PER LA CONSERVAZIONE  
DEL PATRIMONIO CULTURALE

UN WORKSHOP DI ERPANET E DEL CENTRO LINCEO INTERDISCIPLINARE

*1. Introduzione.*

L'organizzazione e la gestione di depositi digitali affidabili costituisce un tema centrale nell'ambito delle questioni connesse alla conservazione delle memorie digitali. Gli ultimi anni hanno conosciuto notevoli progressi in questo ambito sia dal punto di vista delle riflessioni concettuali e metodologiche che per quanto riguarda le concrete realizzazioni. Il modello OAIS ha svolto un ruolo importante di riferimento in grado di fornire basi comuni e principi generali anche in questo ambito di ricerca. Il problema è tuttavia così complesso da richiedere ancora un lungo processo di sperimentazione, riflessione e monitoraggio prima che si possa ritenere consolidato un modello conservativo adeguato ai nuovi ambienti di produzione e tenuta documentaria. Non c'è tuttavia dubbio che la diffusione massiccia di risorse digitali da parte delle istituzioni pubbliche, dei centri di ricerca, delle imprese e ormai anche dei privati cittadini ha messo la questione al centro della riflessione e del dibattito internazionali al punto da determinare la decisione di organizzare sul tema d'intesa con il Centro Linceo Interdisciplinare un workshop nell'ambito del progetto ERPANET con l'obiettivo di identificare e discutere con esperti del mondo accademico, della ricerca e delle imprese e delle istituzioni pubbliche, provenienti da numerosi Paesi e tradizioni e da diversi ambiti della conservazione del patrimonio culturale di natura digitale i principali nodi scientifici, tecnici, organizzativi e procedurali necessari a una corretta e soddisfacente implementazione di tali depositi.

\* Università degli Studi di Urbino «Carlo Bo» - Via Saffi, 2 - 61029 URBINO PU.

## *2. La struttura del Workshop.*

L'iniziativa si è articolata in quattro sessioni di lavoro, oltre a quella introduttiva che ha visto la partecipazione di Robert Kahn, direttore di uno dei più importanti centri di ricerca nel campo della produzione di risorse digitali, e di Tito Orlandi, coordinatore del workshop per il Centro Linceo, che hanno identificato le principali componenti e le criticità nella creazione e gestione di depositi per la conservazione di risorse digitali. Ad ogni sessione è seguita, nella tradizione di ERPANET, una discussione organizzata per piccoli gruppi di lavoro su alcuni nodi critici affrontati dai relatori.

Come è stato sottolineato nelle relazioni introduttive, il tema è uno dei più complessi nell'ambito della funzione conservativa sia per l'insufficiente esperienza maturata in questo ambito dalle istituzioni cui tradizionalmente è affidato tale compito, sia per la molteplicità delle attività e dei requisiti che devono essere sviluppati per garantire l'adeguatezza e la qualità dei depositi. Il workshop ha consentito – grazie alla presenza di testimoni qualificati – di identificare e discutere solo alcune delle questioni centrali che richiedono di essere indagate: standard per la predisposizione di contenuti digitali qualificati, criteri per l'acquisizione delle risorse, metodi e procedure per l'archiviazione e la gestione dei depositi, tecnologie e piattaforme applicative.

Il primo aspetto, la qualità dei contenuti delle risorse digitali destinate alla conservazione permanente, presenta naturalmente caratteristiche alquanto diversificate in base alla natura del patrimonio conservato e degli obiettivi specifici delle istituzioni preposte al compito conservativo, mentre per gli altri temi le testimonianze e le riflessioni si sono potute maggiormente concentrare sugli aspetti metodologici comuni anche a settori radicalmente diversi, creando i presupposti per una comparazione e una verifica congiunta. Ogni sessione del workshop è stata comunque sviluppata facendo riferimento a casi ed esperienze concrete in grado di fornire materiali per l'analisi e la discussione delle diverse criticità.

Il tema nella sua complessa articolazione presenta molti aspetti che meritano di essere approfonditi e molte domande, in parte affrontate nel corso dell'incontro, la cui risposta convincente avrà tuttavia ancora bisogno di numerose altre esperienze e soprattutto di confronti e di iniziative di ricerca. Alcuni di questi interrogativi – in particolare quelli di maggior impatto or-

ganizzativo – erano stati presentati sin dall'avvio del workshop e sono stati ripresi soprattutto nel corso delle sessioni di *break-out*:

1. è possibile e/o necessario elaborare una definizione condivisa di deposito digitale affidabile/certificato?
2. quanto divergono le tipologie in fase di definizione presso le diverse comunità disciplinari?
3. quali realizzazioni in corso di sviluppo a livello internazionale possono meritatamente denominarsi «trusted digital repositories»?
4. quanto diversificate sono le metodologie che assicurano l'autenticità e l'affidabilità della conservazione dei diversi patrimoni culturali nei contesti differenziati delle comunità di pratiche oggi in via di affermazione?
5. quali sono le responsabilità e i ruoli che non possono essere trascurati?
6. sono sufficientemente espresse e valutate le potenzialità dei nuovi istituti?
7. quali aspetti richiedono di essere ancora identificati nell'ambito dei progetti di ricerca e di implementazione in corso?

Ad alcuni di questi interrogativi, soprattutto a quelli che più si riferiscono a questioni di metodo e di impostazione generale, hanno dato una prima risposta di orientamento le due relazioni introduttive: Tito Orlandi affrontando il tema della qualità e della normalizzazione dei contenuti e Robert Kahn presentando un ricchissimo quadro complessivo dei principi, delle componenti, delle funzioni dei depositi digitali.

Il problema complesso di individuare, nei nuovi ambienti digitali, la quantità e la natura delle informazioni che individuano e accompagnano il documento nel passaggio dalla dimensione tradizionale a quella della produzione e conservazione digitale richiedono – ha ricordato Orlandi – ben altra attenzione rispetto a quella finora dedicata alla questione da parte degli autori dei contenuti medesimi e dei loro editori. Il primo passo è quello di definire in modo chiaro le fasi della procedura di «pubblicazione» delle risorse digitali con particolare riferimento al *riconoscimento dell'informazione* rispetto alla sua *rappresentazione*. Non diversamente da quanto avviene nel mondo della stampa tradizionale, è indispensabile individuare nel caso della riproduzione (e quindi nelle fasi necessarie ad assicurare la conservazione digitale) quali elementi vadano mantenuti con esattezza e quali possono essere cambiati. Come ha ricordato il relatore, non c'è su questo

aspetto neanche nell'edizione cartacea dei testi una teoria esplicita ma solo un consenso generalizzato. L'introduzione di strumenti informatici in questo ambito rischia di creare un clima di fiducia mal riposta a meno che non sia accompagnata da una consapevolezza dei molteplici nodi che rendono la questione di particolare rilevanza: tra questi, ad esempio, la distinzione sul piano concettuale tra le due funzioni tradizionali, in parte contraddittorie, dell'edizione a stampa: «quella di restituzione di un testo autentico e quella di fruizione di tale testo da parte dei lettori; in termini informatici, la conservazione dell'informazione contenuta nel documento, e la sua trasmissione ad uno specifico destinatario». L'errore più comune e più grave in questo ambito è la sottovalutazione della molteplicità dei piani di analisi e trattamento del documento digitale rispetto alla sua visualizzazione. La soluzione – ha ricordato Orlandi – non è univoca: può essere affidata a imprese centralizzate, con forti finanziamenti, che si dedicano a ingenti *corpora* omogenei; oppure può far riferimento a una modalità di lavoro distribuito che necessita soltanto di un accordo su standard formali, che garantiscano l'interoperabilità e il riuso in modi e ambienti differenziati.

L'alternativa tra la centralizzazione delle funzioni in grandi istituti pubblici e la creazione di una rete affidabile di centri di ricerca e di servizio ha costituito uno dei punti di riflessione e discussione del workshop, ripreso da molti relatori, a cominciare dalla seconda relazione introduttiva, grazie alla quale Robert Kahn ha offerto ai partecipanti – come si è detto – un esauritivo esame della struttura, delle componenti, delle funzionalità, del sistema di gestione e descrizione dei depositi digitali e delle risorse conservate. In particolare Kahn ha concentrato la sua attenzione sul sistema di gestione e sull'architettura dei depositi che richiedono la identificazione delle loro componenti tecniche fondamentali secondo criteri di efficienza e qualità: oggetti digitali, identificatore univoco di strutture e contenitori di dati sviluppato come servizio Internet distribuito, in grado anche di localizzare in rete i depositi che contengono oggetti digitali, strumenti di registrazione dei metadati. La rilevanza di dar vita a un sistema di depositi federati viene considerata un elemento critico di successo sia nel caso di depositi eterogenei sia nel caso di sistemi di depositi integrati. Tra gli altri temi affrontati Kahn ha sottolineato la centralità della gestione dei metadati sia per quanto riguarda la loro funzione identificativa e descrittiva sia con riferimento alle condizioni di accesso e uso degli oggetti medesimi.

### 3. Gli standard e le procedure per la qualità dei contenuti digitali.

Come già anticipato nell'intervento introduttivo di Tito Orlandi, la qualità dell'archivio e delle produzioni e riproduzioni digitali si basa su principi e standard che tuttavia devono rispondere alle specificità di ciascun settore. La prima sessione di lavoro ha avuto il compito di analizzare la questione per alcuni ambiti disciplinari e di settore particolari: quello della creazione di depositi digitali di edizioni elettroniche (Peter Robinson, *Presenting a text in many versions: an electronic edition of Dante's Commedia*, Manfred Thaller, *From national cultural heritages in the digital world to a digital European cultural heritage*), quello della produzione di contenuti digitali normalizzati e strutturati per la ricerca archeologica (Paola Moscati, *Linguaggi di marcatura per la conservazione e la valorizzazione dell'informazione archeologica*), nonché quello della certificazione, autenticazione e riconoscimento scientifico delle pubblicazioni e dei risultati delle attività di ricerca in ambiente digitale (Raul Mordenti, *Pubblicazione delle ricerche umanistiche in ambiente digitale*), delle nuove strutture cognitive necessarie a restituire nella forma digitale i tradizionali modelli ipertestuali (Claudio Leonardi ed Emiliano Degl'Innocenti, *Fili e trame di Arianna nella massa digitale*), delle nuove forme di pubblicazione e delle trasformazioni che oggetti nuovi, quale l'*e-book*, determinano nel rapporto tra produzione culturale, fruizione e mercato editoriale (Gino Roncaglia, *From printed books to e-books: (cultural) costs and benefits*). Dino Buzzetti (*Biblioteche digitali e oggetti digitali complessi*) si è concentrato sul problema dell'adeguatezza in termini di esaustività e funzionalità dei nuovi oggetti digitali rispetto a quelli tradizionali sia per quanto riguarda la completezza dell'informazione che con riferimento ai requisiti per una corretta lettura e interpretazione dei testi disponibili. Semantica e sintassi acquistano oggi un nuovo ruolo e nuove possibilità grazie ai linguaggi di marcatura. Nuove criticità e nuovi interrogativi devono essere tuttavia affrontati a cominciare dalla necessità di sviluppare metodi adeguati sul piano scientifico per l'utilizzo di linguaggi di marcatura, in particolare la possibilità di sviluppare regole e strumenti di elaborazione automatica per promuovere una «semantica formale» del mark-up e un metodo di rappresentazione nei diversi ambiti disciplinari con riferimento alla natura degli schemi e delle DTD da identificare, alla elaborazione di forme di rappresentazione che distinguano le strutture dai modelli di dati, intesi come algoritmi eseguibili utilizzando i dati disponibili, all'esigenza di comporre l'obiettivo dell'esaustività con quello della funzio-

nalità, alla capacità di sviluppare architetture differenziate in base alla natura dei documenti digitali complessi conservati (risorse digitali originarie, risorse digitali predisposte per la conservazione o per la distribuzione). La complessità che non può essere ignorata riguarda anche l'identità stessa dell'oggetto digitale che non necessariamente coincide con una identificazione univoca del documento rappresentato sia perché le riproduzioni non sempre includono le funzionalità collegate, sia perché si rendono necessarie per obiettivi funzionali diversi riproduzioni diverse dello stesso oggetto (ad esempio in formato XML e/o in formato PDF). Nel complesso tutte le relazioni della sessione hanno riconosciuto quanto esplicitato nella presentazione di Dino Buzzetti, il fatto cioè che il contenuto dei depositi digitali, qualunque sia la loro natura e indipendentemente dai loro obiettivi specifici, non potrà che includere oggetti multipli e varianti che facciano riferimento a un elemento invariante rappresentato dal contenuto informativo. Variabilità e invarianza costituiscono quindi, nel mondo digitale, molto di più rispetto all'ambiente tradizionale, il nodo difficile da sciogliere in termine di identità del documento.

Lo sviluppo di XML, ovvero di un meccanismo di linearizzazione dell'informazione, di un metodo di rappresentazione seriale che consente allo stesso tempo di ridurre a testo, a stringhe di bit il documento, pur identificandone anche gli elementi informativi interni/esterni di auto-riferimento, costituisce una nuova importante tappa di questo processo che, naturalmente, presenta anche numerose e non ancora sufficientemente chiarite criticità a partire dall'esigenza di verificare funzionalmente e scientificamente la natura e la qualità degli elementi informativi utilizzati e utilizzabili. Altrettanto importante – ed altrettanto complesso – è l'obiettivo sottolineato in particolare da Paola Moscati di salvaguardare l'integrità dei materiali anche in riferimento alla loro specifica complessità e alla complessità delle attività scientifiche di verifica, valutazione, confronto che la conservazione del patrimonio culturale digitale deve assicurare. Come ha infatti ricordato Manfred Thaller, le risorse culturali hanno successo solo quando riescono a influenzare la ricerca scientifica e, quindi, contribuirne allo sviluppo. È pertanto indispensabile – come del resto è emerso dalla discussione – che le nuove generazioni di studiosi dedichino attenzione particolare e cura alla fase di creazione degli oggetti digitali, né d'altra parte è possibile sottovalutare l'impatto enorme che il nuovo ambiente di produzione, gestione, fruizione avrà sul problema dei costi in fase di creazione e di accesso alle risorse. Anche il sistema delle responsabilità in questo ambito è destinato a modificarsi mediante un allargamento degli at-

tori e una frammentazione delle attività di documentazione e verifica che non potranno più limitarsi ai soli istituti di conservazione tradizionali (archivi e biblioteche).

#### *4. Le attività di acquisizione.*

È con la seconda sessione che il seminario è entrato nel vivo della questione conservativa e del rapporto difficile che lega la creazione di depositi digitali alla necessità di disporre a lungo termine di risorse informative accessibili, qualificate, integre e autentiche. Complessità e criticità sono state sempre affrontate nel corso di questa sessione e delle successive dal punto di vista duplice della riflessione teorica sui principi e sul metodo per la realizzazione di modelli conservativi e dell'analisi di concrete esperienze. È stato così possibile misurare il cammino effettivamente percorso in questi anni dagli operatori di settore e valutare lo stato dell'arte e i punti di eccellenza raggiunti da alcune istituzioni.

Il primo caso analizzato nella sessione destinata a trattare metodi e strumenti per l'acquisizione permanente a fini conservativi di risorse digitali è stato presentato da Seamus Ross e Adam Rusbridge (*Ingest strategies for digital Libraries*) e riguarda il tentativo condotto presso l'Università di Glasgow di creare un archivio digitale dei materiali scientifici presentati in occasione di convegni e congressi. La presentazione si è soprattutto concentrata sui nodi critici dell'operazione e ha cercato di illuminare gli aspetti metodologici dell'attività realizzata: la definizione dei flussi di lavoro, l'esigenza di rappresentare oggetti complessi, l'individuazione e il rispetto dei requisiti descrittivi con particolare riferimento alla identificazione dei metadati necessari a garantire conservazione e accesso, il tema centrale della fattibilità della funzione conservativa soprattutto quando entrano in gioco le implicazioni legate all'acquisizione manuale delle risorse e la relazione costi/benefici valutata con particolare riferimento ai tempi (e quindi alle risorse umane) necessari ad acquisire correttamente i materiali da conservare. Nel caso specifico i relatori hanno sottolineato le difficoltà legate soprattutto al trattamento di materiali non pubblicati: eterogenei e differenziati anche in relazione ai prodotti applicativi utilizzati per la loro produzione (a differenza di quanto avviene nel caso di lavori editoriali gestiti da personale specializzato), di difficile individuazione anche in termini di provenienza, spesso non strutturati e senz'altro privi

di documentazione originaria o di qualunque forma di documentazione in grado di sostenere il processo di descrizione e trattamento necessari. La relazione ha fornito un quadro alquanto esaustivo e piuttosto inquietante delle difficoltà che attendono le istituzioni di conservazione nei casi in cui gli archivi digitali non abbiano ricevuto alcun trattamento originario finalizzato al trattamento successivo di metadati descrittivi sia gestionali che scientifici: il lavoro preliminare necessario, la ricerca delle informazioni integrative attraverso canali non strutturati, l'esigenza in alcuni casi di individuare hardware non più disponibile per eseguire programmi in ambienti nativi, lo sforzo di recuperare collegamenti tra oggetti che hanno perso nel tempo le informazioni di interrelazione, la difficoltà di utilizzare software proprietari non più disponibili illustrano con sufficienza di elementi la difficoltà con cui sempre più frequentemente gli archivisti e i bibliotecari dovranno confrontarsi nei prossimi anni, in un ambiente che alle criticità tradizionali dovute al disordine e all'assenza di informazioni disponibili aggiunge la difficoltà di individuare gli stessi strumenti di lettura e accesso alle risorse medesime.

Il modello di workflow individuato da Rusbridge e Ross ha costituito un primo momento di verifica – nel corso del workshop – per affrontare uno dei nodi centrali dell'incontro: la definizione di un modello di flusso e delle attività necessarie alla tenuta del patrimonio digitale: selezione, registrazione, acquisizione, gestione e verifica dei rischi di virus, valutazione, preparazione fisica dell'acquisizione (struttura del disco, contenuto dei file, ecc.), descrizione, archiviazione ovvero identificazione e uso di metodi di conservazione appropriati. Una specifica attenzione è stata dedicata al problema della identificazione ed estrazione dei metadati e ai diversi gradi di difficoltà che si legano alla natura dei formati e delle piattaforme applicative utilizzate originariamente. Numerosi gli interrogativi di ricerca e di discussione che la relazione ha consentito di sollevare e di affrontare, a cominciare dai modi per garantire la corretta rappresentazione degli oggetti, dalla eventualità di definire una grammatica descrittiva, dalla possibilità di individuare e utilizzare strumenti che pur pensati per altri scopi siano in grado di estrarre attributi e proprietà dai file. È emersa con chiarezza la necessità di disporre di registri dei formati dei file e di archivi di metadati in quanto componenti essenziali di qualunque possibile strategia conservativa. Allo stesso modo è stata sottolineata la difficoltà di definire i costi, insieme all'esigenza imprescindibile di disporre di tecnici molto preparati e di standard che ancora richiedono di essere sviluppati adeguatamente, a partire da regole condivise per la categorizzazione e la conservazione dei metadati.

Il tentativo di sviluppare un approccio sistematico – nell’ambito specifico della documentazione archivistica – alla creazione di depositi per le risorse digitali è stato presentato con grande ricchezza di dettagli tecnici e con una significativa attenzione alle criticità metodologiche da Andrew Wilson con riferimento all’esperienza ormai pluriennale condotta dal National Archives of Australia (*Access across time: how the National Archives of Australia preserves digital records*). La strategia sviluppata dal NAA è sicuramente destinata a suscitare l’interesse crescente della comunità archivistica, sia per l’approccio realistico e insieme innovativo al problema che per lo sforzo di sistematicità e di coerenza teorica che l’esperienza illustrata da Wilson testimonia. Più che orientata alla costruzione di un *archivio digitale*, il progetto australiano è infatti orientato allo sviluppo in ambiente digitale della funzione conservativa. Una funzione che dovrebbe distinguersi, secondo Wilson, per l’indipendenza dei formati utilizzati e per lo sviluppo adeguato delle attività di valutazione, trasferimento, descrizione, ricerca. La conservazione – ha sottolineato il relatore – è tutt’altro che garantita dall’acquisizione di sistemi applicativi originari, che peraltro proprio per le restrizioni imposte dal mercato non sono in vendita se non nella forma riduttiva della possibilità d’uso. Poiché, a differenza di quanto avviene nell’ambiente tradizionale, le criticità della conservazione di risorse digitali non riguardano tanto gli oggetti ma la capacità di garantire la possibilità di riprodurne l’utilizzo, è sulla *performance* più che sull’*oggetto* medesimo che gli esperti devono concentrare i loro sforzi. Sulla base di queste considerazioni, la soluzione proposta dal NAA presenta due componenti essenziali: la tenuta di una copia master nella forma di una semplice stringa di bit per ciascuna risorsa acquisita e lo sviluppo – maturato nel caso specifico grazie all’esperienza condotta nel settore audiovisivo – di un approccio conservativo orientato al processo di fruizione della fonte («active intervention to recreate the performance»), che naturalmente richiede un’analisi preliminare degli elementi essenziali del documento che si intendono conservare. In ogni caso – ha ricordato Wilson che ha anche presentato in questa occasione le funzionalità del software applicativo (XENA<sup>1</sup> sviluppato

<sup>1</sup> XML Electronic Normalising of Archives. La presentazione ha consentito di illustrare le modalità operative della soluzione australiana: l’uso di un doppio sistema di *checksum* a fini di validazione, l’uno per il flusso di bit originario (*wrapped bitstream*), l’altro per la versione normalizzata; l’attenzione riservata alla documentazione delle azioni relative al trattamento dell’oggetto e necessarie per ricostruire e comprenderne il contesto e l’utilizzo; le modalità della codifica XML orientata a sviluppare schemi gestiti attraverso moduli *plug-in* che definiscono l’essenza della fonte documentaria.

dall'amministrazione archivistica australiana – è indispensabile l'identificazione di formati standard ben documentati (che nel caso specifico si basano sui linguaggi di marcatura XML), nonché gestibili con strumenti applicativi aperti, per la migrazione di tutto il materiale da conservare.

L'altro aspetto centrale nella costruzione e gestione dei depositi digitali, proprio in riferimento ai processi di acquisizione, riguarda la registrazione di formati digitali ed è stato discusso con una dettagliata relazione ricca di spunti di riflessione e di concreti riferimenti all'esperienza maturata in questo campo da alcune importanti istituzioni da Robin Dale (*Global registry of digital formats*). La centralità dello sviluppo e dell'implementazione di un sistema che individui e conservi i formati dei file («a format is a fixed, byte-serialized encoding of an *information model*») non è – secondo il relatore – in discussione, considerata la forte dipendenza dei depositi digitali da una conoscenza e un uso corretto nel lungo termine dei formati dei file in larga parte delle fasi della gestione (acquisizione, archiviazione, accesso). È tuttavia opportuno avere ben chiari gli obiettivi di un sistema di registrazione globale soprattutto se si tiene presente la complessità di tale funzione («maintain persistent, unambiguous bindings between public *identifiers* for digital formats and *representation information* for those formats») in relazione alla frammentazione e alla continua evoluzione delle tecnologie, ma anche alla necessità di catturare una molteplicità di *informazioni di rappresentazione* di ciascun formato, ovvero di identificarne le proprietà *sintattiche* e *semantiche* rilevanti, sia in termini descrittivi che in termini di *processo* e di *gestione*. Numerose sono ormai le ragioni per sostenere un simile sistema di registrazione: esigenze di identificazione, validazione, trasformazione, caratterizzazione, valutazione dei rischi, esibizione. A queste esigenze cerca di rispondere il modello in corso di sviluppo per iniziativa di una rete di istituzioni che operano nella forma di un comitato ad hoc cui partecipano università e centri di ricerca, archivi e biblioteche di grande prestigio. In particolare, oltre a definire una ontologia dei formati e individuare le proprietà di base per la loro caratterizzazione, il comitato ha individuato i servizi fondamentali connessi a un sistema di registrazione: l'approvazione delle entità medesime con particolare attenzione al grado di revisione tecnica e al livello di trasparenza e accesso, la tenuta inclusiva dell'aggiornamento, la notifica agli utenti di modifiche ed evoluzioni dei formati (ad esempio, dei rischi di obsolescenza), la definizione di politiche di importazione dei formati, la descrizione ovvero le informazioni di rappresentazione relative a ciascun formato, il trasferimento presso depositi esterni del sistema

di registrazione o di una sua parte. Attività sussidiarie includono il servizio di identificazione di un formato per uno specifico oggetto digitale mediante il confronto tra gli attributi di profilo della risorsa esaminata, il servizio di validazione in grado di verificare il formato di uno specifico documento digitale analizzandone gli attributi, il servizio di conversione dal formato sorgente a quello richiesto, il servizio di estrazione automatica dei metadati per determinati formati. Nella relazione si sono naturalmente individuate alcune delle principali criticità che caratterizzano la funzione esaminata:

- il rischio che un eccesso di granularità stabilita in modo discrezionale (ad esempio individuando versioni e profili locali) possa determinare una crescita abnorme di formati destinati alla registrazione;
- la necessità che il sistema di registrazione sia affidabile e sostenibile sia al fine di incoraggiare il deposito di informazioni proprietarie, sia al fine di giustificare i costi, il cui livello dipende dalla qualità e dal grado di autorevolezza previsti;
- l'esigenza di definire il modello di sviluppo del sistema operativo di registrazione rispetto alla duplice possibilità di un'autonoma iniziativa da parte degli utenti o di un intervento centralizzato affidato in esclusiva al personale specializzato del deposito;
- l'esigenza di definire il modello di governo di un sistema sufficientemente affidabile valutando adeguatamente i rischi connessi al collocamento di tale iniziativa all'interno di un'organizzazione già operativa e alla compatibilità di obiettivi globali con l'eventuale circoscrizione nazionale del deposito.

La complessità della funzione, i molteplici requisiti da rispettare, gli impegnativi interrogativi da sciogliere non devono tuttavia far rinunciare – ha concluso Robin Dale – alla implementazione del sistema che è invece destinato ad acquisire un ruolo crescente soprattutto con riferimento alle impegnative esigenze della conservazione a lungo termine, in quanto strumento idoneo ad assicurare la persistenza delle risorse.

##### *5. Memorizzazione e depositi: metodi e procedure.*

L'esperienza ormai ultradecennale degli Archivi nazionali di Washington – una delle più lunghe e significative nel campo della conservazione digitale – è stata presentata da Fynnette Eaton (*Repository design and creation: a*

*case study*), che ha innanzi tutto fornito un'accurata descrizione della natura della funzione conservativa sviluppata dal NARA e della sua evoluzione nel corso dell'ultimo decennio con specifico riferimento all'approfondimento dei problemi teorici e metodologici connessi al trattamento delle risorse digitali reso possibile dalla partecipazione diretta dell'istituto statunitense ai progetti di ricerca InterPARES 1 e 2, alla collaborazione con il San Diego Supercomputer Center e al coinvolgimento nello sviluppo dello standard ISO OAIS in partnership con la National Aeronautics and Space Administration. La relazione si è tuttavia concentrata soprattutto sulla identificazione delle nuove strategie operative presenti nel progetto ERA sviluppato con l'obiettivo specifico di costruire un servizio dinamico e scalabile per la conservazione autentica di lungo termine di risorse digitali differenziate nei diversi formati, ma in modo da assicurare il più alto grado possibile di indipendenza dalle specifiche infrastrutture hardware e software esistenti. Nell'ambito del progetto – ha sottolineato Fynnette Eaton – tutte le attività archivistiche sono oggetto di implementazione:

- l'acquisizione e archiviazione in regime di sicurezza;
- l'ordinamento;
- la descrizione;
- l'accesso che privilegia l'ambiente web;
- la selezione.

È degno di nota – ed è emerso chiaramente nella presentazione – l'impegno pluriennale dedicato alla ricerca, oltre alla determinazione di dedicare al progetto risorse umane altamente qualificate. Lo sforzo di analisi e di interrelazione con il mercato e la scalabilità del progetto, nonché la capacità di gestire facilmente interfaccia personalizzabili e documenti anche informali e temporanei, sono stati ricordati nella relazione e hanno costituito elemento di discussione successiva in quanto ritenuti elementi innovativi di un progetto alquanto complesso in tutte le sue componenti.

A Robin Dale è poi toccato il compito di affrontare il tema cruciale dell'attività di certificazione e audit dei depositi digitali (*Certification and audit*), ritenuto da tutti i professionisti coinvolti nei progetti di conservazione digitale, ma anche dagli studiosi più attenti, come una delle sfide più impegnative in questo ambito. La presenza di una rete di organizzazioni affidabili e fidate non a caso è stata definita nel 1996 dalla Task Force on Archiving of Digital Information – e ricordata dal relatore in apertura al suo intervento – «a critical component of the digital archiving» e la certificazione un processo necessario

per creare «an overall climate of trust about the prospects of preserving digital information».

Le criticità in questo ambito emergono non appena si cerca di definire il concetto stesso di certificazione e le modalità della sua trasposizione in ambiente digitale con riferimento all'organizzazione di depositi digitali fidati. Indipendenza nella gestione del processo di certificazione e autorevolezza delle istituzioni sono secondo Robin Dale componenti essenziali di tale processo, con la conseguenza che non tutti gli enti che gestiscono memorie digitali saranno in grado di trasformarsi in depositi certificati, ma dovranno acquistare – come peraltro indicato dal rapporto RLG-OCLC – servizi certificati da terzi, capaci di assicurare la predisposizione e l'utilizzo di un modello funzionale e di un modello informativo adeguati (ad esempio rispondenti ai requisiti dello standard OAIS), l'impiego avanzato e diffuso di standard, procedure di sicurezza e sostenibilità sia finanziaria che tecnica per le funzioni svolte. Il gruppo di lavoro ([www.rlg.org/longterm/certification.html](http://www.rlg.org/longterm/certification.html)), che si è dedicato in questi ultimi anni alla definizione delle peculiarità e dei requisiti della certificazione grazie al supporto di importanti istituzioni di livello internazionale soprattutto nordamericane e britanniche (US National Library, NARA, agenzie aerospaziali, università e centri di ricerca scientifici, Internet Archive, RLG, DPC, OCLC), ha identificato le diverse tipologie di elementi che consentono la valutazione e lo sviluppo di un sistema di certificazione in termini di attributi del deposito, di politiche adottate, di funzioni svolte e ha in fase di sviluppo l'individuazione di una struttura di riferimento e un rapporto documentato sulle componenti essenziali ora ricordate. Una serie di interrogativi rilevanti sono ancora inesplorati: ad esempio, ricorda Dale, l'esistenza di livelli diversi di certificazione, la possibilità di processi di auto-valutazione o di valutazione incrociata, la natura della conservazione assicurata (a livello di bit o inclusiva delle funzionalità per la completa restituzione dell'oggetto digitale?), il riconoscimento dell'auto-certificazione (a proposito del quale il relatore ha espresso forti riserve), la considerazione dei costi del processo in relazione alla sua efficacia. Per quanto riguarda in particolare il modello di conservazione tra i requisiti minimi individuati il relatore ha sottolineato la necessità che gli oggetti siano archiviati correttamente (ovvero insieme ai metadati di contesto), siano ricercabili e accessibili.

Il caso concreto di un deposito digitale sviluppato in ambito scientifico con il compito della disseminazione delle informazioni conservate è stato presentato da Henk Harmsen (*Dissemination of scientific data: the DARE com-*

*munity*) nell'ambito di un progetto di ricerca e sviluppo realizzato dal governo olandese d'intesa con le istituzioni accademiche nazionali con l'obiettivo di creare l'esperienza e definire le condizioni perché ogni istituzione di ricerca disponga di un proprio deposito digitale aperto nell'accesso e qualificato nella conservazione, destinati peraltro a trasformarsi sensibilmente rispetto al contesto tradizionale. Il relatore ha messo soprattutto l'accento sulle numerose difficoltà che accompagnano tale investimento sia per quanto riguarda la gestione dei materiali sul web, sia per quanto concerne le modalità di cooperazione nell'ambiente scientifico e lo sviluppo di strumenti di «pubblicazione», che naturalmente dovranno basarsi su standard aperti (OAI, XML, PDF, ecc.) e utilizzare strategie complessive di comunicazione e condivisione adeguate al patrimonio da gestire e in base a criteri finalizzati alla facilità d'uso e alla flessibilità, oltre che coerenti con le funzioni previste. Se infatti il progetto DARE ha sviluppato un sistema aperto per l'accesso, ha scelto un modello accentrato di conservazione affidandone i compiti esclusivamente alla Koninklijke Bibliotheek. Questa decisione non ha tuttavia implicato la rinuncia a condividere le responsabilità con le istituzioni che producono gli oggetti digitali: la strada del coinvolgimento dei soggetti produttori, ritenuto un requisito di successo del progetto, prevede infatti che il concetto di conservazione sia incorporato nel modello stesso di «open access». Le questioni che in questo ambito sono ancora irrisolte – e che la discussione che è seguita alla presentazione non ha mancato di ricordare – riguardano tra l'altro la difficoltà di promuovere il concetto della conservazione (ma anche dell'*open access* che per ora sembra svilupparsi quasi esclusivamente nella forma di vere e proprie pubblicazioni elettroniche) e la responsabilità e i tempi per la selezione dei materiali.

#### 6. *Tecnologie e applicazioni emergenti: studi di casi.*

Gli studi di casi dell'ultima sessione sono stati allo stesso tempo un'occasione per illustrare e discutere esperienze e riprendere e approfondire i tanti temi di riflessione che le sessioni precedenti avevano già in parte affrontato. I risultati del programma LOCKSS, presentato da Victoria Reich (*Custodianship of cultural web content: the LOCKSS Program*), hanno in particolare messo l'accento sulle necessità (e sulla possibilità dimostrata proprio dal prodotto in questione) di sostenere in questa difficile fase di transizione i centri

tradizionali di conservazione, tra cui in particolare le biblioteche per quanto riguarda i contenuti del web (tesi e dissertazioni elettroniche, documenti di *e-government*, letteratura grigia, documenti e pubblicazioni diverse, altri materiali web), sviluppando un modello semplice, di facile utilizzo, basso costo e affidabile, soprattutto grazie all'uso di prodotti *open source* e alle procedure seguite basate sulla ridondanza dei materiali raccolti oltre che sulla verifica e comparazione delle diverse versioni di contenuto catturate e sulla distribuzione/condivisione delle attività che implicano quindi forme interne di continua validazione dei materiali e garanzia contro le perdite.

Anche questo programma presenta naturalmente criticità che non possono essere ignorate:

- richiede l'accesso autorizzato degli editori;
- ha bisogno di versioni certificate del materiale catturato;
- non consente la conservazione del web dinamico;
- implica una notevole memoria di archiviazione;
- deve sviluppare un sistema rilevante di alleanze;
- non è per ora in grado di superare l'ambiente http.

Anche il progetto i-Tor, presentato da Henk Harmsen nella sua seconda relazione (*i-Tor: tools and technologies for open repositories*) ha messo l'accento sulle possibilità e sui vantaggi degli standard e dei sistemi aperti e interoperabili e sull'architettura seguita nell'ambito del modello OAI per lo sviluppo dello strumento di acquisizione degli oggetti digitali prevista dal progetto DARE ricordato in precedenza. Il problema della definizione di schemi di metadati condivisi, ma anche sufficientemente specifici e adeguati alla complessità del patrimonio documentario è stato oggetto di un'ampia discussione che ha peraltro trovato un'eco rilevante nella relazione di Sergio Bernardi dedicata all'uso della tecnologia DOI (*Digital voluntary deposit and preservation: the use of DOI technology*) e in quella di Marco Veneziani (*MAG: a metadata set for managing and preserving Italian digital resources*). In tutti i casi descritti si è focalizzata l'attenzione – sia pure con impostazioni diverse – sul ruolo centrale dei metadati per la identificazione univoca delle risorse e sulla diversità di approccio tra chi (soluzione DOI) punta sostanzialmente sulla localizzazione permanente e chi invece intende formare oggetti in formati standard (soluzione i-Tor). Una via intermedia è invece quella seguita dalla Biblioteca digitale italiana per il trattamento e la conservazione dei prodotti digitalizzati che comunque sembra ancora finalizzata – almeno in questa

fase peraltro ancora interlocutoria – alla localizzazione permanente dell’oggetto più che a crearne un surrogato in formato standard.

A Stephan Heuscher (*Workflow modeling language evaluation for an archival environment*) è toccato infine il compito di affrontare il tema dell’autenticità degli archivi/depositi digitali nei progetti di conservazione permanente offrendo alla discussione le riflessioni maturate nel corso del progetto ARELDA condotto dagli archivi nazionali svizzeri.

Definendo innanzi tutto con chiarezza la distinzione tra biblioteche e archivi digitali (che includono oltre al valore informativo anche il valore di testimonianza affidabile), il relatore ha ricordato le caratteristiche essenziali che garantiscono in ambiente tradizionale l’affidabilità del patrimonio documentario (integrità, intelligibilità, originalità, accessibilità) e che per i documenti digitali presentano non pochi elementi contraddittori tutt’altro che risolti in questa fase dalle infrastrutture tecniche disponibili in grado di garantire la sicurezza dei dati solo per tempi limitati, come nel caso della firma digitale. Tra le difficoltà connesse alla conservazione autentica di risorse digitali un rilievo particolare – ha sottolineato Heuscher – assume il problema della perdita delle caratteristiche di originalità a seguito degli inevitabili processi di migrazione, che può essere affrontato solo con la completa documentazione del processo conservativo, la definizione di procedure rigorose e automatizzate, il mantenimento di tutti i dati di contesto che consentano la verifica, l’interpretabilità, l’utilizzo nel tempo delle risorse medesime, lo sviluppo di infrastrutture adeguate e sostenibili sul piano finanziario e organizzativo.

## 6. Conclusioni.

Se gran parte delle questioni più rilevanti e cruciali sono state oggetto delle relazioni presentate nel corso del workshop, non c’è dubbio tuttavia che molti aspetti richiedano ulteriori approfondimenti e analisi, come del resto è emerso con chiarezza nelle sessioni di *break-out* seguite alla fine del secondo e terzo giorno di lavoro e in tutte le numerose occasioni di discussione che hanno accompagnato il suo svolgimento. Il dibattito che ha avuto luogo, intenso e ricco di spunti innovativi, ha permesso di individuare alcune raccomandazioni volte soprattutto a sviluppare la ricerca e a far maturare esperienze positive in questo campo sempre più rilevante nello stesso processo di

creazione – oltre che di mantenimento – delle risorse digitali. In sintesi, tra le conclusioni condivise del workshop sono emersi alcuni requisiti essenziali che è opportuno qui ricordare:

- la necessità di far maturare la consapevolezza del problema conservativo e della sua criticità presso tutti i portatori di interesse con l’obiettivo di trasformare quanto prima l’eccezione del deposito volontario in luoghi fidati in un comportamento abituale;
- l’importanza di sviluppare schemi e strumenti che facilitino il deposito degli oggetti e la loro descrizione mediante l’implementazione di sistemi automatici per l’estrazione dei metadati;
- la rilevanza dei criteri di flessibilità delle funzioni sempre più complesse connesse alla conservazione e all’accesso delle memorie digitali che richiedono comunque nuovi livelli di mediazione;
- l’esigenza di sviluppare:
  - un’ampia scelta di standard;
  - soluzioni integrate con altri sistemi di registrazione (dei formati e degli schemi di metadati);
  - strumenti di estrazione dei metadati;
  - strumenti di migrazione;
  - un livello minimo di attributi per la descrizione;
  - convenzioni di deposito che siano il risultato di una precisa negoziazione e che determinino una chiara assunzione di responsabilità.

Tra le specifiche raccomandazioni elaborate in occasione del workshop assume infine un ruolo significativo la necessità di ottenere un ruolo più attivo delle istituzioni di conservazione anche nell’attività volta a illustrare i benefici dei nuovi servizi, di sviluppare livelli di servizio coerenti con la qualità degli oggetti depositati, di definire priorità nella produzione di standard e nei processi di normalizzazione e di predisporre reti di depositi in grado di sostenersi e integrarsi reciprocamente.



TITO ORLANDI\*

## ARCHIVI ED INFORMAZIONE

È opportuno premettere alla mia comunicazione l'avvertenza che i problemi riguardanti propriamente e tecnicamente gli archivi saranno introdotti nella prossima comunicazione preliminare, di Robert Kahn, e poi trattati nelle sezioni terza-quinta, legate alla funzione e alla preziosa attività di Erpanet. Il problema che intendo affrontare rientra piuttosto nell'ambito di una delle attività di ricerca del Centro Linceo Interdisciplinare. Esso sarà il tema della seconda sezione, e riguarda archivi di documenti che non nascono digitali, ed in particolare il delicato passaggio dalla individuazione del documento, dalla presa di coscienza del suo essere documento, alla sua collocazione nell'archivio digitale.

La differenza fra l'archivio digitale e quello convenzionale è qui particolarmente evidente. Mentre nel secondo caso il manufatto stesso entra a far parte dell'archivio, e dunque si tratterà solo di organizzare il modo di recuperarlo in base all'informazione che esso contiene, nel caso dell'archivio digitale si tratta di trasformare il manufatto, il documento, nella sua *rappresentazione digitale*. Il tema è analogo a quello delle edizioni digitali<sup>1</sup>, ma presenta sue caratteristiche specifiche, in quanto l'edizione digitale di per sé obbedisce allo scopo che si prefigge l'editore, e l'informazione che conserva e trasmette si valuta sulla base di quello scopo. L'archiviazione invece deve *tendere* (il traguardo non si raggiunge mai a causa della nota legge dell'entropia dell'informazione, o se si preferisce, del fiume eracliteo) a conservare l'informazione che possa essere utile per scopi che possono essere al momento imprevedibili, e dunque *tutta* l'informazione che il documento contiene.

Si aggiunge che il documento tradizionale, in sostanza il documento cartaceo o comunque costituito di segni su una superficie, presenta un aspetto di

\* Facoltà di Lettere - Università degli Studi di Roma «La Sapienza» - Piazzale A. Moro, 5 - 00185 ROMA.

<sup>1</sup> Cf. *Scholarly Electronic Publishing Bibliography*: <http://epress.lib.uh.edu/sepb/sepb.html>

grave ambiguità, in quanto per un verso è esso stesso informazione; per un altro *trasmette, veicola* informazione. Nella procedura di rappresentazione digitale del documento cartaceo occorre prima di tutto chiarire e distinguere nel modo più esauriente possibile (anche in questo caso la completezza non sarà mai raggiunta) i tipi di informazione che esso contiene. Come ci insegnano la semiotica e la linguistica, i segni di un linguaggio possono avere, e di solito hanno, più livelli di lettura, di significato, di interpretazione, che pongono il problema di analisi, diciamo, di I grado, II grado, ecc., e del riconoscimento e del recupero della specifica informazione di ciascuno di questi tipi. Osserva Cesare Segre (1985: 29):

«... il testo è l'invariante, la successione di valori, rispetto alle variabili dei caratteri, della scrittura, ecc. Possiamo anche parlare di significati, se si precisa che si allude a significati grafici, quelli della serie di lettere e segni d'interpunzione che costituiscono il testo è dunque una successione fissa di significati grafici. Questi significati grafici sono poi portatori di significati semantici, come vedremo subito; ma occorre insistere in partenza su questa costituzione originaria»

e ad uno stesso ordine di problemi riconduce la teoria della «doppia articolazione» del linguaggio<sup>2</sup>.

Si evidenzia in questo modo una distinzione molto netta fra le due fasi della procedura di digitalizzazione del documento: da un lato il riconoscimento dell'informazione in esso contenuta; dall'altro la realizzazione della rappresentazione (digitale) dell'informazione. È naturale, ma non per questo meno errato, che anche da parte dei responsabili di parte umanistica si dia, in ambito informatico, importanza quasi esclusiva alla seconda fase, in quanto la prima sembra tutta appartenente ad un ambito disciplinare non informatico, sia esso della linguistica, della filologia, o altro, e dunque viene dato per scontato. In realtà, senza una stretta correlazione fra le due fasi, in cui cioè ciascuna nell'esplicarsi tenga conto dell'altra, il risultato è sostanzialmente improduttivo, come conto di dimostrare.

Per esemplificare concretamente quello che intendo, mi riferirò ad un caso specifico, rappresentato nella fattispecie dal testo dei *Sepolcri* del Foscolo. Esso fornisce un esempio, fra i tanti che si potrebbero portare, di come la mancanza di una vera cooperazione fra livello filologico e livello informatico possa dare risultati insoddisfacenti. Occorre prima di tutto ricordare,

<sup>2</sup> Cf. p. es. Malmberg (1994: 49-50), sulla scorta di Hjelmslev e Martinet.



generali, su cui il Foscolo, in collaborazione con il tipografo, basò la comunicazione del suo testo.

*Caratteri.* L'edizione nazionale così si esprime: «eleganti caratteri bodoniani nitidissimi di occhio marcato, impressi su carta a mano di qualità fine, bianca e spessa». Il corpo del testo «normale», se possiamo basarci sulle riproduzioni inviate dalla Biblioteca Braidense di Milano, è 14; quello dell'epigrafe 23 con le I a 31; quello del titolo 22; quello della dedica 14; quello della prima lettera del testo 23, seguita da due L maiuscoletto in corpo 8. Si noti che, se la riproduzione non fosse esattamente uguale all'originale, i rapporti rimarrebbero comunque validi.

*Grafemi.* Sono utilizzate le maiuscole e minuscole con esclusione di K, W, X, Y, ed inclusa J, che pone problemi per l'uso ortografico (cfr. sotto). Sono usati i politipi per fi, fl, ma non ffi, ffl. È usata la crenatura (e dunque politipi?) per fo, fa, fe, fu, el, rb, gl, gh, ch, rl (per quanto possiamo vedere). Sono usati i caratteri vocalici con accento acuto (salvo la u) e grave, il circonflesso su ô, e la dieresi su ï, ë, ä. Mi sembra di aver notato che spesso i due punti sulla ï non sono allineati, come se questo carattere avesse dovuto essere prodotto fuori del normale complesso di caratteri, in modo non professionale. È possibile che Foscolo l'abbia sollecitato? Non sono usate la lineetta (- o - o —) e le virgolette o i caporali. Se si ricorda la mania lineettaria del Foscolo nello Jacopo Ortis, questo segnala probabilmente una sua distinzione ortografica fra prosa e poesia, o meglio generi di poesia, perché in altre poesie la lineetta è usata.

*Distanza delle unità grafematiche.* **spazi** Nessuna osservazione da fare sulla divisione delle parole. Interessante è invece il sistema usato per i segni di punteggiatura. Sembra che la regola generale sia:

|carattere|mezzo-spazio|segno|spazio-intero|carattere|

con eccezioni che si riscontrano per: . , ; (punto, virgola, punto e virgola). Da notare che per il segno ' (apostrofe o elisione) la regola è invece:

|carattere'|mezzo-spazio|carattere|,

cosa che assimila fortemente casi come «de' cipressi» «l' urne» «die' Dárdano», e li differenzia dalla separazione per punteggiatura. Un caso particolare di distanza delle unità grafematiche può essere considerata l'indentazione, che è qui usata, salvo che all'inizio del carne. Non è usata invece una eventuale distanza fra le linee, maggiore di quella «normale».

Questo risulta essere l'insieme dei segni che il Foscolo volle o comunque ebbe a disposizione per comporre la stampa dei Sepolcri. Che la sua coerenza, e le regole secondo le quali è stato utilizzato, siano un oggetto d'indagine meno banale di quanto possa sembrare, credo risulti dalle seguenti constatazioni. Le maiuscole sono usate all'inizio di ogni verso, all'inizio del periodo, all'inizio del discorso diretto (preceduto dai due punti, ma non, come abbiamo notato, dalla lineetta, che alcuni editori aggiungono). Inoltre sono usate sia per i nomi propri o assimilati: Sole, Musa/e, Speme, Dea, Natura, Nume/i, Genj, Orco, Celeste/i, Ghibellin, Dei, Ninfa, Giulia (gente), Olimpio, Pelídi, Prenci; sia per per gli aggettivi derivati da nome proprio di genti o luoghi, salvo che per abduani (60, forse una svista).

La j è tendenzialmente usata per il plurale ii: effluj, Genj, auspicj; ma: ozi, silenzi, e (interessante) patrii. Inoltre in mezzo a due vocali: gioja, Ajace, Troja.

L'apostrofo è usato per l'elisione. Da notare che il tipo l'uomo, è trattato (cfr. sopra) esattamente come de' suoi, die', fe' cari.

L'accento è normalmente quello grave, usato per i plurisillabi con accento finale (udrò, perchè) o con accento sulla terzultima (ùpupa) e nei monosillabi per distinzione: nè, è, più, di. Tuttavia la distinzione fra che e chè suscita dei dubbi. L'uso dell'accento acuto pone molti problemi, oltre al fatto che non è mai usato sulla u. Per i casi di dieresi si vedrà sotto; altri casi sono di difficile spiegazione: Egée, Retée, Pimplée, uscían, Rapían, Sentía, Tróade, Dárdano, Assáraco, Tidíde, Pelídi, Oceáno, fuggía, nutría; ma si noti: armonia.

Il fenomeno della dieresi (diciamo così) poetica è indicato in due modi. Possono esserci i due punti ä, ë; ï (per questo cfr. anche sopra), ovvero quello che chiamiamo accento grave, nel solo caso della ù (luttuoso, abduani). Non è segnalata la dieresi in obbligo 18 (ma cfr. obbligate 86), Talia 54, funerea 83, santuario 114, beata 165.

Per quale motivo si trovi il circonflesso piuttosto che il grave su allôr non saprei dire.

Sulla punteggiatura preferisco non esprimermi, ma certamente il Foscolo ha avuto le sue ragioni quando ha messo due punti o punto e virgola piuttosto che punto, ecc.

Per quanto riguarda l'impaginazione, riteniamo prima di tutto che anche la scelta dei caratteri non sia stata casuale. L'indentatura è utilizzata per distinguere le sezioni, salvo che all'inizio del carme. Le linee (che cominciano sempre con la maiuscola) non sono numerate. Il titolo corrente (DEI SE-

POLCRI) sta sia sulla pagina pari sia su quella dispari; la numerazione delle pagine sta rispettivamente sulla sinistra o sulla destra. A parte il frontespizio, che a mio avviso in questi casi non conta, il carne include un'epigrafe, che è stampata in modo da imitare i glifi delle iscrizioni latine, ed una dedica.

A me sembra che da questa indagine emerga abbastanza chiaramente che il Foscolo aveva in mente un sistema ortografico e di impaginazione che non era una pura accettazione del sistema convenzionale (anche se non formalmente esplicitato) dei suoi tempi, che del resto non era ancora fissato come avverrà in séguito, ma obbediva, almeno in parte, ad alcune concezioni personali del poeta, le quali a loro volta interferirono con la concezione più propriamente poetica del carne. Di qui l'esigenza, secondo me imprescindibile, di includere nel documento elettronico anche questo tipo di informazione.

È vero che l'*ortografia* curata dallo stesso Foscolo è almeno in parte difforme dall'uso moderno, e dunque da quanto il lettore moderno si attende. Secondo una teorizzazione assai condivisa gli editori sono autorizzati, anzi invitati ad intervenire, al fine di presentare un testo accettabile per tale lettore, annullando però con questo tutto il sistema dell'autore, p. es. quello che abbiamo tentato di ricostruire sopra. Altrettanto condivisa è l'autorizzazione, anzi l'esortazione, ad intervenire sulla *punteggiatura*. Sono particolarmente significative le parole di Contini:

«Di un autografo (o suo equivalente) l'edizione interpretativa riproduce ciò che interessa e omette, intenzionalmente o spontaneamente, ciò che non interessa. In sostanza essa è la traduzione o adattamento di un sistema, storicamente individuato, in altro sistema» (Contini, 1986: 15-16);

«E tutti sanno che anche in epoca più recente fini letterati non disdegnarono di limare dall'esterno le scritture di autori provvisti di forte personalità poetica ma non di robusta cultura alfabetica ...» (Contini, 1986: 20);

di Stussi:

«Cautele editoriali del tipo indicato sono utili, tutte o in parte, per testi che vanno dal XIII al XV secolo ...; successivamente si possono adottare criteri più elastici arrivando anche ad interventi che uniformino grafie oscillanti o le modernizzino ... Condizioni irrinunciabili sono tuttavia, anche per un testo moderno: a) il non travisamento della realtà fonetica sottostante ai fatti grafici; b) l'enunciazione

esplicita dei criteri seguiti, soprattutto là dove si è avuto un intervento modificatore» (Stussi, 1994: 151);

di Bentivogli e Vecchi Galli:

«Il passaggio dal manoscritto al libro contemporaneo comporta ... un inevitabile ammodernamento dell'uso grafico ... Tali interventi, del resto, sono forse più semplici, ma non assenti del tutto, persino nei confronti di un'opera moderna o contemporanea (punteggiatura, accenti, apostrofi, grafemi ecc.). Si consideri poi che, fra queste operazioni, che sembrano quasi irrilevanti, alcune (come la scelta della punteggiatura) sono veicoli decisivi per l'interpretazione del testo» (Bentivogli e Vecchi Galli, 2002: 83).

Altrettanto vale per alcuni aspetti dell'*impaginazione*, che finiscono per essere strettamente legati all'ortografia: gli editori si sentono autorizzati ad intervenire con la sostituzione delle maiuscole con le minuscole all'inizio dei versi, l'uso di distanza doppia fra le righe per segnalare le sezioni, la numerazione dei versi, ecc.

Ciò che importa, per comprendere le considerazioni che seguiranno, è prendere come punto di partenza il fatto che il procedimento a stampa ha introdotto in questo campo (come p. es. in quello dell'ortografia o in quello della separazione delle parole) elementi di normalizzazione, di prescrittività, di rigidità, prima sconosciuti. Tali elementi sono entrati col tempo a far parte dell'ovvia coscienza e competenza dei lettori (oltre che degli autori), i quali si sono perciò abituati a distinguere implicitamente, senza soffermarvisi più che tanto, fra elementi grafici con valenze, diciamo, sopra-segmentali, e dunque significativi, e dunque sottoposti a norme precise di mantenimento/conservazione nelle riproduzioni, ed elementi grafici puramente, diciamo, esornativi, e dunque sottoposti al puro giudizio estetico del compositore attuale della pagina.

In altre parole, è invalsa una prassi consolidata, secondo la quale, di un testo da riprodurre, cioè di cui prodorre una nuova edizione, è implicitamente inteso quali elementi vadano conservati con esattezza, e quali possono essere cambiati. Non c'è su questo (a quanto mi consta) una teoria esplicita (salvo gli accenni riportati sopra), ma una prassi relativamente consolidata. Su di essa interviene in maniera dirompente, tanto più in quanto ancora non apprezzata nella sua vera sostanza, l'informatica, con procedure che vanno ben oltre l'aspetto cosiddetto ipertestuale, che attualmente gode di un'attenzione privilegiata e di una fiducia mal riposta.

L'edizione cartacea non ignora, ma sostanzialmente rende implicito il problema, offrendo una riproduzione del testo di tipo analogico, che include implicitamente quella informazione, anche sul carattere esteriore del testo, che l'editore ritiene opportuno mantenere. Si noti che in questo momento non ci interessano i problemi relativi alla fedeltà convenzionale al testo, cioè alla liceità data all'editore di intervenire con correzioni per restituire una lezione originaria o autentica rispetto ad una corrotta. Noi ci riferiamo propriamente ad una situazione che precede questi problemi, cioè all'informazione contenuta nel documento in quanto tale, di là dal cosiddetto testo che esso contiene. È peraltro vero che i problemi sono connessi, e vale la pena di chiarire il loro rapporto, che non sempre è bene inteso.

La necessità di intervenire, per l'editore tradizionale, deriva dalla commistione in uno stesso prodotto (il prodotto a stampa) di due funzioni diverse, in parte contraddittorie: quella di restituzione di un testo autentico, e quella di fruizione di tale testo da parte dei lettori; in termini informatici, la conservazione dell'informazione contenuta nel documento, e la sua trasmissione ad uno specifico destinatario. Connessa con ambedue le funzioni è la costrizione che impongono le possibilità tecnologiche della stampa. Occorre comunque liberarsi subito di un falso problema: la funzione storico-riproduttiva non può essere risolta neppure in ambito informatico fornendo una riproduzione puramente iconografica, per immagine, del testo che si suppone autentico, sia esso manoscritto o a stampa. Infatti il lavoro editoriale si basa sempre su scelte, che non possono essere demandate al lettore. Inoltre, per quanto riguarda l'ambiente informatico, essa non permette una analisi automatizzata (linguistica, letteraria, storica) del testo stesso.

Prima di illustrare quale dovrebbe essere la prassi corretta per l'ambiente informatico, sembra utile esaminare la situazione quale si presenta attualmente nel caso che abbiamo scelto come esempio, ribadendo comunque che lo riteniamo rappresentativo di quanto accade anche per gli altri testi della letteratura italiana, per i testi classici, e per quelli in altre lingue moderne.

Osserveremo prima di tutto che nessuna delle edizioni elettroniche è basata sull'edizione Bettoni. Per questo motivo il testo che esse offrono è già stato filtrato attraverso edizioni moderne, con interventi su ortografia e punteggiatura, oltre che sull'impaginazione, a cui i responsabili delle edizioni elettroniche hanno poi aggiunto specifiche brutture (caratteri improbabili, impaginazione sconcertante, ecc.) e specifici errori. Tralasciando le

questioni relative al testo comunemente inteso, concentreremo l'attenzione su caratteristiche considerate esterne, che invece contengono informazione direttamente e concettualmente legata al testo per la cui descrizione ci siamo basati su quelli che consideriamo i capostipiti dei due rami, Einaudi e Liz. L'informazione testuale, come è rappresentata nelle edizioni elettroniche, dipende soprattutto da tre fattori: gli interventi rispetto al testo originale fatti nell'edizione a stampa che è servita di modello per il testo digitale; la disponibilità (o creduta tale) di caratteri nella tabella di codifica secondo la quale è stato prodotto il testo digitale; le caratteristiche dei filtri di impaginazione che sono stati usati.

### **Ramo Einaudi**

Elimina intitolazione, epigrafe, e dedica

Elimina particolarità grafiche

Elimina maiuscole inizio verso

Elimina le indentature

Introduce numerazione delle linee

Separa le sezioni con spazio interlineare invece che con indentatura (le versioni derivate hanno generalmente corretto questo punto)

Inserzione di una lineetta all'inizio e alla fine del discorso diretto

Modifica del punto e virgola in due punti, 17, 164, 251

Eliminazione (o dimenticanza) della virgola, 58, 83, 146, 234

Aggiunta di una virgola, 238, 252, 267, 277

Modifica dell'accento grave in acuto: piú, né, dí

Accento grave invece di acuto in egèe, retèe, Pimplèe

Modifica della j in i

Maiuscola dei nomi del tipo Greca, Nume, Luna, Prenci, ecc. modificata in minuscola

Casi particolari: obblío per obblío, Troade per Tróade, diè per die'

### **Liz**

Elimina epigrafe e dedica

Elimina particolarità grafiche

Introduce numerazione delle linee

i invece di ì: obbliate, religion, effigiati, preziosi; ma: rapian, sentia, nutria, fuggia (per fuggía)

Eliminazione (o dimenticanza) della virgola:

### Casi interessanti

- 60 abdùani: Liz abduani; Ein etc. abdüani  
 83 funerea: Liz funerêa; Ein etc. funerëa  
 84 luttùoso: Liz luttuoso; Ein etc. luttüoso  
 157 allôr: Liz allor; Ein etc. allòr  
 273 Príamo: Liz Priamo; Ein etc. Priamo  
 291 Oceáno: Liz Oceàno; Ein etc. Oceàno

Quest'ultimo caso merita forse una riflessione. L'accento acuto su á è contrario alla regola del Foscolo di porre come tonico l'accento grave; potrebbe dunque essere interpretato come dieresi, dal momento che non sembra che Bettoni disponesse del carattere a con dieresi, e tale dieresi sarebbe correttamente posta sulla vocale *non accentata*. Secondo questa interpretazione l'accento di Oceano sarebbe quello usuale, e l'endecasillabo sarebbe ricostruito ammettendo una cesura fra padre e Oceano: abbraccia terre il gran padre | Ocèano.

Aggiungeremo che si dovrebbero notare i differenti tipi di spazio fra le parole, secondo quanto abbiamo notato sopra; il differente valore dei cosiddetti accenti: tonico, per la dieresi (v. abduani, Oceano, ecc.), ecc.; i salti pagina; i rapporti di grandezza dei caratteri, e forse la loro estetica.

Queste le discrasie nella trasposizione digitale del testo, che sono, per quanto posso vedere, e *mutatis mutandis*, comuni a quasi tutti gli altri testi digitali, in particolar modo quelli di carattere letterario. Ma è importante comprendere che esse non dipendono da trascuratezze individuali e soggettive, ma proprio da una errata visione metodica del procedimento digitale stesso, sulla quale vale la pena di soffermarsi. La prima causa di errore deriva dal perverso criterio per cui chi produce un testo elettronico (termine che in realtà nasconde vari tipi di testualità) concentra l'attenzione sul prodotto visibile del procedimento, mentre quella che si può definire la *base digitale* che consente di ottenerlo viene considerata esclusivamente un mezzo, che viene giudicato appunto esclusivamente in base a quello che si ottiene, si vede, sullo schermo o viene stampato dalle stampanti.

Ora, come in altre cose di questo mondo, nell'informatica c'è qualcosa che si vede e qualcosa che non si vede, che pure sta alla base di ciò che si vede. Si badi che «non si vede» è da intendere qui sotto due aspetti: da un lato si tratta di *realia* microscopici, che sfuggono all'occhio; ma dall'altro si tratta di astrazioni. È normale che chi viene in contatto con l'informatica senza un'adeguata preparazione, si lasci impressionare da ciò che si vede, e

trascuri ciò che non si vede, ritenendo che questa parte vada lasciata alle cure e all'attenzione dei soli professionisti dell'informatica. C'è molta diffidenza in questo, da parte degli studiosi umanistici, e inoltre la convinzione che siano cose troppo complicate, e difformi dai propri interessi. Non solo questo non è vero, ma purtroppo gli specialisti informatici non hanno sempre e del tutto la volontà e la capacità di piegare ciò che non si vede alle esigenze degli utenti. Non sempre, in quanto questo compito li può costringere a mutare procedimenti abitudinari che essi non intendono mutare (p. es. l'uso di certi programmi di scrittura); e non del tutto, in quanto essi non sono alieni dal temere (peraltro a ragione) di perdere il dominio sulla progettazione delle procedure.

La situazione è dunque assai complicata, ma fino a quando non sarà compresa chiaramente da parte degli editori di testi elettronici, è vano sperare che possano essere progettate e prodotte edizioni decenti. Occorre prima di tutto aver presenti due concetti fondamentali: (1) l'informatica intesa in senso stretto, indipendentemente dalle sue applicazioni, non impone altre restrizioni se non quella di pensare e di agire (concepire gli strumenti applicativi) in base ad elementi discreti di informazione. Una volta individuati tali elementi, essi potranno *sempre* essere rappresentati mediante simboli informatici, cioè trattabili da parte di un calcolatore, ed inoltre essere sottoposti a procedimenti algoritmici. Poiché un algoritmo è in sostanza la risoluzione di un problema, questo significa che se gli elementi discreti costituiscono i dati di un problema, tale problema può essere risolto dalla macchina. A scanso di equivoci, chiariremo che la soluzione deve essere concepita dall'uomo (algoritmo), ma il procedimento può essere eseguito dalla macchina, cioè automatizzato.

(2) In secondo luogo, occorre comprendere che la comunicazione uomo-macchina (input) è momento anche concettualmente separato da quello della comunicazione macchina-uomo (output). Quello che intendiamo sottolineare è che gli strumenti di input che sono stati via via escogitati e prodotti, dalle perforatrici di schede fino alle nostre attuali tastiere (e ciò vale anche per gli OCR e i sintetizzatori vocali, ma il discorso sarebbe più complesso) hanno di per sé il solo scopo di rendere più facile la gestione in input delle sequenze di bit che in ultima analisi sono i simboli che dobbiamo utilizzare per rappresentare gli elementi discreti d'informazione di cui abbiamo parlato sopra. L'aspetto concreto, fisico, visibile (o udibile) che i simboli assumeranno in uscita (output) può essere totalmente indipendente dall'aspetto che essi hanno sulla tastiera o anche, al momento dell'input, sullo schermo. Il codice standard ASCII è stato inventato (come tanti altri prima) per facilitare la comu-



un prodotto, un testo. Ma quando si vogliono vedere anche le marcature, utilizzando un programma-filtro apposito, a sua volta quello che si vede non corrisponde a quanto sta sul supporto magnetico, bensì è una rappresentazione delle sequenze di bit che stanno sul supporto magnetico.

È dunque fondamentale la triplice distinzione: dati elettronici invisibili – rappresentazione/gestione ASCII dei dati elettronici – gestione dell’impaginatore degli stessi dati. La gestione ASCII è più diretta e fa meglio comprendere il «vero» (invisibile) contenuto dell’insieme dei dati di origine (il *file*); gli impaginatori sono fuorvianti. Ma anche il file ASCII rappresenta un secondo livello, mentre quello che conta veramente è il primo livello. Esso è un livello astratto, costituito da una serie di segni appartenenti ad un codice formato di un numero di segni grande a piacere. Questi segni (astratti!) corrispondono ad altrettanti elementi/fenomeni del documento da trasformare in documento digitale. È chiaro a questo punto che il limite al numero ed alla quantità dei fenomeni che si possono trasportare e rappresentare in ambiente digitale dipende *non* dall’apparecchiatura digitale che si usa o si userà, ma soltanto dalla capacità dell’editore di riconoscerli, enuclearli, e renderli espliciti, cioè nel renderli digitali, prima ancora che elettronici. È insomma una prerogativa degli studiosi umanistici, non degli informatici.

Soltanto a questo punto entra in gioco il codice ASCII, il cui uso permette di dare una corrispondenza simbolica digitale a ciascun fenomeno, in forma evidente per una persona umana, che allo stesso tempo sarà «compresa» anche dalla macchina. Si forma dunque un alfabeto capace di esprimere compiutamente il testo del documento sotto tutti i suoi aspetti. E conseguentemente entrano in gioco i linguaggi di marcatura, che, a loro volta, tramite segni espressi in codice ASCII, consentono di progettare, produrre, e gestire il testo del documento a vari livelli: di impaginazione, di significati linguistici, sintattici, stilistici, storici, ecc., di nuovo senza alcun limite, se non la volontà e la capacità dell’editore, e poi del linguista, del critico, dello storico, ecc.

Questo tipo di testo, nel quale sono posti in evidenza anche quegli elementi che lo compongono, ma sono altri rispetto ai segni alfa-numeriche e dia-critici utilizzati nella stampa, può essere concretizzato, visualizzato nel modo che abbiamo visto sopra, quando abbiamo mostrato un sorgente html (fig. 2). A questo punto entrano in gioco i browser o filtri, che devono interpretare i segni in modo da presentare il testo, diciamo così, propriamente detto, sullo schermo o sulla stampante, secondo la volontà di chi ha prodotto il sorgente,

cioè l'editore. E questo, cioè la produzione dei browser, è appunto e solo il compito degli ingegneri del software.

Abbiamo così delineato un modello ideale di procedura per rappresentare elettronicamente i testi letterari e poi ottenerne una rappresentazione su uno schermo (monitor) o attraverso una stampante, cioè quello che viene normalmente riassunto col termine, purtroppo ingannevole, di edizione elettronica. Ci siamo soffermati in particolare su edizioni che rappresentano testi a stampa, ma per quelle che rappresentano manoscritti il sistema non cambia. In effetti, delle imprese attualmente in corso, e di cui abbiamo notizia, e che più si avvicinano al modello che abbiamo proposto: l'edizione di Chaucer curata da Peter Robinson (1996), quella del *Chevalier de la Charrette* curata da Karl D. Uitti (1994-2003), quella dei *Renaissance Electronic Texts*, curata da Ian Lancashire (1997), e quella del «Durham Liber Vitae»<sup>3</sup>, tre su quattro sono dedicate a manoscritti, perché evidentemente in questo campo c'è maggiore consapevolezza filologica; ma nessuna si adegua pienamente ad esso. Così come i sistemi software che intendono rendere possibile, o meglio più praticamente agevole, la produzione di edizioni elettroniche, Kleio, Tustep, e Bambi (cfr. Thaller, 1991, 1993a, 1993b; Ott, 1987, 1992, 1993 e Bozzi, 1997), in parte riflettono quel modello, ma mai compiutamente.

La renitenza ad adeguarsi ad uno standard scientifico come quello che abbiamo delineato, sia dal punto di vista ecdotico, sia da quello informatico, si può ricondurre ad un certo numero di pregiudizi condivisi purtroppo largamente fra chi lavora in questo campo. Segneremo, per semplificare: la «dipendenza da software», cioè la convinzione che la possibilità di produrre un'edizione elettronica dipenda in primo luogo dalla disponibilità di un particolare software; l'attenzione esclusiva alla manifestazione finale, quella che viene proposta concretamente al lettore, trascurando i caratteri intrinseci del «sorgente» (cfr. sopra; questo oltretutto favorisce i ciarlatani); il principio «meglio un uovo oggi che una gallina domani», che raccomanda la quantità a scapito della qualità, e fra l'altro contribuisce a screditare i prodotti informatici.

Sotto questo atteggiamento falsamente pragmatico c'è indubbiamente anche un problema economico, ma credo sia un falso problema. In realtà si tratta di valutare qual è il sistema migliore per raggiungere il fine voluto, tenendo conto che tutti, almeno in teoria, concordano che questo consiste nella

<sup>3</sup> *The Durham Liber Vitae Project*, cf. <http://www.kcl.ac.uk/humanities/cch/dlv/index.html>

produzione di testi altrettanto autorevoli di quelli delle migliori edizioni a stampa (io dico, anche di più). Si possono sintetizzare due filosofie: quelle delle imprese centralizzate, con forti finanziamenti, che si dedicano a ingenti *corpora* omogenei (si potrebbe citare l'esempio in ambito tradizionale della celebre operazione dell'abate Migne e delle sue *Patrologiae*); e quella del lavoro distribuito, che necessita soltanto di un accordo su standard formali, che garantiscano la cosiddetta interoperabilità, cioè che i prodotti possano essere riutilizzati da più utenti in modi differenziati.

Io credo che l'unico modo per avere una grande quantità di testi in tempi ragionevoli, di buona affidabilità critica, e corretta qualità informatica, sia il secondo, che richiede la formulazione di standard formali e l'impegno da parte di singoli studiosi, operanti ciascuno per proprio conto, e non all'interno di cosiddette *équipes* di ricerca, ad adeguarvisi. Aggiungo che dovrebbe generalizzarsi una prassi, secondo la quale fra studiosi si mettono a disposizione i «sorgenti» testuali, mentre le rappresentazioni poco devono contare in ambito scientifico; e anche la valutazione dell'opera di uno studioso nell'ambito dell'informatica umanistica (testuale) deve essere fatta sui «sorgenti». È evidente quanto grande in questo tipo di filosofia possa essere il ruolo dei *trusted repositories* di cui si discute in questo convegno. Essi possono essere il bacino di raccolta delle singole produzioni di studiosi altamente qualificati e specializzati; e contemporaneamente uno stimolo per questi studiosi ad accettare gli standard proposti, che in quanto formali non impegnano ad alcun particolare modello di metodologia scientifica per quanto riguarda la costituzione di testi criticamente accettabili, secondo le buone tradizioni della filologia, indipendentemente dall'applicazione di sistemi informatici.

#### BIBLIOGRAFIA

- BENTIVOGLI B., VECCHI GALLI P., 2002. *Filologia italiana*. Mondadori, Milano.
- BOZZI A. (ed.), 1997. *Better Access to Manuscripts and Browsing of Images*. Clueb, Bologna.
- CONTINI G., 1986. *Breviario di ecdotica*. Einaudi, Torino (1992).
- LANCASHIRE I. (gen. ed.), 1997. *Renaissance Electronic Texts*. Web Development Group University of Toronto Library.  
<http://www.library.utoronto.ca/utel/ret/ret.html>
- MACRÌ O., 1978. *Semantica e metrica dei «Sepolcri» del Foscolo, con una teoria dell'endecasillabo*. Bulzoni, Roma.
- MALMBERG B., 1994. *Manuale di fonetica generale*. Il Mulino, Bologna.

- OTT W., 1987. *TUSTEP. Tübinger System von Textverarbeitungs-Programmen*. Zentrum für Datenverarbeitung der Universität, Tübingen.
- OTT W., 1992. *Computer and Textual Editing*. In: C.S. BUTLER (ed.), *Computers and Written Texts (Applied Language Studies)*. Blackwell, Oxford: 205-226.
- OTT W., 1993. *Edizione critica e gestione di testi. Il pacchetto TUSTEP* In: T. ORLANDI (ed.), *Discipline umanistiche e informatica. Il problema dell'integrazione* (Roma, 8 ottobre 1991). Contributi del Centro Linceo Interdisciplinare «B. Segre», 87, Accademia Nazionale dei Lincei, Roma: 17-28.  
<http://www.uni-tuebingen.de/zdv/tustep/index.html>
- ROBINSON P. (ed.), 1996. *Chaucer: The Wife of Bath's Prologue on CD-ROM*. Oxford Univ. Press.
- SEGRE C., 1985. *Avviamento all'analisi del testo letterario*. Einaudi, Torino.
- STUSSI A., 1994. *Introduzione agli studi di filologia italiana*. Il Mulino, Bologna.
- THALLER M., 1991. *The Historical Workstation Project*. *Computers and the Humanities*, 25: 149-162.
- THALLER M., 1993a. *Historical Information Science: Is There Such a Thing? New Comments on an Old Idea*. In: T. ORLANDI (ed.), *Discipline umanistiche e informatica. Il problema dell'integrazione* (Roma, 8 ottobre 1991). Contributi del Centro Linceo Interdisciplinare «B. Segre», 87, Accademia Nazionale dei Lincei, Roma: 51-86.
- THALLER M., 1993b. *KLEIO. A Database System*. Göttingen, St. Katharinen.
- UETTI K.D., 1994-2003. *The Charrette Project*. Princeton University.  
<http://www.princeton.edu/~lancelot/>

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 traggio 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



Fig. 1.

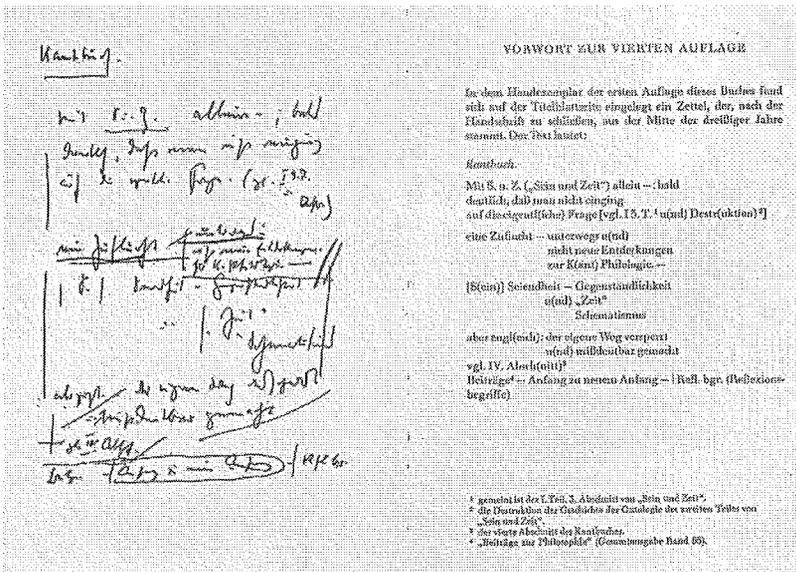


Fig. 2.

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'esaustività, 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).

<i>Numerazione</i>		<i>Notazione</i>
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 trascegliendone 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»<sup>1</sup>. 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-

<sup>1</sup> «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)<sup>2</sup> e la riproduzione del documento fisico originale (fig. 5)<sup>3</sup>.

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*<sup>4</sup>. 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.

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

<sup>3</sup> [http://portal.acm.org/ft\\_gateway.cfm?id=827192&type=pdf&coll=GUIDE&dl=GUIDE&CFID=14057334&CFTOKEN=5744073](http://portal.acm.org/ft_gateway.cfm?id=827192&type=pdf&coll=GUIDE&dl=GUIDE&CFID=14057334&CFTOKEN=5744073).

<sup>4</sup> Cfr. <http://www.ceec.uni-koeln.de>

The screenshot shows a web browser window displaying the ACM Portal. The page title is 'THE GUIDE TO COMPUTING LITERATURE'. The search results show an entry for 'XML semantics and digital libraries' by David Dubin, Allen Rensar, and Claus Hiltfeldt. The entry includes source information (International Conference on Digital Libraries), authors, sponsors (IEEE Technical Committee on Digital Libraries), and publisher (IEEE Computer Society). Below the entry, there is an abstract section starting with 'The lack of a standard formalism for expressing the semantics of an XML vocabulary is a major obstacle to the development of...'. A 'REFERENCES' section follows, listing several academic papers. At the bottom, there is a 'Collaborative Colleagues' section listing the names of the authors and other researchers in the field. The page footer contains the ACM logo and copyright information.

**XML semantics and digital libraries**

Full text [\(45 KB\)](#)

**Source** International Conference on Digital Libraries [getinfo](#)  
 Proceedings of the third ACM/IEEE-CS Joint conference on Digital Libraries [table of contents](#)  
 Houston, Texas  
 SESSION: Standards, mark-up, and metadata [table of contents](#)  
 Pages: 303 - 306  
 Year of Publication: 2003  
 ISBN: 0-7695-1938-3

**Authors** [Allen Rensar](#) University of Illinois at Urbana-Champaign  
[David Dubin](#) University of Illinois at Urbana-Champaign  
[C. M. Spensberg-McQueen](#) MIT Laboratory for Computer Science  
[Claus Hiltfeldt](#) Oregon University Research Foundation

**Sponsors** : IEEE Technical Committee on Digital Libraries (TC DL)  
 SIGCHI: ACM Special Interest Group on Information Retrieval  
 SIGUCCS: ACM Special Interest Group on Hypertext, Hypermedia, and Web

**Publisher** IEEE Computer Society Washington, DC, USA

**Additional Information:** [abstract](#) [references](#) [collaborative colleagues](#)

**Tools and Actions:** [Discussion](#) [Find similar Articles](#) [Review this Article](#)  
[Save this Article to a Binder](#) [Display in Bibtex Format](#)

↑ **ABSTRACT**

The lack of a standard formalism for expressing the semantics of an XML vocabulary is a major obstacle to the development of

[ ... ]

substantial loss of functionality and interoperability in digital libraries results from not having a common machine-readable formalism for expressing these relationships for the XML vocabularies currently being used to encode content. Recently a number of projects and standards have begun taking up related topics. We describe the problem and our own project.

↑ **REFERENCES**

Note: OCR errors may be found in this Reference List extracted from the full text article. ACM has opted to expose the complete list rather than only correct and linked references.

- 1 James H. Combs, Allen H. Rensar, Steven J. DeRosa, Marking systems and the future of scholarly text processing, *Communications of the ACM*, v.39, n.11, p.933-947, Nov. 1997.
- 2 D. Dubin, C. M. Spensberg-McQueen, A. Rensar, and C. Hiltfeldt, A logic programming environment for document semantics and inference, *Journal of Literary and Linguistic Computing*, Forthcoming in 2005.
- 3 D. R. Raymond and F. W. Tompa, Markup reconsidered. Technical Report 356, Department of Computer Science, The University of Western Ontario, 1995. An earlier version was circulated privately as "Markup Considered Harmful" in the late 1980s.

[ ... ]

13 C. Wally and N. Ide. Using the right tools: Enhancing retrieval from marked-up documents. *Computers and the Humanities*, 33(1-2):59-84, 1999. Originally delivered in 1997 at the TEI 10 conference in Providence, RI.

14 Mike Wuytsingha, Chaitoon Anutaniya, Nivetha Akema, Ekwin Nantaisawatwatt, XML Declarative Descriptions: A Language for the Semantic Web. *Intell. Intelligent Systems*, v.16, n.3, p.64-65, May 2004

↑ **Collaborative Colleagues:**

**David Dubin:** [Claus Hiltfeldt](#)  
[Josephson Liu](#)  
[Allen Rensar](#)  
[C. M. Spensberg-McQueen](#)

**Claus Hiltfeldt:** [David Dubin](#)  
[Allen Rensar](#)  
[C. M. Spensberg-McQueen](#)

**Allen Rensar:** [David Dubin](#)  
[Claus Hiltfeldt](#)  
[Eli Hiltfeldt](#)  
[Dorothea Sabo](#)  
[Rosemary Simpson](#)  
[C. M. Spensberg-McQueen](#)  
[Andreas van Dam](#)

**C. M. Spensberg-McQueen:** [Tom Bray](#)  
[Lou Bernard](#)  
[David Dubin](#)  
[Robert E. Colquhoun](#)  
[Claus Hiltfeldt](#)  
[Juan Pablo](#)  
[Allen Rensar](#)  
[B. Terence Liddy](#)  
[B. Terence Liddy](#)

The ACM Portal is published by the Association for Computing Machinery, Copyright © 2003 ACM, Inc.  
[Home](#) [Contact](#) [Privacy Policy](#) [Code of Ethics](#) [Feedback](#)

Useful downloads: [Adobe Acrobat](#) [QuickTime](#) [Windows Media Player](#) [RealPlayer](#)

Fig. 4.

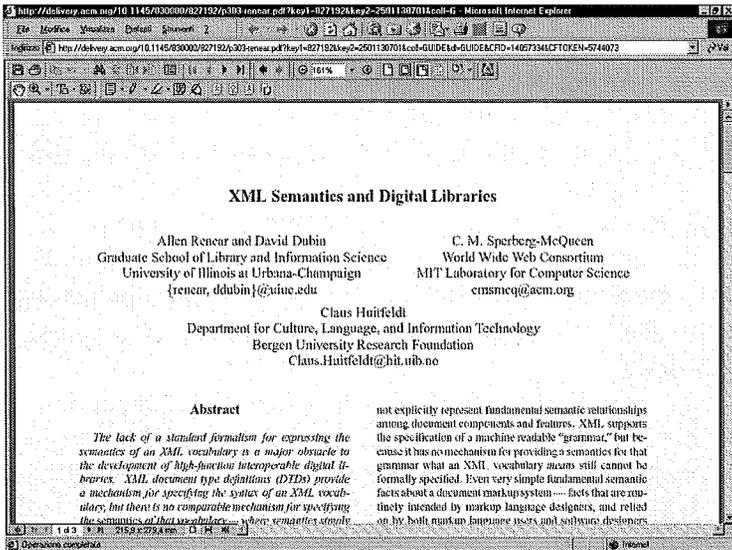


Fig. 5.

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'esattività 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'esattività, 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 autorizza<sup>5</sup> 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,

<sup>5</sup> «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 «fluido» 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 fluidi» ad un tempo. Ma l'opposizione non pare ridursi solo ad una questione di grado<sup>6</sup>, 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.

<sup>6</sup> «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 palesamente «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 mostrano<sup>7</sup>, oppure descrivono, la forma logica del testo, così come di volta in volta possono rappresentare un operatore funzionale, oppure il valore che ne dipende<sup>8</sup>. Intese come espressioni autoriflessive del linguaggio-oggetto, esse comportano forme di predicazione di ordine superiore.

<sup>7</sup> 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».

<sup>8</sup> 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'insieme 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 contemperamento 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 invariante 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.

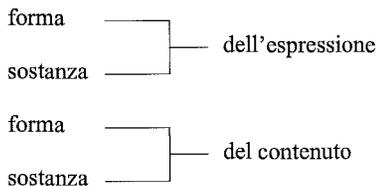
9. Certamente «un oggetto digitale» può essere definito, con Kenneth Thibodeau, come «un oggetto informativo, qualunque sia il tipo di informazione rappresentata o qualunque ne sia il formato, purché essa sia espressa in forma digitale» (Thibodeau, 2002: 6). Ma anche l'analisi che egli ne propone, ispirandosi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

<sup>9</sup> 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’esaustività 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*<sup>10</sup>, 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*<sup>11</sup>. 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 formalismo<sup>12</sup>, 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

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

<sup>11</sup> 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.

<sup>12</sup> 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.*, 2001b), 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.*, 2001a: 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.*, 2001a: 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 semistrutturato» (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» (Reneau *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 parti<sup>13</sup>. 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

<sup>13</sup> 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 XML<sup>14</sup>; 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 caratteri<sup>15</sup>. 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 letterari<sup>16</sup>. 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

<sup>14</sup> 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.

<sup>15</sup> 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».

<sup>16</sup> «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](http://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/diplslamm/Pubblicazioni-telematiche/dino\\_buz-zetti.htm](http://dobc.unipv.it/diplslamm/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., DEROSE 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.
- DEROSE 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>
- FOLDOC, 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, 1961<sup>2</sup>). Trad. it., Einaudi, Torino.
- HOFMAN H., 2002a. *Some comments on preservation metadata and the OAIIS model*. DigiCULT. Info, Newsletter, 2: 15-20.  
[http://www.digicult.info/downloads/digicult\\_info2.pdf](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](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](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/cstr/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/lagoze-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](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](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](http://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 Quædarni 1914-1916* (trad. it. di A.G. Conte). Einaudi, Torino.



PAOLA MOSCATI\*

## LINGUAGGI DI MARCATURA PER LA CONSERVAZIONE E LA VALORIZZAZIONE DELL'INFORMAZIONE ARCHEOLOGICA

### *1. Introduzione.*

Nel novembre del 2000 ho avuto occasione di illustrare, in questa stessa sede prestigiosa, alcuni aspetti del Progetto Caere, dedicato alla realizzazione di un sistema informativo territoriale per l'acquisizione, la gestione e l'elaborazione informatizzate delle testimonianze archeologiche relative alla città etrusca di Cerveteri provenienti dall'attività di ricerca sul campo, dalle ricognizioni di superficie, alle prospezioni geofisiche e allo scavo (Moscati, 2003). Trattandosi allora di un Convegno dedicato al ruolo dei modelli nella ricerca archeologica, l'attenzione era stata posta soprattutto sulle fasi di rappresentazione e interpretazione delle informazioni, quali momenti strettamente connessi del processo di ricerca e ricostruzione storica del passato.

Tornando oggi sull'argomento, grazie al cortese invito di Tito Orlandi, in un contesto diverso e specificamente indirizzato alle problematiche connesse con l'archiviazione digitale, è mia intenzione, dopo aver brevemente riassunto per motivi di chiarezza gli intenti generali del Progetto Caere, soffermarmi sulla codifica delle informazioni archeologiche mediante l'uso di linguaggi di marcatura e soprattutto sulle prospettive di ricerca connesse con la conservazione e la fruizione di tali informazioni.

### *2. Il Progetto Caere: dal territorio allo scavo.*

Il Progetto Caere ha preso avvio nel 1997 in seno al più ampio e ambizioso Progetto Finalizzato Beni Culturali del CNR (<http://www.progettocaere.rm.cnr.it/>, con bibliografia relativa; per una sintesi Moscati, 2002). Promosso

\* CNR - Istituto di Studi sulle Civiltà Italiane e del Mediterraneo Antico (ISCIMA) - Viale di Villa Massimo, 29 - 00161 ROMA. [paola.moscati@iscima.cnr.it](mailto:paola.moscati@iscima.cnr.it)

da Mauro Cristofani poco prima della sua prematura scomparsa, esso ha previsto un'intensa e diversificata attività di ricerca, resa possibile dal concorso di specifiche competenze. Il percorso adottato per lo sviluppo del Progetto ha seguito un criterio estremamente semplice, ma assai efficace: dal generale al particolare, dal contesto all'informazione e da quest'ultima al dato atomico.

La rinnovata base cartografica digitale del pianoro urbano ha costituito il punto di partenza per la georeferenziazione delle testimonianze archeologiche e per la realizzazione di modelli digitali del terreno, quali tecniche di visualizzazione scientifica delle caratteristiche geomorfologiche e altimetriche del pianoro e spesso fonti di indicazioni non deducibili dalla comune osservazione (Ceccarelli, 2001). L'implementazione di una piattaforma GIS (fig. 1) ha consentito quindi l'integrazione dei dati spaziali con quelli alfanumerici, provenienti per la massima parte dalle ricognizioni e dalle campagne di scavo realizzate dall'Istituto per l'archeologia etrusco-italica (oggi Istituto di studi sulle civiltà italiche e del Mediterraneo antico) negli anni Ottanta nell'area centrale del pianoro urbano, denominata Vigna Parrocchiale (*Caere I*, 3.1, 3.2, 4), e riprese nel 2003 sotto la direzione di Francesco Roncalli.

La documentazione di quest'area nodale per lo sviluppo della città etrusca risultava assai ricca e dettagliata e consisteva sostanzialmente in immagini, rilievi e rapporti di scavo, insieme a schede di catalogazione dei reperti. Per l'archivio grafico e fotografico e per le schede di catalogo, la scelta di inserire le informazioni all'interno di un database relazionale si è dimostrata assai proficua e non ha creato alcuna difficoltà di integrazione all'interno della piattaforma GIS. Invece, come abbiamo già avuto occasione di esporre in altre sedi nelle prime fasi del progetto (Moscati *et al.*, 1999; Moscati, 2000), di fronte alla ricchezza e alla complessità delle informazioni contenute nei diari di scavo manoscritti, sono sorte alcune perplessità riguardo alla procedura da adottare. La scelta di un database, infatti, avrebbe richiesto una strutturazione dei dati piuttosto rigida, soprattutto per quanto concerne le scelte terminologiche, in genere sottoposte a una selezione a priori, e avrebbe posto alcune difficoltà per la consultazione integrale e trasparente dei dati una volta immessi nella rete, per soddisfare un'altra delle finalità previste dal progetto.

Si è pertanto deciso di sperimentare una metodologia allora nuova per il settore dell'informatizzazione dei dati archeologici, e cioè l'uso di un linguaggio di marcatura quale SGML prima e XML poi, per la codifica e la conseguente formalizzazione delle informazioni contenute nei diari di scavo, fonte di riferimento essenziale, oltre che per le annotazioni di carattere ope-

Georeferenziazione  
dei dati

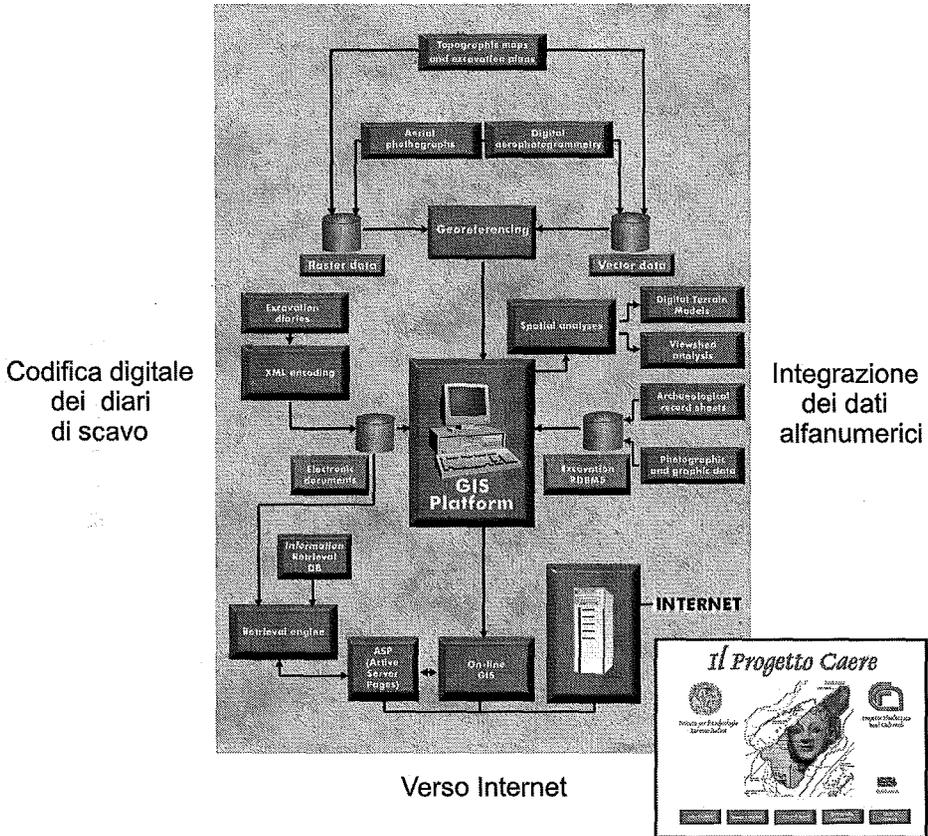


Fig. 1. – L'integrazione dei dati all'interno del Progetto Caere.

rativo, per la ricostruzione dei dati topografici, stratigrafici e materiali rilevati nel corso delle campagne (per un'altra recente esperienza cfr. Meckseper e Warwick, 2003). Il modello del testo, riassunto nella DTD appositamente definita, ha consentito da un lato di descrivere la struttura logica dei documenti e dall'altro lato di evidenziare gli elementi costituenti. Al di là degli esiti positivi raggiunti attraverso la sistematizzazione della documentazione, di cui si stanno già raccogliendo i frutti nella ripresa degli scavi (fig. 2), è interessante evidenziare che il modello messo a punto, a partire dalla sua semplicità

iniziale, si è dimostrato passibile di ulteriori ampliamenti, quasi una sorta di esplosione dei diversi elementi selezionati che ha consentito una più dettagliata descrizione dei dati stratigrafici, delle strutture e soprattutto dei reperti (Bonincontro, 2001).

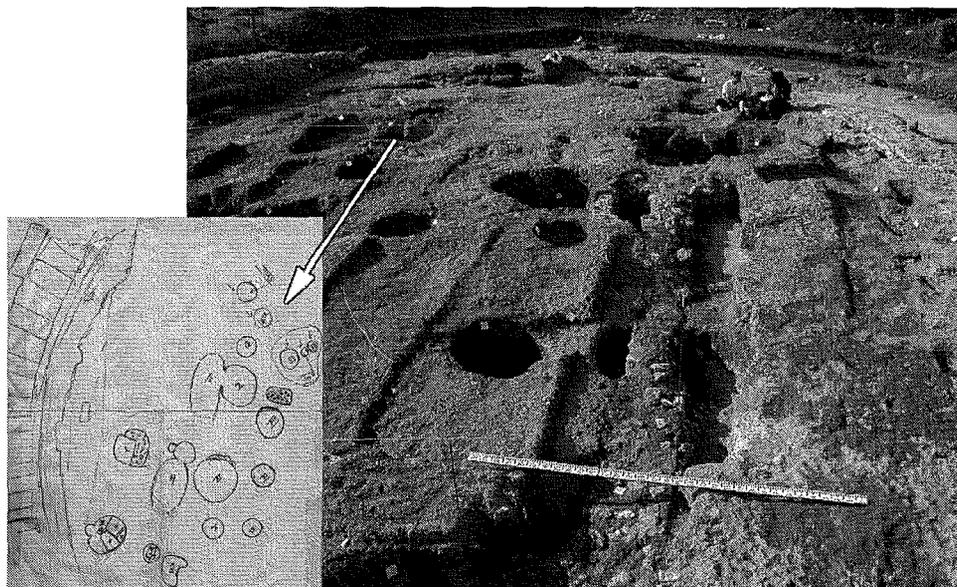


Fig. 2. – Cerveteri (Vigna Parrocchiale). La ripresa degli scavi e il confronto con i dati relativi alle campagne degli anni Ottanta.

### 3. La marcatura di rapporti di scavo editi: il ricorso alla TEI Lite.

Una volta definito questo modello innovativo di informatizzazione dei dati di scavo, si è subito pensato di adottarlo per la descrizione formalizzata di documenti archeologici di carattere diverso. È stato quindi elaborato un sistema di registrazione e di interrogazione dei dati provenienti da rapporti editi, attraverso l'integrazione della DTD realizzata per la marcatura dei diari di scavo con lo schema di codifica della TEI Lite (Mariotti, 2001). Le pubblicazioni prescelte, rispettivamente editate nelle riviste *Studi Etruschi* (Mengarelli, 1936; fig. 3) e *Notizie degli Scavi di Antichità* (Mengarelli, 1937), con-

cernono sia interventi di scavo condotti agli inizi del Novecento da Raniero Mengarelli, sempre a Cerveteri, nell'area della Vigna Parrocchiale, sia reperti iscritti provenienti da contesti urbani e funerari ceretani.

Nella naturale evoluzione del progetto, l'adozione della TEI Lite quale modello standard di marcatura di documenti di carattere archeologico, e soprattutto la possibilità di inglobare i dati marcati all'interno di un Sistema Informativo Territoriale che ruota intorno a una piattaforma GIS, hanno suggerito di approfondire la possibilità di realizzare un modello informatico per il trattamento di dati archeologici testuali non strutturati.



Fig. 3. – La versione elettronica del testo di R. Mengarelli (*Studi Etruschi*, X, 1936).

#### 4. Il ritorno al territorio.

Il progetto si è così indirizzato verso l'elaborazione elettronica di documenti di archivio dell'Ottocento, contenenti itinerari archeologici nel territorio circostante Roma, che ha subito in tempi recenti notevoli trasformazioni ur-

banistiche. Per ulteriori ampliamenti applicativi, grazie anche al materiale documentario che Lucos Cozza sta donando alla Biblioteca Massimo Pallottino che ha sede presso l'Istituto di studi sulle civiltà italiche e del Mediterraneo antico, ci si va orientando verso la codifica digitale della documentazione testuale prodotta per la Carta Archeologica d'Italia nel corso delle indagini condotte soprattutto nell'Etruria tiberina e nell'Agro falisco (Gamurrini *et al.*, 1972; Castagnoli, 1978).

L'esemplificazione che qui si mostra è stata realizzata sul manoscritto di Ercole Nardi dal titolo *Ruderi delle Ville Romano-Sabine nei dintorni di Poggio Mirteto* (Barchesi *et al.*, 2003). L'area presa in esame (fig. 4) rientra nella porzione del territorio della Sabina tiberina in cui l'Istituto, sotto la responsabilità scientifica di Paola Santoro, sta conducendo, d'intesa con l'Università degli Studi di Verona e la British School at Rome, un progetto di ricognizione di superficie denominato Progetto Galantina, al fine di delineare la storia dell'occupazione e dell'uso del territorio nelle diverse fasi crono-

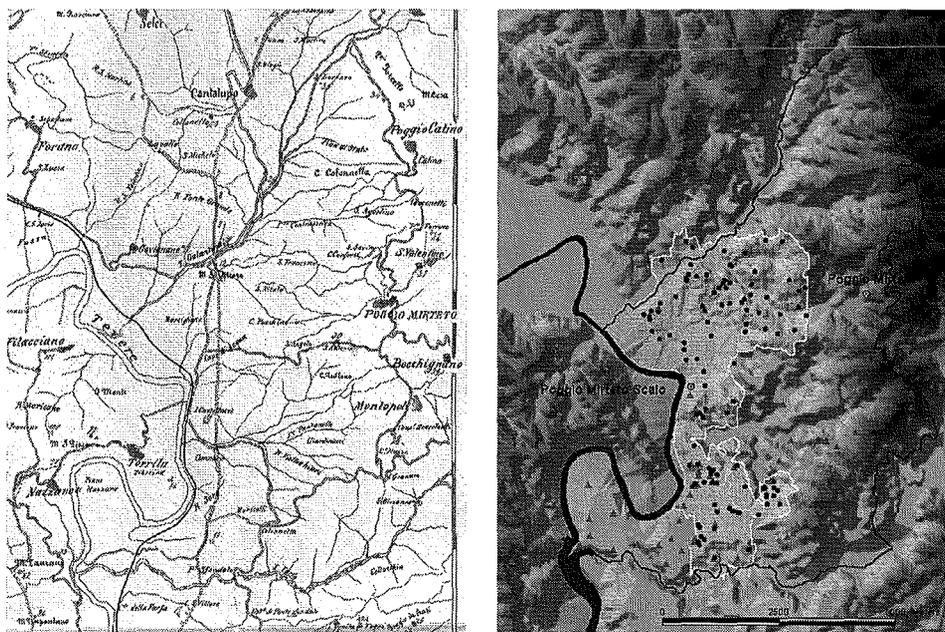


Fig. 4. – Sabina tiberina: area delle ricognizioni per la Carta Archeologica d'Italia e per il Progetto Galantina.

giche, dal paleolitico all'alto medioevo (Guidi e Santoro, 2003; Santoro *et al.*, 2005).

Ercole Nardi, originario di Poggio Mirteto e Ispettore Onorario della Sovrintendenza ai Beni Archeologici, si dedicò in età avanzata alla documentazione dei resti archeologici dell'area comunale per salvaguardare la memoria minacciata dal tempo o meglio, come dice il Nardi stesso, «per conservare ad ogni costo quei benché meschini ruderi, agli amatori, alla storia ed al decoro della mia terra Sabina» (P. Santoro, in Barchesi *et al.*, 2003: 298). Il manoscritto, ultimato nel 1885, è conservato presso la Biblioteca di Archeologia e Storia dell'Arte di Palazzo Venezia in Roma e consta di 22 fascicoli, ognuno dei quali descrive meticolosamente, e con un ricco apparato grafico e fotografico, i resti di una villa rustica romano-sabina (Scarpati, in corso di stampa).

I luoghi visitati, che sono inseriti in tre percorsi, o itinerari principali, risultano accuratamente descritti attraverso un'introduzione generale (vie di accesso, storia del territorio, riferimenti cartografici e toponomastici) e un'analisi più dettagliata delle evidenze archeologiche che parte dalla descrizione delle strutture nel loro aspetto generale per giungere alle singole parti costituenti. La marcatura, che ha finora riguardato il testo del Primo Itinerario in cui, tra i principali monumenti descritti, si trovano i «Bagni di Lucilla» e la «Villa dei Casoni» o «Villa di Varrone» (fig. 5), ha lo scopo di definire da un lato un primo inquadramento del sito, attraverso indicazioni relative all'ubicazione, ai toponimi, alle vie d'accesso, ai proprietari, e dall'altro lato di legare le singole strutture al contesto storico-geografico di appartenenza.

### *5. I metadati e la fruizione dei dati in rete.*

Per la codifica elettronica del manoscritto del Nardi si è partiti dal presupposto che, per esprimere in modo strutturato le caratteristiche relative al documento, è necessario far ricorso ai metadati, cioè a un insieme molteplice di informazioni da fornire per una corretta documentazione del testo elettronico (indicazione della fonte, esplicitazione del modello adottato nella rappresentazione della sua struttura, specificazione delle modalità di trascrizione e della natura degli interventi di codifica adottati). Così, spinti dalla volontà di promuovere un innovativo approccio integrato alle informazioni e di confrontarsi con le tendenze emerse nel recente dibattito sul «semantic web» (Signore, 2001, 2002; AA.VV., 2003), si è scelto di inserire gli elementi costituenti l'in-

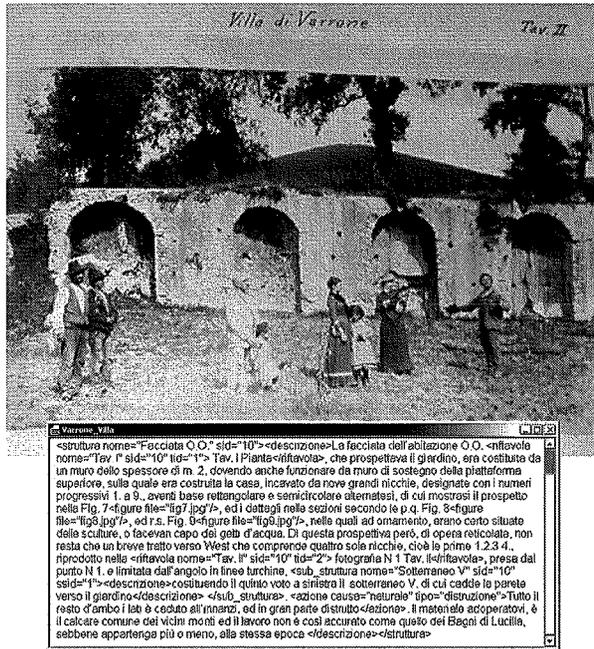


Fig. 5. – “Villa di Varrone”: marcatura del testo descrittivo di Ercole Nardi.

sieme Dublin Core espressi in RDF all'interno del documento marcato TEI Lite, e di aggiungere, sulla scia dell'esperienza realizzata nel caso ceretano, gli elementi della nostra DTD specificamente dedicati alla marcatura di testi di carattere storico-archeologico (C. Barchesi, in Barchesi *et al.*, 2003: 309-325; fig. 6). L'integrazione tra la complessa struttura TEI Lite e un ulteriore livello di formalizzazione ha avuto lo scopo di trasformare il testo narrativo in una fonte di informazioni strutturate, fruibili da applicazioni diverse, non solo orientate alla visualizzazione sullo schermo, ma accessibili anche attraverso software cartografici e database.

Le caratteristiche tecniche del progetto, infatti, sono sicuramente rese più complesse dalla volontà sia di integrare i dati testuali all'interno della piattaforma GIS, per inserire e interpretare le testimonianze archeologiche all'interno del loro contesto geografico, sia di diffondere le informazioni in rete facilitandone lo scambio anche fra studiosi che utilizzano sistemi informatici diversi. Nella prassi comune della metodologia GIS viene attribuito un codice agli oggetti vettoriali presenti sulla cartografia, allo scopo di creare un colle-

```

<TEI.2>
<teiHeader></teiHeader>
<text>
<front></front>
<body>
<div1 type=sito n=Bagni di Lucilla>
<metadati geografici>...</metadati geografici>
<introduzione>... Introduzione che precede la narrazione analitica...</introduzione>
<area id=1 nome=Parte esterna della villa>
  <struttura id=1 nome=Mura di cinta A A' A" A">
    <descrizione coordx=... coordy=... coordz=...>
      <substruttura id=1 nome=Spezzatura mura in reticolato>
        <descrizione>
          <unità id=1> Tratto A...</unità>
          <unità id=2> Tratto B...</unità>
          ...
        </descrizione>
      </substruttura>
      ...
      <substruttura id=n nome=Mura di sostegno B e f'>
        ...
      </substruttura>
    </descrizione>
  </struttura>
  <struttura id=2>...</struttura>
  <struttura id=3>...</struttura>
  ...
  <struttura id=n>...</struttura>
</area>
<area id=2 nome=Ambienti ipogei>
...
</area>
</div1>
...
<divn>...</divn>
</body>
<back></back>
</text>
</TEI.2>

```

Fig. 6. – Struttura della semantica archeologica inserita nella TEI Lite.

gamento fra le mappe e i dati alfanumerici inseriti all'interno di un database; nel nostro caso, invece, sono stati collegati i dati testuali agli oggetti grafici rappresentati da un lato sulla planimetria digitalizzata dello scavo (fig. 7) e dall'altro lato sulla cartografia numerica elaborata nell'ambito del Progetto Galantina.

Per quanto attiene alle problematiche della trasformazione e fruizione dei dati in rete, si è operato offrendo la possibilità di personalizzare la presentazione delle informazioni. Per i diari di scavo prima, e per il testo del Nardi poi, è stato utilizzato un approccio ipertestuale come strumento di interrogazione, attraverso una collezione di pagine ASP, che permettono all'utente di effettuare ricerche e recuperare i dati alfanumerici, grafici e iconografici presenti nel file marcato e quindi visualizzare i risultati in un browser (Barchesi,

2001). Inoltre la codifica XML consente, tramite l'applicazione dei fogli di stile XSL, di creare molteplici forme di rappresentazione dei documenti, nella forma più adeguata alla fruizione.

I testi, infatti, possono essere facilmente composti nella veste grafica tradizionale, adatta a un lettore generico, ma al contempo possono essere presentati in forme diverse, che pongono in luce gli elementi della semantica archeologica e le relazioni tra loro esistenti (fig. 8); il documento diviene dunque una sorgente di informazioni correlate e al contempo di nuovi dati che possono interagire sia con archivi digitali relazionali sia con sistemi GIS, per giungere a un sistema di conoscenza, in cui si fondono e si completano l'indagine passata e quella recente.

The screenshot displays a web application interface titled "Progetto Casere" with the subtitle "Diari di scavo della Vigna Parrocchiale". A warning message states: "Attenzione: Per l'interrogazione on-line di utenti 'visitatori' è accessibile un set di dati dimostrativo relativo al piano del 1983 dell'area dello scavo della 'Cisterna'".

The interface is divided into several sections:

- 1. Selezionare i diari:** A list of years (1983, 1984, 1985, 1986) with a dropdown menu currently set to 1983. Buttons include "Leggi il testo->" and "Interroga diario". A note below reads: "Questo set di dati contiene solo il diario del 1983".
- 2. Selezionare un settore?** Radio buttons for "No" and "Sì".
- 3. Selezionare un'area:** A dropdown menu set to "Cisterna" and a "Seleziona" button. A note below reads: "Per questo set di dati è interrogabile solo l'area della Cisterna".
- 4. Ricerche possibili in:**
  - Area = "Cisterna"
  - Settore = "E7IV - ✕"
  - Strutture: "Blocchi di tufo" (dropdown)
- 5. Risultati ricerca:**
  - n.1 Data: 20 giugno 1983 7 operai --- Area: [Cisterna]---Nome Settore: E7IV ✕
  - Struttura= [blocchi di tufo]

The main content area shows a window titled "Pianta dello scavo della Vigna Parrocchiale - Cerveteri -". It contains a planimetric map of the excavation site with a grid overlay. The map is labeled "Settori attivi" and shows various structural elements and their coordinates.

Fig. 7. – L'interrogazione in rete dei diari di scavo: collegamento tra dati planimetrici e contenuto testuale.

## 6. Integrità e integrazione.

Al di là degli aspetti più spiccatamente tecnici, è interessante soffermarsi specificamente sugli aspetti connessi con l'integrità dell'informazione archeo-

logica. L'obiettivo perseguito, infatti, non era solo quello di utilizzare un supporto digitale per preservare documenti cartacei deperibili o utilizzare uno standard per catalogarli, quanto piuttosto analizzare il contenuto dei testi per passare dalle fasi di acquisizione e rappresentazione dei dati a quelle di elaborazione e interpretazione: attenzione quindi non solo all'integrità «fisica» del bene, nel senso del suo recupero su supporto elettronico, ma anche e soprattutto a quella dei dati in esso contenuti. Ciò ha comportato di non limitarsi solamente alla preservazione e alla comunicazione delle informazioni, ma in

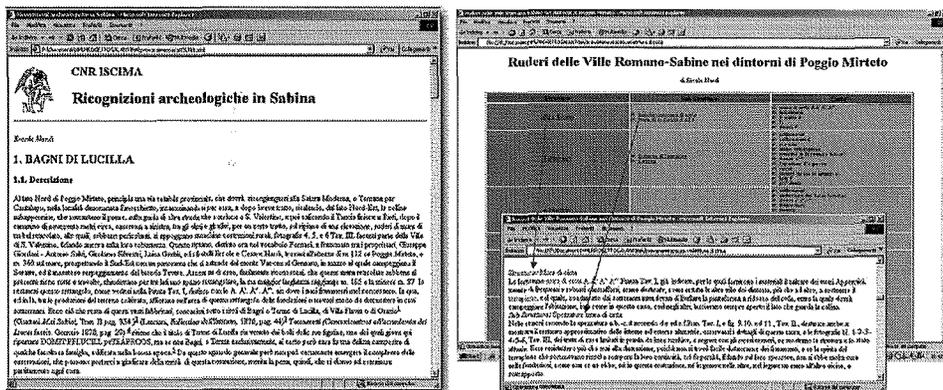


Fig. 8. – Diverse possibilità di resa del testo del Nardi in Internet.

qualche modo anche alla loro valorizzazione. Infatti, durante l'attività di ricerca e la realizzazione delle fasi successive del progetto, l'approccio adottato ha esaltato alcuni aspetti di carattere metodologico relativi all'analisi delle informazioni.

La selezione e la marcatura di singoli elementi presenti nel testo ha comportato un'attenta e accurata lettura dei documenti. Se questa centralità della fase di codifica del testo acquista negli studi letterari e linguistici una valenza del tutto particolare (Adamo, 1987 e da ultimo Orlandi, 1999; Gigliozzi, 2003), in ambito archeologico essa consente di concentrare l'attenzione su quelli che Gardin chiamerebbe «les faits», cioè i dati di base da cui ha origine ogni successivo passaggio interpretativo o, parimenti, nella sua concezione spiraliforme dell'acquisizione della conoscenza, punto di arrivo, e insieme di nuova partenza, di un approccio di tipo ipotetico-deduttivo (cfr. da ultimo Gardin, 2002, 2003, con bibliografia precedente). È ovvio che questi

dati hanno assunto fisionomie diverse nell'analisi dei rapporti di scavo o degli itinerari archeologici, ma in ogni caso essi hanno costituito il punto di partenza per successive interpretazioni, concernenti lo sviluppo di un sito come la conoscenza di un territorio.

Quali sono, dunque, gli aspetti della ricerca intesi a preservare l'informazione archeologica? Per prima cosa citerei la salvaguardia dell'integrità dei testi, che ha consentito di conservare la terminologia originaria utilizzata nella fase di redazione dei documenti. Un esempio chiarificatore è offerto proprio dai diari di scavo (fig. 9): nel corso della prima campagna, nel 1983, fu riportato alla luce un «recinto» eseguito a blocchi di tufo che ben presto si rivelò una sorta di coronamento che regolarizzava i bordi di una «struttura» di notevoli proporzioni e di forma parallelepipedica, interamente scavata nel tufo e completamente colmata di terra.

Nel corso delle operazioni di scavo e di svuotamento della terra di riempimento, tale «cavità artificiale», profonda circa 11 metri, fu definita dagli archeologi con termini diversi, strettamente connessi con la sua destinazione d'uso: «cisterna», per il confronto con altri monumenti simili in ambito etrusco; «cava» a cielo aperto, per la presenza sul fondo di gradoni con tracce di lavorazione connesse con il distacco di blocchi di tufo; «vasca», per la sua evidente funzione di deposito e smaltimento di acque impure, ivi convogliate da un complesso sistema idrico, riportato alla luce nel corso delle successive campagne di scavo.

Basti questa complessità terminologica e insieme interpretativa per dimostrare come sarebbe stato limitativo scegliere un unico vocabolo descrittivo, frutto di un'analisi a posteriori e quindi privo di utilità per chiarire i momenti successivi e i meccanismi della progressiva costruzione delle conoscenze archeologiche. Non sempre, infatti, è possibile ricondurre in schede standard di catalogazione dati che non sono solo il frutto di operazioni volte alla documentazione oggettiva e sintetica, ma che implicano processi interpretativi diacronici, valutazioni soggettive, intuizioni (Hodder, 1992; Wolle e Tringham, 2000).

Oltre all'aspetto terminologico, la procedura adottata consente di recuperare frammenti del passato, a volte perduti – anche lo scavo in fondo è un'operazione distruttiva – dando loro un'esatta collocazione spaziale, elemento fondamentale per ogni studio di carattere topografico e per la salvaguardia del patrimonio storico-archeologico. Ad esempio, nel caso dell'opera di Ercole Nardi, l'intento perseguito è di ricollocare con esattezza, tramite un

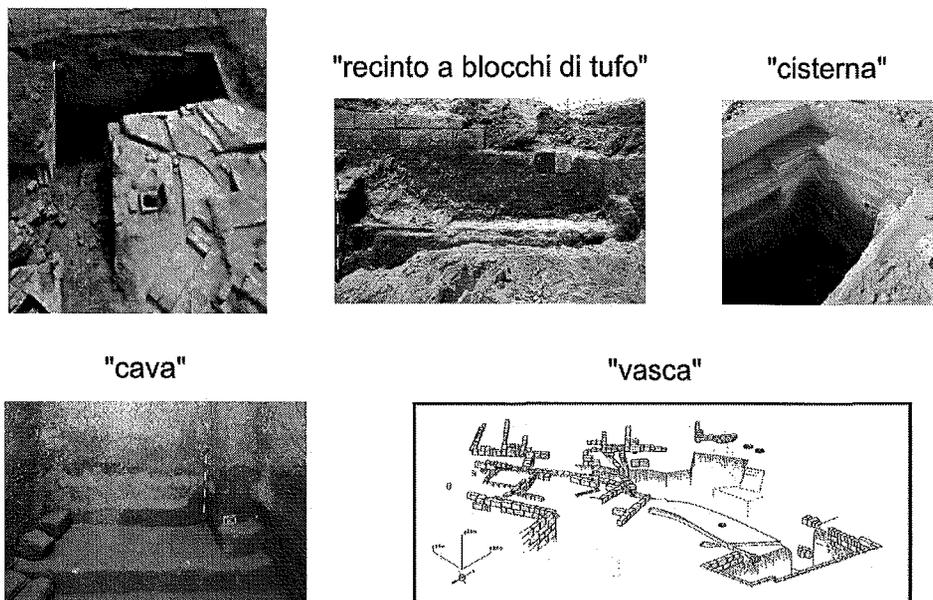


Fig. 9. – Cerveteri (Vigna Parrocchiale): la varietà terminologica nei diari di scavo come testimonianza del processo interpretativo archeologico.

processo di georeferenziazione dei dati, le evidenze archeologiche citate nel testo, e sistematicamente marcate, posizionandole nel loro contesto territoriale e su una base cartografica moderna, a integrazione dei dati provenienti dalle ricognizioni in corso. La conservazione, in questo caso, è intesa dunque non solo come rivalutazione dell'antica percezione del territorio, ma anche come registrazione delle trasformazioni subite dal paesaggio, in un processo di valorizzazione dei dati, che alla ricostruzione diacronica delle strutture archeologiche aggiunge quella del loro studio nel corso del tempo.

Anche la diffusione delle conoscenze contribuisce alla fruizione e alla valorizzazione delle informazioni. Ad esempio, nel quadro del Progetto Valle del Tevere e del previsto sistema museale della Sabina tiberina, la versione elettronica del testo del Nardi e delle relative immagini potrebbe costituire un punto di partenza per una ricerca volta all'analisi dei rapporti tra gli strumenti della comunicazione multimediale e la loro influenza sulla formazione culturale, da sperimentare sui fruitori dei musei locali.

### 7. Dalla marcatura alla «riscrittura».

A conclusione di questa breve presentazione di un'applicazione dei linguaggi di marcatura a documenti archeologici non strutturati o semistrutturati di carattere narrativo, mi sia consentito un ultimo breve commento generale. Ci sembra che l'approccio metodologico adottato per la codifica elettronica dei testi archeologici sia originale e innovativo. Tale approccio, infatti, mentre ha trovato un'ampia diffusione nella codifica di testi letterari e di documenti storici e linguistici, non risulta finora ampiamente diffuso negli studi archeologici dedicati alle ricerche sul territorio, soprattutto se comparato con altri sistemi quali i database e i GIS.

E questo appare singolare, se si pensa che già alla fine degli anni Ottanta Sebastian Rahtz (cfr. in particolare Rahtz *et al.*, 1992) proponeva un modello ipertestuale dinamico di informatizzazione dei rapporti di scavo mediante l'uso di SGML, nell'ambito di un approccio più generale dedicato alle potenzialità dei sistemi multimediali per la pubblicazione elettronica dei dati e per la didattica in archeologia. Del pari, però, va notato che circa un decennio dopo c'era chi riteneva, nell'ambito delle problematiche di archiviazione dei dati archeologici, che «even good standards in the computer world, such as SGML, have a short, brutish life» (Eiteljorg, 1999: 119).

L'uso di XML, ancor più che quello di SGML, si è comunque andato diffondendo nel corso degli ultimi anni in relazione alla codifica e alla strutturazione dei dati archeologici, come forma innovativa di codifica di database testuali, soprattutto al fine della descrizione e della catalogazione delle risorse in formato digitale e della loro distribuzione in rete (Niccolucci, 2002), mentre più complesso appare lo sviluppo di procedure di information retrieval (per il settore dei beni culturali cfr. ad esempio Corti *et al.*, 1998 e da ultimo Colazzo *et al.*, 2002; cfr. anche Schloen, 2001: 123-152 e il Progetto XSTAR, oggi OCHRE, promosso dall'Oriental Institute of the University of Chicago: <http://ochre.lib.uchicago.edu/>). In generale, comunque, i linguaggi di marcatura in archeologia sono presi in considerazione in relazione ai possibili sviluppi della pubblicazione elettronica dei dati e ai diversi aspetti della comunicazione in rete, in quanto offrono un più facile accesso alle informazioni e uno scambio più trasparente non solo dei dati ma anche dei modelli con cui sono stati strutturati: ad esempio, un settore in cui essi hanno avuto una rapida affermazione è quello della Realtà Virtuale, dove se ne auspica l'utilizzazione per rendere disponibili i dati e i metodi che hanno contribuito

alla distribuzione dei modelli virtuali attraverso Internet e che sono quindi alla base della loro visualizzazione (Ryan, 2001).

La distribuzione dei dati in rete apre un altro motivo di riflessione. Nella storia dell'informatica archeologica, fin dalle prime applicazioni statistiche per la realizzazione automatica di liste tipologiche o fin dai pionieristici sforzi di Jean-Claude Gardin per la realizzazione di codifiche descrittive dei materiali archeologici, l'intento perseguito, forse per la natura stessa della ricerca archeologica che procede attraverso classificazioni, confronti e analogie, è stata la ricerca della sintesi delle informazioni, che alcuni hanno anche inteso come ricerca dell'oggettività. Ne sono derivati, come diretta conseguenza, una particolare attenzione alle questioni relative al linguaggio scientifico utilizzato dagli archeologi e un acceso dibattito, che ha coinvolto a più riprese e con esiti diversi varie generazioni di studiosi, dagli esponenti della *New Archaeology*, ai fautori della scuola post-processuale e infine ai sostenitori dell'«archeologia cognitiva».

Ora, le nuove forme di comunicazione interattiva e multimediale sono state accolte da alcuni come possibili soluzioni di questo annoso problema, in quanto esse consentono di rendere più accessibili i documenti prodotti dalla ricerca, secondo modelli maggiormente versatili rispetto alle forme tradizionali di pubblicazione dei dati. Il web, d'altronde, si sta rapidamente affermando come strumento privilegiato per la trasmissione delle informazioni e ciò comporta inevitabilmente una più approfondita esplorazione di nuove forme del discorso archeologico.

Così, negli ultimi anni, a fronte non solo della crescita esponenziale dei costi delle pubblicazioni, ma a ragione della consapevolezza che si scrive più di quanto si legga e che ormai è più frequente la consultazione dei testi piuttosto che la loro lettura analitica, sono stati avviati progetti per sperimentare nuove forme di scrittura, in cui il supporto digitale costituisce lo strumento, cioè la tecnologia, che consente di porre in atto idee di riorganizzazione dei testi archeologici già precedentemente proposte. L'esempio più evidente, incentrato sulla volontà di separare le parti narrative del discorso storico da quelle cognitive, è stato promosso dal programma logicista di Gardin fin dalla fine degli anni Settanta e si è oggi materializzato nella collezione *Référentiels*, pubblicata dalla Maison des Sciences de l'Homme di Parigi, e nel Progetto ARKEOTEK (Gardin, 2004).

In questo contesto più specifico di riferimento, inteso non tanto alla pubblicazione di testi elettronici interrogabili in rete quanto piuttosto alla loro

«riscrittura», per esaltarne la struttura logico-discorsiva e rendere consultabili i dati che ne sono alla base, la nostra limitata esperienza si pone come una via intermedia, intesa al recupero e alla formalizzazione di testi archeologici narrativi, redatti in un passato anche recente, per valorizzarne i contenuti. Ciò non vuol dire che l'approccio adottato non possa essere utilizzato anche nella stesura di nuovi testi – come d'altronde si sta già sperimentando nella redazione dei diari dei recenti scavi ceretani – in cui vengano integrate al contempo le qualità narrative, e perché no didattiche, con una più attenta analisi dei contenuti. Quest'ultima dovrebbe tendere a sintetizzare e rendere esplicite, proporrei ancora una volta attraverso la marcatura, le unità strutturali del testo, per evidenziarne il carattere di elementi descrittivi ma anche di elementi interpretativi, cioè di documentazione e di rappresentazione, di espressione e di contenuto.

#### BIBLIOGRAFIA

- AA.VV., 2003. *Towards a Semantic Web for Heritage Resources*. DigiCULT Project, Thematic Issue 3; [http://www.digicult.info/downloads/ti3\\_high.pdf](http://www.digicult.info/downloads/ti3_high.pdf)
- ADAMO G., 1987. *La codifica come rappresentazione. Trasmissione e trattamento dell'informazione nell'elaborazione automatica di dati in ambito umanistico*. In: G. GIGLIOZZI (ed.), *Studi di codifica e trattamento automatico di testi*. Bulzoni, Roma: 39-63.
- BARCHESI C., 2001. *Progetto Caere: un'applicazione Internet attiva per l'Information Retrieval di documenti SGML*. *Archeologia e Calcolatori*, 12: 71-90.
- BARCHESI C., MOSCATI P., SANTORO P., SCARPATI D., 2003. *Ricerche archeologiche sul campo e archivi digitali: il manoscritto di Ercole Nardi*. *Archeologia e Calcolatori*, 14: 295-325.
- BONINCONTRO I., 2001. *Progetto Caere: prospettive di applicazione degli standard internazionali per la codifica dei dati testuali*. *Archeologia e Calcolatori*, 12: 55-70.
- CASTAGNOLI F., 1978. *La Carta Archeologica d'Italia*. *La Parola del Passato*, 33: 78-80.
- CECCARELLI L., 2001. *Progetto Caere: dallo scavo al territorio. Una soluzione per la distribuzione dei dati mediante un GIS on-line*. *Archeologia e Calcolatori*, 12: 105-121.
- COLAZZO D., SARTIANI C., ALBANO A., GHELLI G., MANGHI P., LINI L., PAOLI M., 2002. *A typed text retrieval query language for XML documents*. *Journal of American Society for Information Science and Technology*, 53, 6: 467-488.
- CORTI F., PAOLI M., SPRUGNOLI R., 1998. *TreSy un motore di ricerca testuale per documenti strutturati*. *Bollettino d'Informazioni del Centro di Ricerche Informatiche per i Beni Culturali*, 8, 1: 49-71.
- EITELJORG II H., 1999. *Archiving archaeological data*. In: L. DINGWELL et al. (eds.), *Computer Applications and Quantitative Methods in Archaeology, Proceedings of the 25<sup>th</sup> Anniversary Conference (Birmingham 1997)*. BAR International Series, 750, Archaeopress, Oxford: 117-120.
- GAMURRINI G.F., COZZA A., PASQUI A., MENGARELLI R., 1972. *Carta Archeologica d'Italia (1881-1897). Materiali per l'Etruria e la Sabina*. Leo S. Olschki, Firenze.

- GARDIN J.-C., 2002. *Les modèles logico-discursifs en archéologie*. In: F. DJINDJIAN, P. MOSCATI (eds.), *XIV Congress of the UISPP (Liège 2001), Commission IV, Data Management and Mathematical Methods in Archaeology*. Archeologia e Calcolatori, 13: 19-30.
- GARDIN J.-C., 2003. *Archaeological discourse, conceptual modelling and digitalisation: an interim report of the logicist program*. In: M. DOERR, A. SARRIS (eds.), *The Digital Heritage of Archaeology, Computer Applications and Quantitative Methods in Archaeology, Proceedings of the 30<sup>th</sup> Conference (Heraklion, Crete 2002)*. Archive of Monuments and Publications, Hellenic Ministry of Culture: 5-11.
- GARDIN J.-C., 2004. *The ARKEOTEK Project: A European network of knowledge bases in the archaeology of techniques*. In: P. MOSCATI (ed.), *New Frontiers of Archaeological Research. Languages, Communication, Information Technology*. Archeologia e Calcolatori, 15: 25-40.
- GIGLIOZZI G., 2003. *L'informatica, la didattica e il grillo parlante*. In: D. FIORMONTE (ed.), *Informatica umanistica. Dalla ricerca all'insegnamento*. Bulzoni, Roma: 11-18.
- GUIDI A., SANTORO P., 2003. *Il Progetto Galantina*. In: AA.VV., *Lazio e Sabina. Atti del Convegno*. De Luca Editore, Roma: 109-114.
- HODDER I., 1992. *Theory and practice in Archaeology*. Routledge, London-New York: 263-274.
- MARIOTTI S., 2001. *Progetto Caere: proposta di un modello per il trattamento e la codifica di documenti archeologici editi*. Archeologia e Calcolatori, 12: 91-104.
- MECKSEPER C., WARWICK C., 2003. *The electronic publication of archaeological field reports using XML*. *Literary and Linguistic Computing*, 18, 1: 63-75.
- MENGARELLI R., 1936. *Il luogo e i materiali del tempio di Hera a Caere*. *Studi Etruschi*, 10: 67-86.
- MENGARELLI R., 1937. *Iscrizioni etrusche e latine sui cippi sepolcrali, su vasi fittili e oggetti diversi trovate negli scavi della città e della necropoli di Caere*. *Notizie degli Scavi di Antichità*: 355-455.
- MOSCATI P., 2000. *The "Caere Project": Methodological and technical considerations*. In: A. GUARINO (ed.), *Proceedings of the II International Congress Science and Technology for the Safeguard of Cultural Heritage in the Mediterranean Basin (Paris 1999)*. Elsevier, Paris: 119-128.
- MOSCATI P., 2002. *From an Etruscan town to modern technologies: new advancements in the "Caere Project"*. In: F. DJINDJIAN, P. MOSCATI (eds.), *cit.* Archeologia e Calcolatori, 13: 135-149.
- MOSCATI P., 2003. *Dal dato al modello: l'approccio informatico alla ricerca archeologica sul campo*. In: T. ORLANDI (ed.), *I modelli nella ricerca archeologica. Il ruolo dell'informatica. Convegno Internazionale (Roma, 23-24 novembre 2000)*. Contributi del Centro Linceo Interdisciplinare "B. Segre", 107, Accademia Nazionale dei Lincei, Roma: 55-76.
- MOSCATI P., MARIOTTI S., LIMATA B., 1999. *Il "Progetto Caere": un esempio di informatizzazione dei diari di scavo*. Archeologia e Calcolatori, 10: 165-188.
- NICCOLUCCI F. (ed.), 2002. *Dalla fonte alla rete. Il linguaggio XML e la codifica dei documenti storici, archeologici e archivistici*. Bollettino d'Informazioni del Centro Ricerche Informatiche per i Beni Culturali, 12, 1.
- ORLANDI T., 1999. *Testi, modelli, e sistemi*. In: AA.VV., *Il ruolo del modello nella scienza e nel sapere (Roma, 27-28 ottobre 1998)*. Contributi del Centro Linceo Interdisciplinare "B. Segre", 100, Accademia Nazionale dei Lincei, Roma: 73-90.
- RAHTZ S., HALL W., ALLEN T., 1992. *The development of dynamic archaeological publications*. In: P. REILLY, S. RAHTZ (eds.), *Archaeology and the Information Age. A Global Perspective*. Routledge, London-New York: 360-386.
- RYAN N.S., 2001. *Documenting and validating Virtual Archaeology*. Archeologia e Calcolatori, 12: 245-273.
- SANTORO P. et al., 2005. *Il Progetto Galantina*. In: P. ATTEMA, A. NIJBOER, A. ZIFFERERO (eds.), *Papers in Italian Archaeology VI. Communities and Settlements from the Neolithic to the Early Medieval*

*Period. Proceedings of the 6<sup>th</sup> Conference of Italian Archaeology (Groningen 2003)*, BAR International Series, 1452, Archaeopress, Oxford: 993-1007.

SCARPATI D. (ed.), in corso di stampa. *Ruderi delle Ville Romano - Sabine nei dintorni di Poggio Mirteto, illustrati dal Prof. Ercole Nardi, 1885. Itinerario Primo.*

SCHLOEN J.D., 2001. *Archaeological data models and web publication using XML*. Computers and the Humanities, 35, 2: 123-152.

SIGNORE O., 2001. *Il ruolo centrale di XML nell'evoluzione del web*. <http://www.w3c.it/papers/>

SIGNORE O., 2002. *Una panoramica delle tecnologie W3C, RDF per la rappresentazione della conoscenza*. <http://www.w3c.it/papers/>

WOLLE A.-C., TRINGHAM R.I., 2000. *Multiple Çatalhöyüks on the World Wide Web*. In: I. HODDER (ed.), *Towards Reflexive Method in Archaeology: the Example at Çatalhöyük*. Oxbow, Oxford: 207-217.

*Caere 1* = CRISTOFANI M., NARDI G., RIZZO M.A., 1988. *Caere 1. Il parco archeologico*. CNR, Roma.

*Caere 3.1* = CRISTOFANI M. (ed.), 1992. *Caere 3.1. Lo scarico arcaico della Vigna Parrocchiale*. CNR, Roma.

*Caere 3.2* = CRISTOFANI M. (ed.), 1993. *Caere 3.2. Lo scarico arcaico della Vigna Parrocchiale*. CNR, Roma.

*Caere 4* = CRISTOFANI M. et al., 2003. *Caere 4. Vigna Parrocchiale: scavi 1983-1989. Il santuario, la "residenza" e l'edificio ellittico*. CNR, Roma.

RAUL MORDENTI\*

## PUBBLICAZIONE DELLE RICERCHE UMANISTICHE IN AMBIENTE DIGITALE

### *1. Una delimitazione del campo, che invece di chiuderlo lo apre.*

Occorre, anzitutto, un tentativo di delimitazione del nostro tema, costituito dalle pubblicazioni umanistiche in ambiente digitale<sup>1</sup>: con la parola «pubblicazioni» si intende qui riferirsi esclusivamente alle *pubblicazioni scientifiche*, cioè a quelle pubblicazioni che hanno per oggetto l'incremento della conoscenza e che a tale scopo sono essenzialmente rivolte e ordinate. Più banalmente potremmo dire che useremo il termine esattamente nel senso in cui si parla di «pubblicazioni» da presentare ad un concorso o da elencare in un curriculum.

Dunque il nostro problema è da distinguere, in via di principio, da quello del *copyright* e dei diritti d'Autore, un problema che, benché appaia strettamente intrecciato al nostro, rappresenta in realtà una questione diversa, soprattutto economica e giuridica (benché forse oggi *de iure condendo*), da trattarsi evidentemente in altra sede e con competenze ben diverse dalle mie<sup>2</sup>.

\* Università degli Studi di Roma «Tor Vergata».

<sup>1</sup> Questo intervento riutilizza una relazione da me svolta presso il Centro Linceo Interdisciplinare «B. Segre», nell'ambito di un Seminario dedicato al problema del «Riconoscimento legale della pubblicazione elettronica di ricerche scientifiche»; l'occasione mi è propizia per ringraziare l'infaticabile animatore delle nostre riflessioni, Tito Orlandi, e anche i colleghi presenti a quel seminario, come Mariella Guercio, Paola Moscati, Gianni Adamo, Nicola Tangari e Dino Buzzetti; un grazie particolare va a quest'ultimo che ha avuto la bontà di farmi pervenire le sue preziose osservazioni dopo quegli incontri.

<sup>2</sup> Nella vastissima bibliografia sull'argomento (che si arricchisce ora, e non certo per caso, di una sconfinata sitografia in continuo aumento disponibile in rete) ci limiteremo a segnalare il recente lavoro di sintesi rivolto al versante storico-letterario che più ci riguarda: Borghi (2003); si veda anche l'acuta messa a punto di Palazzolo (2003). È prezioso, sia per gli aspetti strettamente informatici del problema che per una serie di casi giurisprudenziali presentati, il lavoro di Sordi (1999). La classica ricerca dello storico e giurista Jacques Boncompain (2002) risale alle origini storiche della questione che pare stret-

In effetti proprio questa rinuncia a considerare sotto la specie giuridica ed economica la questione, che potrebbe apparire come una drastica delimitazione di campo, «apre» invece dal punto di vista teorico il nostro problema, perché ci spinge a considerare il concetto di pubblicazione scientifica *in quanto tale*.

Ciò significa che per «pubblicazione» noi intenderemo qui la *procedura* (generalissima) tramite la quale all'interno di una società (di una comunità scientifica) data vengono comunicati gli incrementi di conoscenza, cioè qualsiasi innovazione (o scoperta) nel campo del sapere umano<sup>3</sup>.

Tale procedura di comunicazione, a ben vedere, è *duplice*, giacché essa riguarda sia (a) la *diffusione pubblica*, dell'innovazione conoscitiva ottenuta, sia (b) la sua *validazione* (o il suo *riconoscimento*) da parte della comunità scientifica.

## 2. *Quattro domande, fra le altre, ancora senza risposta.*

Le prime domande che vengono alla mente (domande che mi paiono particolarmente consone alla natura di una sede multidisciplinare e interdisciplinare per antonomasia, quale l'Accademia che ci ospita) riguardano dunque la individuazione più precisa di tale procedura nei diversi tempi e nei diversi settori della produzione di conoscenza. Come «pubblicava» Aristotele? Solo attraverso gli appunti dei suoi allievi grazie ai quali conosciamo le sue opere? E come «pubblicavano» San Tommaso o Galileo? Non solo gli storici della scienza ma anche gli scienziati potrebbero aiutarci a rispondere ad una serie ordinata di domande, a cominciare dalle seguenti quattro:

tamente connessa all'invenzione della stampa: il primo «privilegio» editoriale è quello ottenuto da Aldo Manuzio nel 1495 dal Senato di Venezia per la stampa delle opere di Ariosto. È ora da vedere l'importante ricostruzione storica e giuridica del problema svolta da Laura Moscati (2001) lungo l'asse franco-italiano (che evidenzia il punto di svolta rappresentato dalla legislazione napoleonica e dall'idea di una *propriété incorporelle*). Segnalo anche *Filosofia e diritto* (2003).

<sup>3</sup> Non vale dunque qui la distinzione fra scienze umane e scienze della natura; per quanto possa sembrare strano ad alcuni colleghi scienziati, anche le pubblicazioni degli umanisti hanno (o dovrebbero avere) per scopo la comunicazione di incrementi di conoscenza.

1. *La produzione dell'innovazione.* Come veniva prodotta in passato la ricerca, o più semplicemente l'innovazione conoscitiva, nei vari e diversi campi del sapere umano (scienza e tecnologia, medicina e filosofia naturale, filosofia, filologia ecc.)? Pure ammettendo che l'innovazione in quanto tale non costituisse affatto un valore nelle società di *ancien régime*, tuttavia non è dubbio che innovazione sia stata sempre prodotta. Ma in quali forme? Attraverso quali istituti? Secondo quali procedure? In applicazione di quali statuti?

2. *La trasmissione dell'innovazione.* Come veniva scambiata/trasmessa l'informazione relativa alla conoscenza innovativa prodotta? Da quando esiste, ammesso che esista per tutti i settori, un ambito comunicativo definito per cui si possa usare l'espressione «comunità scientifica»? E come circolava in altri tempi e in altri assetti sociali la conoscenza innovativa all'interno di una tale comunità?

3. *L'«autorialità», o la proprietà individuale dell'innovazione.* Sappiamo bene che non sempre è esistito il concetto di «autore» individuale quale noi lo intendiamo, e che dunque tale assenza riguarda sia le opere letterarie che le scoperte scientifiche. Ma allora, in quegli assetti storico-sociali privi del concetto di «autore», chi era considerato il responsabile delle innovazioni? Esisteva una qualche forma di riconoscimento positivo per l'innovazione? E, se sì, cosa comportava tale riconoscimento?

4. *Il conflitto fra diffusione e proprietà dell'innovazione.* Infine, e soprattutto: come veniva risolto in passato il *conflitto* fra la diffusione della nuova conoscenza e il riconoscimento sociale della sua appartenenza allo scopritore?

Richiamo l'attenzione su questo ultimo punto: sembra infatti evidente che fra questi due gesti (la diffusione della scoperta e l'attribuzione esclusiva allo scopritore, il quale si deve almeno difendere dal plagio) esista addirittura una sorta di *contraddizione di principio*, difficilmente aggirabile.

### 3. *Modalità pre-moderne della «pubblicazione».*

Sappiamo che la lezione universitaria e le sue *reportationes*, il trattato e la lettera scientifica (personale e privata o *erga omnes*? Diffusa come? Con quali modalità e quali limiti?), infine la relazione svolta presso l'accademia ecc., sono state in passato modalità di pubblicazione in grado di risolvere entrambe le esigenze contraddittorie ricordate nella nostra quarta e ultima domanda,

cioè garantire davvero *al tempo stesso* l'esigenza della diffusione e quella del riconoscimento della paternità dell'innovazione.

Sappiamo che anche il moderno sistema della *rivista scientifica* (già presente nei secoli XVII-XVIII, e poi diffusasi dall'Inghilterra e dalla Francia a tutta l'Europa) risolveva più o meno efficacemente il nostro problema, perché garantiva, al tempo stesso, la diffusione della scoperta nella comunità scientifica (le riviste di cui parliamo ebbero presto diffusione pressoché universale all'interno di una comunità scientifica di proporzioni ristrette) e l'attribuzione dell'innovazione al suo Autore.

Ma in età pre-moderna come funzionava esattamente la faccenda? Il paleografo iberico Francisco Gimeno Blay<sup>4</sup>, riferendosi alla situazione del medioevo europeo, ha attribuito una nuova importanza al dono dell'esemplare d'Autore a un'autorità mecenatizia, un'offerta che seguiva immediatamente alla composizione e precedeva ogni ulteriore diffusione dell'opera; non si tratterebbe solo di un omaggio (sia pure interessato) ma anche di una precisa *procedura di pubblicazione*, giacché quel dono (appunto: assolutamente *pubblico*) a un'autorità riconosciuta consentiva di radicare l'autorialità e l'originalità dell'opera in modo ufficiale, e però anche di diffondere quel lavoro, giacché proprio a partire dall'esemplare di dono (come ha dimostrato Gimeno Blay) venivano poi richieste, autorizzate ed eseguite le copie.

Le modalità concrete di tali processi sarebbero da indagare più approfonditamente; in particolare esiste una diffusione della conoscenza che prescinde del tutto dalla pubblicazione scritta: ad es. tutta l'area vastissima del sapere che potremmo definire tecnico-artigianale, trasmesso direttamente fra persone e corpi, da una bocca ad un orecchio o forse da una mano ad un'altra, un sapere che sarebbe importante poter conoscere meglio<sup>5</sup>.

<sup>4</sup> Dell'Università di Valencia. Cfr. Gimeno Blay (in corso di stampa).

<sup>5</sup> Per secoli le informazioni tecnologiche si sono diffuse attraverso lo spostamento fisico degli artigiani da un paese all'altro: è questo il caso della «importazione» in Gran Bretagna dei mattonai fiamminghi e dei vetrai francesi (ed è anche il caso più noto della diffusione in tutta Europa della tecnologia della stampa a caratteri mobili grazie alla diaspora degli artigiani di Magonza); cfr. Ubbelohde (1964: 677-695, in particolare il § 1 «Esperienza e segreti artigianali»: 677-685). Lo stesso Ubbelohde, dopo aver definito «uno dei problemi che non hanno trovato soluzione» quello della trasmissione dei segreti artigianali (specie di più alto livello), scrive purtroppo: «(...) non sarà intrapresa qui nessuna indagine completa sullo scambio e la diffusione delle tecniche artigianali nelle età passate, per quanto affascinante possa essere questa ricerca» (*ibid.*: 678).

#### 4. *Il «sistema dei diritti d'Autore» e la sua crisi.*

Negli ultimi secoli il nostro problema si intreccia con il moderno sistema dei diritti d'Autore e dei brevetti che fra la fine del secolo XVIII e quella del XX sembra aver garantito (pur fra mille incertezze e difficoltà) un certo punto di equilibrio fra le due esigenze contraddittorie di cui si è detto, l'esigenza della circolazione pubblica dell'innovazione e quella della sua appropriazione personale da parte dell'autore. In questo assetto, infatti, è la stessa proprietà dell'innovazione che spinge il suo autore-proprietario a favorirne la diffusione, dato che questa (e questa soltanto) garantisce in effetti la remuneratività (morale o materiale che sia) dell'opera dell'ingegno.

Eppure anche questo punto di equilibrio sembra oggi essere scosso da una crisi irreversibile, e secondo molti è anzi diventato del tutto insostenibile<sup>6</sup>: non solo perché, come scrive Maria Iolanda Palazzolo, il diritto d'Autore si trova di nuovo come schiacciato dai suoi «due nemici storici oggettivamente alleati, la censura e la pirateria libraria» (Palazzolo, 2003: 2), ma soprattutto perché l'intero meccanismo di produzione/appropriazione della conoscenza viene stravolto e ridefinito in un'epoca che è stata chiamata «età dell'informazione», cioè nella nostra epoca in cui l'informazione stessa si rivela essere una forza produttiva, ed anzi la più formidabile delle forze produttive, la vera «materia prima della competizione mondiale», come afferma Zaki Laïdi su *Esprit*<sup>7</sup>. Lo stesso Laïdi ricorda che gli Stati Uniti hanno incassato solo nel corso dell'anno 2000 ben trentotto miliardi di dollari a titolo di diritti della proprietà intellettuale<sup>8</sup> (una cifra di cui il *software* informatico rappresenta gran parte), e ciò fa della proprietà intellettuale la prima voce delle esportazioni di quel paese; d'altra parte (è questo l'inevitabile rovescio della medaglia) la medesima voce diritti della proprietà intellettuale rappresenta il più decisivo impedimento allo sviluppo dei paesi poveri del mondo.

<sup>6</sup> Come è noto il sistema dei brevetti, assieme alla proprietà privata del *software*, rappresenta uno degli obiettivi polemici principali del movimento di contestazione della globalizzazione capitalistica.

<sup>7</sup> «Le savoir est devenu la matière première de la compétition mondiale. (...) Dans le système capitaliste actuel, la propriété intellectuelle équivaut au capital, qui équivaut à son tour au pouvoir» (Laïdi, 2003: 116-131, 119).

<sup>8</sup> *Ibid.*: 119. I dati sono tratti dal *Report of the Commission on Intellectual Property Rights* del 2002; disponibile in rete all'indirizzo [www.upi-commission.org](http://www.upi-commission.org)

Eppure nel luglio del 2000 la «Carta di Okinawa», che concluse il G8 svoltosi in Giappone, aveva affermato che proprio le *Information and Communication Technologies* avrebbero non solo dovuto «offrire agli individui la possibilità di raggiungere il proprio potenziale e di realizzare le proprie aspirazioni», ma anche permettere a tutti i popoli del mondo «una crescita economica sostenibile e il rafforzamento del *welfare*». Come si vede la contraddizione non potrebbe essere più evidente: il sistema delle ICT basato ancora sui brevetti e sulle proprietà intellettuali da un lato viene invocato per risolvere le disuguaglianze del mondo (il cosiddetto «*computer based welfare*») dall'altro rappresenta però di fatto il più formidabile fattore di ostacolo allo sviluppo e anzi di crescente impoverimento del mondo povero.

Ci troviamo forse di fronte ad un nuovo sistema di *enclosures*, invisibili ma ferree «recinzioni» del patrimonio informativo dell'umanità, del tutto analoghe alle recinzioni delle terre che accompagnarono la rivoluzione industriale e la connotarono di miserie indicibili?

Non si tratta, tuttavia, solo di obiezioni di tipo etico-politico bensì di una contraddizione ancora più radicale che oppone il vincolo del *copyright*<sup>9</sup> alle modalità di fruizione libera e gratuita che sembrano caratterizzare la rete Internet. Lo confermano alcuni esperimenti condotti negli Stati Uniti: l'autore di fantascienza Eric Flint ha convinto nel 2001 la propria casa editrice a offrire in rete ai lettori, con accesso del tutto libero e gratuito<sup>10</sup>, una biblioteca virtuale dove è possibile non solo leggere ma anche scaricare i suoi romanzi, gli stessi che venivano contemporaneamente distribuiti a pagamento in modo tradizionale nelle librerie; ebbene, la disponibilità gratuita nella rete di quei romanzi ha avuto come esito paradossale il sensibile aumento (e non la diminuzione!) delle vendite dei libri. Analogo risultato ha conseguito la casa editrice O' Reilly (specializzata in manuali, cartacei e *on line*, sui linguaggi di programmazione, il *software* e le nuove tecnologie) la quale propone una sezione del suo sito intitolata *Safari Bookshelf*.

Gli Autori italiani che si celano dietro il nome collettivo di Wu Ming/Luther Blissett (a cui si deve il fortunato romanzo *Q* edito da Einaudi; v.

<sup>9</sup> La sfumatura di significato fra «copyright» («diritto di *copia*») e «diritto d'*Autore*» sembra però tutt'altro che trascurabile.

<sup>10</sup> Cfr.: <http://www.baen.com>

Blissett, 1999) testimoniano il medesimo risultato per ciò che li riguarda<sup>11</sup> e ne forniscono una spiegazione: la diffusione gratuita tramite la rete non deve essere messa in rapporto contrastivo con la vendita in libreria, perché semmai essa rappresenta un'evoluzione della biblioteca pubblica, o del prestito amicale, o del «passaparola»; e dunque, così come la presenza di un libro nella biblioteca pubblica o il suo prestito da parte di un amico non ostacola le vendite in libreria (ma semmai le incrementa «facendo parlare» del libro in questione, e facendolo apprezzare) lo stesso può accadere anche nel caso dei testi offerti gratis in formato digitale.

Tanto più l'abbandono del vincolo del *copyright* sembra naturale nelle comunità scientifiche; così scrive Wu Ming 2 citando l'esperienza del Massachusetts Institute of Technology:

«Il progetto OpenCourseWare del MIT è un tentativo di riportare l'Università al suo scopo originario: rendere universale la cultura, piuttosto che razionarla e rivenderla solo a chi può permettersela. All'indirizzo <http://ocw.mit.edu>, fin dall'ottobre dello scorso anno [2003, NdR], sono consultabili i materiali di insegnamento utilizzati dai docenti di numerosi corsi tenuti al MIT. Nel giro di quattro o cinque anni il progetto conta di coprire l'intero spettro dei 2000 corsi che si tengono nell'Istituto. (...) molte persone potranno accedere al contenuto delle lezioni di uno degli istituti più prestigiosi del mondo. Tra queste persone sono compresi anche i professori di altre Università, che potranno utilizzare quei materiali, nonché integrarli ulteriormente, in una sorta di "docenza *open source*" dalla quale gli studenti hanno tutto da guadagnare»<sup>12</sup>.

E, tornando al nostro problema, sembra fuori di dubbio che anche questi lavori scientifici messi in rete generosamente dai docenti del MIT senza *copyright* debbano considerarsi a tutti gli effetti delle «pubblicazioni» scientifiche.

Ma ciò dimostra che i due elementi, la pubblicazione e il diritto d'Autore, non sono affatto legati fra loro necessariamente.

<sup>11</sup> «(...) Il romanzo che da più tempo è disponibile gratis in diversi formati sul sito [www.wumingfoundation.com](http://www.wumingfoundation.com) continua a vendere molto bene» (così scrive Wu Ming 2, *No copyright: le case editrici hanno tutto da guadagnarci*, nello *Speciale copyright*, inviato alla *mailing list* da [giap@wumingfoundation.com](mailto:giap@wumingfoundation.com)).

<sup>12</sup> *Ibid.*

### 5. Il dibattito fra Diderot e Condorcet, e le conoscenze come «beni pubblici».

Si riapre forse davanti a noi il dibattito che vide contrapposti Denis Diderot e il marchese di Condorcet (Laïdi, 2003: 120-121). Il primo, nella celebre *Lettre sur le commerce de librairie* del 1764 argomentò la necessità di trarre un beneficio economico dalle opere dell'ingegno (da dividersi, si noti, fra l'Autore e l'Editore, ma non fra i discendenti: nel suo caso i discendenti di La Fontaine a cui il re aveva concesso i privilegi per le favole). Una proposta che doveva servire secondo Diderot, ed effettivamente servì, a sviluppare l'industria editoriale e a permettere che gli autori vivessero autonomamente dei frutti delle loro opere, rompendo il secolare nesso servile del mecenatismo. Condorcet, al contrario, sostenne che i diritti privati dei singoli fossero secondari di fronte alla primaria esigenza sociale della diffusione del sapere, e dunque che anche i diritti degli autori dovessero e potessero essere limitati nell'interesse della collettività.

Su questa linea di Condorcet si mosse (sarebbe bene ricordarlo oggi) tutto il pensiero liberale, in nome della lotta contro il monopolio, da Jefferson a Madison, da Adam Smith al Congresso degli economisti liberali tedeschi del 1863 per culminare nell'art. 27 della *Dichiarazione universale dei diritti dell'uomo* del 1948<sup>13</sup>.

Il citato Zaki Laïdi (rielaborando il concetto di «bene pubblico» già proposto da Garrett Hardin, 1968, su *Science*) sostiene che, trattandosi di informazioni, ci troviamo di fronte a beni che *per loro stessa intrinseca natura* sono *pubblici*, e non possono non esserlo, giacché le informazioni presentano le due caratteristiche peculiari dei beni pubblici, cioè la «non-rivalità» e la «non-esclusione»<sup>14</sup>. In altri termini, beni come la bellezza dei fuochi d'arti-

<sup>13</sup> Dove sarà da notare una certa articolazione (se non si vuol parlare di contraddizione) fra il primo comma (ispirato a Condorcet?), comunque prevalente, e il secondo comma (ispirato a Diderot?) dell'art. 27: «(1) Ogni individuo ha diritto di prendere parte liberamente alla vita culturale della comunità, di godere delle arti e di partecipare al progresso scientifico ed ai suoi benefici. (2) Ogni individuo ha diritto alla protezione degli interessi morali e materiali derivanti da ogni produzione scientifica, letteraria o artistica di cui egli sia autore».

<sup>14</sup> «(...) L'information entre parfaitement dans la définition classique des biens publics. Elle repose sur le principe de non-rivalité qui signifie que l'accroissement du nombre des utilisateurs n'entraîne pas de dégradation de ce bien. Ce qui n'est pas le cas, par exemple, d'une terre arable. Elle invoque ensuite le principe de non-exclusion: il est difficile d'exclure certains utilisateurs de ressources offertes sur le Net» (Laïdi, 2003: 123).

ficio o il poter disporre di strade pulite (sono due esempi che Laïdi allega, ma si potrebbe aggiungere, da italiani: il godimento dei beni culturali e architettonici delle nostre città, o della luce del sole) non sono beni per i quali la fruizione da parte di un individuo entri in conflitto, «rivaleggi», con la fruizione da parte di qualsiasi altro, o la «escluda» (semmai è vero il contrario: si tratta di beni che tanto più sono godibili da uno e da ciascuno quanto più sono condivisi da tutti). Non sono forse proprio di questo tipo la conoscenza e la cultura se considerati in quanto «beni»? Non è forse così anche per la lettura di un libro<sup>15</sup>?

Il problema è attualmente all'ordine del giorno della Comunità Europea; ma quale che sia la conclusione del dibattito politico-giuridico, sembra che questa antica contraddizione fra le due diverse esigenze di cui si fecero storicamente portatori Diderot e Condorcet<sup>16</sup> debba e possa essere risolta in radice almeno nel caso delle ricerche che si svolgono in ambito universitario, cioè all'interno e per conto di istituzioni che hanno il loro fondamento e il loro principio ispiratore proprio nel dovere di contribuire al progresso delle conoscenze umane e alla loro universale diffusione.

#### *6. Le pubblicazioni dell'era di Gutenberg e «il quadrato dell'identificazione».*

A partire da una prospettiva diacronica del nostro problema, sembra necessario indagare l'influenza che l'epistemologia di Gutenberg (insisto: l'epistemologia, non solo la tecnologia di Gutenberg) ha esercitato a proposito del concetto di pubblicazione.

Consideriamo ad esempio ciò che possiamo definire «il quadrato dell'identificazione bibliografica», cioè la modalità vigente di identificazione di

<sup>15</sup> Specie dopo che le moderne tecnologie della riproduzione digitale promettono di risolvere perfino il rischio (peraltro il più delle volte marginale) della consunzione fisica delle pagine del libro ad opera di chi lo legge.

<sup>16</sup> Come scrive Georges Lefebvre: «Nessuno espresse meglio di Condorcet le speranze che, agli occhi dei discepoli degli Enciclopedisti, nobilitavano la missione assegnata all'insegnamento nazionale. (...) Interprete di un umanesimo abbastanza generoso da ripudiare l'egoismo censitario, il democratico Condorcet, riprendendo l'insegnamento di Descartes, ne ampliava singolarmente la portata, perché all'insegnamento nazionale assegnava la missione di "assicurare a ciascuno la facoltà di sviluppare tutti i doni ricevuti dalla natura e di stabilire così fra i cittadini un'eguaglianza di fatto, rendendo effettiva l'eguaglianza politica riconosciuta dalla legge" (...)» (Lefebvre, 1970: 636-637).

una pubblicazione, che noi tutti usiamo, e anzi consideriamo normale, se non addirittura «naturale»: ebbene tale modalità è invece del tutto guttemberghiana, e come essa non esisteva prima della stampa, cioè nell'ambiente chirografico pre-guttemberghiano, così è assai probabile che sia messa radicalmente in crisi (e non possa conservarsi intatta) nel passaggio dalla stampa alla modalità propriamente informatica<sup>17</sup> di trasmissione delle conoscenze.

Per «quadrato dell'identificazione», intendo i quattro elementi Autore, titolo, editore, anno di edizione, per meglio dire intendo la convenzionale *scelta* storica (scelta: cioè niente affatto naturale né ovvia) secondo cui si assume che:

- (1) un testo (che così diventa, appunto, un'opera) abbia sempre un Autore, ed uno soltanto<sup>18</sup>;
- (2) ogni testo-opera sia e rimanga sempre identico a se stesso, come l'identità del titolo annuncia e garantisce;
- (3) esso sia altresì identificabile in riferimento all'Editore (cioè a chi provvede alla stampa e alla diffusione<sup>19</sup>), a cui viene conferito anche per questa via un ruolo (e un potere) assai cospicuo;
- (4) e, infine, che l'anno (dell'edizione a stampa, si noti!, e non della composizione del testo) possa servire e bastare a collocare saldamente nel tempo l'intero sforzo di identificazione.

Noi sappiamo (soprattutto grazie a Michel Foucault) che questa rigida e costrittiva griglia identificativa risponde, in realtà, anzitutto ad esigenze «disciplinari», cioè in buona sostanza poliziesche, serve cioè a mettere finalmente «ordine nel discorso»; e sappiamo anche che solamente una testualità così organizzata consente (ad esempio) il pieno successo dell'attività della censura controriformista della Chiesa di Roma che, prima dell'insediarsi di tale tassonomia identificativa, si vedeva fatalmente sfuggire da tutte le parti la multiforme e cangiante varietà dei testi, inafferrabili come pesci nel mare.

<sup>17</sup> Con l'espressione «modalità propriamente informatica», cioè non mimetica e debitrice rispetto alla precedente configurazione guttemberghiana, intendo riferirmi essenzialmente alla rete e all'ipertesto.

<sup>18</sup> Per un'acuta quanto persuasiva critica dell'assolutizzazione storica del concetto di Autore (una critica, una volta tanto, proveniente da un approccio storico-filologico, e niente affatto filosofico misticheggiante) si veda ora: Canfora (2002).

<sup>19</sup> Si tratta dunque, anche in questo caso, di una *funzione*, piuttosto che di un individuo.

Ma non è tanto la storia di questo «quadrato» che qui ci interessa, quanto verificare come esso in effetti definisca anche i contorni del concetto vigente di pubblicazione scientifica e ne costituisca tuttora il fondamento:

1. una pubblicazione deve essere chiaramente attribuibile ad un individuo, e ad uno soltanto (se necessario anche tramite appositi attestati nel caso di testi anonimi e/o collettivi);
2. essa deve avere un titolo (anzi, non a caso, nella definizione accademica essa è «un titolo»); d'altra parte proprio il titolo è ciò che rende citabile il lavoro, ed è infatti l'elemento di cui nessuna pubblicazione può mai fare a meno;
3. deve essere *stampata*, cioè deve essere stata accettata da un Editore (e si noti come la «bozza di stampa», ma solo se autenticata dall'Editore!, sia considerata egualmente valida ai fini concorsuali);
4. deve, infine, essere datata (ed anche in questo caso l'elemento è talmente cogente da dover essere sostituito con apposite dichiarazioni legali nel caso di datazione incerta o insufficiente).

### 7. Aporie e insostenibilità del vigente concetto di pubblicazione.

Sembra in effetti che tutti e quattro questi elementi, considerati tuttora necessari per poter parlare di «pubblicazione scientifica» utilizzabile nei nostri concorsi, mostrino oggi la corda<sup>20</sup> e siano difficilmente sostenibili teoricamente (anche in questo caso, come spesso accade, l'informatica sembra aiutarci a rivelare insostenibili aporie già pre-esistenti ad essa, piuttosto che provarle).

Ad esempio il punto (1), l'attribuzione del titolo ad un Autore e ad uno solo, entra in clamorosa contraddizione con le modalità di effettivo svolgimento di molte ricerche, che sono in realtà davvero *collettive*, e non individuali: la soluzione a cui si ricorre per adeguare la verità effettuale del modo di fare ricerca alla norma della pubblicazioni (e non viceversa) è un'insensata,

<sup>20</sup> Già nel maggio 2002 Mario Pozzi dell'Università di Torino si è fatto promotore di un'iniziativa per indurre il Ministero dell'Università e della Ricerca a consentire la valutazione accademica anche delle pubblicazioni in rete.

umiliante (e spesso falsa) attribuzione di porzioni separate del lavoro unitario a diverse persone<sup>21</sup>.

Così come si potrebbe anche notare che il punto (4), la datazione del lavoro, non ha necessariamente a che fare con la data dell'effettiva «scoperta» o innovazione, ma di certo ha più a che fare con i tempi tecnico-commerciali dell'Editore e della sua tipografia (alcune case editrici particolarmente prestigiose pubblicano solo dopo due-tre anni dal ricevimento del libro<sup>22</sup>).

Ma certo il punto oggi più insostenibile è proprio il (3), cioè l'idea che un lavoro scientifico non sia serio (anzi neppure degno di essere preso in considerazione in quanto «pubblicazione») se prima non è stato accettato da un Editore. Questa presupposizione aveva forse qualche valore scientifico-culturale in passato, in un assetto dell'organizzazione della cultura in cui gli Editori (e/o le riviste) pubblicavano poco, e accettavano di pubblicare solo sulla base di Comitati di lettura prestigiosi ed affidabili<sup>23</sup>; di certo oggi questa situazione, soprattutto nel caso della ricerca umanistica, è del tutto superata, e la necessità di fare stampare comunque i propri lavori da un Editore significa solo affidare a quest'ultimo, al di là di qualsiasi controllo culturale, un potere cruciale del tutto ingiustificato dal punto di vista scientifico; è un potere che peraltro gli Editori non esitano a monetizzare, specie in prossimità dei concorsi (facendosi pagare l'edizione direttamente dagli Autori).

Si potrebbe peraltro ipotizzare, per paradosso, che un'opera davvero scientificamente degna ed impegnativa sia oggi proprio quella destinata a restare *inedita*, cioè quella che trova più difficoltà ad essere pubblicata da un Editore, ad esempio per la sua grande mole (tutti sappiamo che le edizioni critiche di opere anche importantissime ma troppo corpose sono praticamente im-pubblicabili) o per il suo stesso carattere teoricamente troppo innovativo (e infatti la storia della cultura del Novecento conosce casi di opere fondamentali che non trovarono un Editore disposto a pubblicarle).

<sup>21</sup> In altri sistemi si attribuisce valore diverso (e decrescente) alle diverse posizioni che gli Autori occupano nel frontespizio.

<sup>22</sup> Si potrebbe giungere addirittura a sostenere che l'importanza di una casa editrice sia direttamente proporzionale alla durata dei tempi di attesa.

<sup>23</sup> Questa situazione è ancora vigente in alcuni settori scientifici, dove la pubblicazione di un articolo accettato dalle (poche) riviste riconosciute del settore significa che esso è effettivamente valido ed originale.

Insomma mi pare che la configurazione vigente del concetto di pubblicazione non sia né oggettiva né razionale, e che non regga da nessun punto di vista al vaglio di un'analisi critica funzionale.

*8. Alcuni problemi che mettono in discussione il concetto tradizionale di Autore.*

Si pongono peraltro problemi teorici di fondo, che mettono in discussione i concetti tradizionali, a cominciare da quelli di Autore e di testo<sup>24</sup>. Tali problemi riguardano la natura stessa degli oggetti a cui ci riferiamo parlando di pubblicazioni informatiche, troppo spesso senza considerare che le modificazioni tecnologiche intervenute pongono anche radicali questioni teoriche (o forse si limitano a rivelarle).

Una pubblicazione informatica somiglia sempre meno a un libro guttenberghiano e sempre più invece a un «oggetto digitale complesso interattivo» il quale è concepito fin dall'inizio, e poi costruito, per essere fruito per mezzo di un computer, della rete e dei relativi programmi; tutto ciò sembra rendere del tutto superato il tradizionale concetto di testo come «deposito» statico (e stabile) di dati, siano essi testuali, iconici o di altra natura. Si noti che una tale definizione di oggetto digitale complesso interattivo non si riferisce solo alle banche dati ma comprende le edizioni, almeno quelle che non si limitino banalmente a digitalizzare su supporto magnetico il lavoro già cartaceo ma che *prendano sul serio* l'informatica e le sue potenzialità euristiche<sup>25</sup>.

Questo cambio di natura dell'oggetto informatico dipende anzitutto (come osserva Dino Buzzetti) dal fatto che sia la creazione sia l'implementazione sia la fruizione lettoriale dei prodotti informatici sono sempre la risultante dell'esecuzione di un *programma*, di un *software*, cioè del modo in cui i meri dati testuali vengono elaborati e gestiti. Senza l'insieme delle procedure e delle istruzioni che accompagnano e organizzano i dati testuali, questi ultimi sarebbero semplicemente inservibili, e dunque privi addirittura di senso, non solo di valore scientifico innovativo. Da questo punto di vista sembrerebbe quasi che siano proprio tali procedure e istruzioni meta-testuali a fornire il *valore ag-*

<sup>24</sup> Mi sia consentito, per l'articolazione di quest'affermazione qui formulata in modo apodittico, il rinvio a Mordenti (2002).

<sup>25</sup> Insomma edizioni informatiche che siano concepite dall'Editore per consentire ulteriori ricerche da parte del Lettore, come la produzione di concordanze, occorrenze, indici, la gestione delle varianti, ecc.

*giunto conoscitivo* dell'oggetto digitale, e dunque a costituirne la scientificità e l'originalità, due caratteristiche che nell'assetto tradizionale (gutemberghiano o pre-informatico) appartenevano invece interamente al testo d'Autore, e anzi lo definivano. Senza considerare che i dati in quanto tali spesso «incorporano», per così dire, i risultati delle ricerche precedenti: si pensi ad esempio ai dati catalografici, paleografici, codicologici ecc. che vengono incorporati nella digitalizzazione dei cataloghi o dei repertori dei fondi manoscritti. In questi casi chi è l'Autore? Non c'è forse una stratificazione complessa di diversi e successivi Autori che consente un accumulo crescente di conoscenze?

E chi rispondesse che non si può considerare Autore chi si limita a recuperare e organizzare e fare interagire i dati pre-esistenti dovrebbe però poi trarre le conseguenze radicali che derivano da una simile affermazione, e allora si troverebbe costretto a escludere dal concetto di Autore chiunque rielabori e riorganizzi diversamente le conoscenze pregresse: ma non è esattamente questa l'attività di *ogni Autore*? Il lavoro di un Autore, anche prescindendo dall'informatica, è appunto sempre «comporre» (s-comporre o ricomporre) l'esistente, dato che assai raramente egli parte dal nulla, e crea *ex nihilo*. Le note, le citazioni, le bibliografie, che sono tanta parte delle nostre pubblicazioni scientifiche (ed anzi che ne segnano la serietà), non alludono forse ad una simile attività essenzialmente *ri-elaborativa*, e non certo in assoluto «creativa»? L'attività della ricerca anche più innovativa sembra essere un intreccio inestricabile in cui lo scarto dell'innovazione è sempre strettamente connesso al padroneggiamento consapevole, ed al riuso innovativo, della tradizione, ma dunque in cui la novità è sempre legata al patrimonio già dato delle conoscenze. Se dunque gli organizzatori e gli implementatori di banche dati telematiche (si noti: un'attività che assorbe energie sempre crescenti anche nelle nostre discipline) non sono da considerarsi Autori di innovazione scientifica, si rischia allora di dover ammettere che nessuno, o quasi nessuno (eccezion fatta, forse, per alcuni artisti) è Autore.

Credo che non si possa non concordare con René Pellen il quale, facendo il punto sulla vera e propria *impasse* che la questione del diritto d'Autore delle opere digitalizzate sta vivendo nella Comunità Europea e nel mondo, afferma:

«Il sera sans doute indispensable, plus profondément, de redéfinir les concepts mêmes auxquels on se réfère pour leur associer des droits; entre autres, les concepts de *texte*, d'*auteur*, d'*édition électronique*, d'*hypertexte*, d'*espace*, de *temps*, etc.» (Pellen, 1999: 47).

In ogni caso, ciò che abbiamo definito «dati» e ciò che abbiamo definito «l'insieme delle procedure e delle istruzioni che accompagnano e organizzano i dati» sono, nelle pubblicazioni informatiche, assai difficilmente separabili, e ciò (una volta di più) dimostra l'impossibilità di limitarsi ad estendere pigramente alla produzione informatica i concetti derivati dal mondo del libro.

Ma se quanto detto finora sfuma (per dir così) *a monte* il concetto di Autore tradizionale, mescolandolo all'Autore del *software* e non solo a lui, occorre dire che la modalità informatica di edizione presenta anche un secondo punto di fuga, o una seconda «contaminazione» del concetto di Autore, questa volta *a valle*, cioè verso il Lettore stesso e la sua cooperazione attiva (o inter-attiva) che la fruizione informatica consente e anzi prevede. Non mi riferisco qui al problema complesso (ma troppo spesso proposto in modo confusionario e confuso) della cosiddetta «morte dell'Autore», che sarebbe il portato di una fruizione post-moderna del testo, del tutto destrutturata e interattiva. Penso invece a cose estremamente utili e serie, come sono le edizioni elettroniche concepite in modo tale che il Lettore stesso possa costruire al loro interno percorsi di ricerca inediti, cioè imprevisi dal Curatore (se non, per dir così, passivamente, cioè come mera potenzialità offerta ai Lettori): ad esempio la possibilità che il Lettore ha di produrre, secondo le proprie personali esigenze, spogli specifici, *Indices locorum* parziali, concordanze ecc., ma anche la possibilità, consentita da edizioni informatiche ben fatte, di seguire rami diversi di uno *stemma* o (ancora più utilmente) di portare a testo alcune varianti che l'Editore invece aveva proposto nell'apparato, e così via. Deriva da qui, come è ben noto agli allievi di Orlandi, l'assoluta necessità di produrre sempre *files* «onesti» dal punto di vista informatico, cioè trasparenti e trasportabili, in cui siano sempre dichiarate, e se necessario anche revocabili, le modalità della loro organizzazione interna, insomma in maniera tale che sia sempre possibile distinguere i dati primari (o propriamente testuali) da quelli secondari (gestionali o interpretativi) e risalire dai secondi ai primi, anche in vista di una, sempre possibile, diversa utilizzazione degli stessi dati primari forniti in formato digitale.

Se vogliamo sforzarci di ricercare (come è sempre utile fare) delle analogie fra le nostre riflessioni e quelle della ricerca umanistica pre-informatica, potremmo dire che ci troviamo qui esattamente di fronte al problema esposto già da Lanfranco Caretti nella sua classica definizione dell'edizione critica «ottima», che sarebbe quella in cui (per ipotesi) venga reso possibile ad un Lettore di giungere ad una diversa edizione critica, a partire però dagli stessi dati forniti dall'edizione critica in questione.

Quell'auspicio, del tutto teorico e tendenziale in Caretti, è reso invece dall'uso dell'informatica completamente attuabile, e ciò si trasforma nel dovere di perseguirlo per chi si accinge a produrre edizioni critiche sostenute dal computer.

Certo è che il concetto di Autore tradizionale (cioè, in buona sostanza, il concetto gutemberghiano di Autore) viene messo in crisi su due versanti: l'Autore di un oggetto digitale complesso interattivo che aspira ad apportare innovazione scientifica e incremento delle conoscenze<sup>26</sup> sembra essere (al tempo stesso) *di più* e *di meno* dell'Autore tradizionale gutemberghiano: è *di più*, perché egli fornisce sia i dati che le procedure che li organizzano (se necessario, anche traendoli da altri) producendo in tal modo un sovrappiù di conoscenza che viene offerto alla fruizione del Lettore. Ma il nostro Autore informatico è al tempo stesso *di meno*, dell'Autore gutemberghiano, dato che egli deve essere capace, per così dire, di *ritirarsi* dalla sua stessa creazione (come il Dio di certa mistica ebraica), appunto per consentire al Lettore una funzione cooperativa forte, tale, per ipotesi, da poter produrre anche una diversa e migliore utilizzazione degli stessi dati forniti dall'edizione.

*9. Nella procedura di ricerca pre-informatica la pubblicazione rappresenta effettivamente il momento centrale.*

Sembra allora che la nostra attenzione si debba concentrare (una volta di più) sulla *procedura* di cui la pubblicazione è parte; per meglio dire, riflettere sulla posizione assai diversa che la pubblicazione del risultato occupa rispettivamente nella procedura della ricerca scientifica pre-informatica (che, come abbiamo visto, è segnata essenzialmente da Gutenberg) e in quella segnata dall'avvento dell'informatica, e della rete.

Nella configurazione attuale (o tradizionale, o pre-informatica, o gutemberghiana che dir si voglia), il carattere cruciale della pubblicazione consiste nel fatto che essa si pone *crono-logicamente* nel momento decisivo, e precisamente quando la ricerca è già conclusa e i suoi risultati vengono offerti alla discussione, al giudizio e alla fruizione da parte della comunità, scientifica e non.

<sup>26</sup> Potrebbe essere proprio questa la definizione di una «pubblicazione scientifica» in ambiente digitale.

Se cerchiamo di delineare la successione in fasi di tale procedura di ricerca, dal momento assolutamente iniziale della prima ideazione (intenzione di ricerca?) al momento finale della sua definitiva attuazione (applicazione?), potremmo descrivere un percorso di questo tipo:

SCHEMA 1. *La procedura di ricerca tradizionale e la centralità della pubblicazione.*

---

(a) Prima ideazione, Ipotesi di ricerca → (b) Riflessione (personale), Discussione (privata) → (c) Progettazione (formulazione dell'ipotesi della ricerca, da verificare) → (d) Messa in opera della ricerca → (e) Svolgimento della ricerca → (f) Conseguimento dei primi risultati (verifica dell'ipotesi della ricerca) → (g) **Pubblicazione** dei risultati della ricerca → (h) Discussione all'interno della comunità scientifica → (i) Verifica della validità, accettazione/confutazione pubblica dei risultati della ricerca → (l) Applicazione → (m = a') Formulazione di nuove ipotesi di ricerca stimulate dall'esito della ricerca in questione → (b') ecc.

---

Questo Schema dovrebbe valere sia per le ricerche scientifiche di argomento umanistico e letterario sia per quelle che hanno per argomento le scienze della natura.

Si noti anche il carattere circolare e ricorsivo della procedura descritta nello Schema 1 la conclusione (coronata da successo) di un'ipotesi di ricerca (a) è tale da provocare altre ipotesi di ricerca (a'), cioè l'inizio di una nuova, diversa e più avanzata, ricerca. Osservando ancora il nostro Schema 1, rileviamo che è possibile distinguere nella procedura descritta una prima fase interna e, per dir così, «privata», da una seconda fase «pubblica», cioè offerta alla discussione, al giudizio (e poi all'eventuale applicazione) da parte della comunità: la prima fase «privata» consisterebbe nei punti (a)-(f), la seconda fase «pubblica» consisterebbe nei punti (h)-(l); il punto (g), rappresentato dalla pubblicazione, costituisce proprio il cruciale momento di snodo fra le due fasi.

Il problema del *conflitto* fra la diffusione della nuova conoscenza e la sua appropriazione esclusiva da parte dello scopritore (richiamato *supra*, cfr. § 2, punto 4) viene affrontato e risolto, nella procedura tradizionale, per l'appunto dalla pubblicazione: chi «pubblica» mette a disposizione della comunità la *propria* scoperta, cioè l'innovazione che reca comunque il suo nome e gli viene attribuita; insomma proprio e solo attraverso la pubblicazione un lavoro viene legato al suo Autore. La pubblicazione coincide così con la legalizzazione. Paradossalmente: solo pubblicando (mettendo in pubblico) ci si ap-propria.

10. Nella procedura di ricerca segnata dall'informatica la pubblicazione non è più centrale.

Ma noi sappiamo che l'informatica modifica anzitutto i processi euristici, non solo i risultati, dunque sconvolge le procedure, le innova radicalmente e sostanzialmente.

Una ricerca *segnata* dall'informatica, e in particolare appoggiata alla modalità *propriamente informatica* della rete Internet, modifica sostanzialmente la procedura di ricerca descritta nello Schema 1.

La ricerca viene infatti ormai annunciata pubblicamente (tramite la rete) già nel momento (b) «Riflessione-Discussione», che precede l'impostazione della ricerca vera e propria e il suo progetto; e anche i momenti (e) «Svolgimento della ricerca» ed (f) «Conseguimento dei primi risultati (verifica dell'ipotesi della ricerca)» possono essere in teoria (ed effettivamente, in molti casi, già sono in pratica) esposti nella rete ben *prima* della pubblicazione vera e propria. Quest'ultima potrebbe, al limite, non avere mai luogo, e il contributo conoscitivo della ricerca proposta potrebbe consistere proprio, per ipotesi estrema, nell'aver provocato una discussione capace di sconsigliare, o escludere, la messa in atto di una determinata linea ricerca.

Riguardando lo Schema 1 (che descrive la procedura che abbiamo definito «tradizionale») sembra anzi che *tutti* i suoi momenti siano sconvolti, soprattutto per quanto concerne la loro *sequenzialità*, logica e cronologica. Ad esempio ora si può sottoporre alla discussione pubblica (e, in questo senso, «pubblicare»: punto g) una ricerca ben prima che essa sia conclusa fornendo risultati (punto f), e anzi anche prima che essa sia compiutamente progettata (punto c), così come si può discutere pubblicamente, e collettivamente, anche una semplice idea o ipotesi (punto a, o b), provocando una discussione collettiva all'interno della comunità scientifica (punto h) che sia, per così dire, preliminare, e così via.

Non sembra assurdo poter ipotizzare uno Schema 2 (o un altro altrettanto immaginario) in cui gli stessi undici momenti considerati nello Schema 1 si presentano però in una *successione* del tutto diversa, e in cui soprattutto il momento (g) pubblicazione (ora, si noti, divenuto *pubblicazione in rete*) non conclude e corona la ricerca, e neppure segna il passaggio dalla prima fase «privata» alla seconda fase «pubblica», ma piuttosto accompagna, ricorsivamente, *ogni passaggio* della nuova procedura.

Si avrebbero in tal caso un numero *n* di atti di (g) pubblicazione: (g'), (g''), (g'''), ..., sempre accompagnati, significativamente, da un corrispon-

dente numero di (*h*), cioè di discussioni all'interno della comunità scientifica e da una continua modifica-riformulazione dell'ipotesi (*a*)-(*b*) di partenza.

Nello Schema 2 qui di seguito proposto i «tornanti» di questo andamento ricorsivo (ciascuno segnato da una diversa pubblicazione) sono cinque (indicati in numeri romani), ma nella realtà essi possono essere evidentemente di numero variabilissimo: la cosa importante è che in questo caso non si tratta più di una schema lineare bensì di una sorta di *spirale*, in cui ogni tornante è segnato da *una* pubblicazione, e forse nessuno più da *la* pubblicazione.

SCHEMA 2. *La procedura di ricerca segnata dall'informatica (e la non-centralità della pubblicazione).*

---

(I)

(a) Prima ideazione, Ipotesi di ricerca → (b) Riflessione (personale), Discussione (privata) → (g) (Prima) Pubblicazione in rete dell'ipotesi della ricerca →

(II)

(h) Discussione all'interno della comunità scientifica → (a') Seconda ideazione, Ri-formulazione dell'ipotesi di ricerca → (b') Seconda riflessione (personale), Discussione (privata) → (g') (Seconda) Pubblicazione in rete delle risultanze di (h)-(b') a proposito della ricerca →

(III)

(c) Progettazione (formulazione dell'ipotesi della ricerca, da verificare) → (g'') Terza Pubblicazione in rete del progetto di ricerca →

(IV)

(h') Discussione all'interno della comunità scientifica → (i) Verifica (iniziale) della validità, accettazione pubblica dell'interesse dei (possibili) risultati della ricerca → (d) Messa in opera della ricerca → (e) Svolgimento (iniziale) della ricerca → (g''') (Quarta) Pubblicazione in rete dell'andamento della ricerca →

(V)

(h'') Discussione all'interno della comunità scientifica → (a'') Terza ideazione. Ri-formulazione dell'ipotesi in base al procedere della ricerca e alla discussione relativa → (e') Svolgimento (fase avanzata) della ricerca → (f) Conseguimento dei primi risultati (verifica dell'ipotesi della ricerca) → (g''') (Quinta) Pubblicazione in rete dei risultati della ricerca → (h''') Discussione all'interno della comunità scientifica → (i'') Verifica (definitiva) della validità, accettazione pubblica dei risultati della ricerca → (l) Applicazione → (m) Produzione di altre ipotesi di ricerca stimolate dall'esito della ricerca in questione → (= a''') Ulteriori ideazioni, nuove ipotesi → ecc.

---

Sembra in conclusione che, nella nuova procedura segnata dall'uso dell'informatica, la pubblicazione non rivesta più *alcun* carattere di centralità, al contrario di quanto accadeva nella procedura guttemberghiana o pre-informatica.

### *11. La proposta di un Centro di Certificazione della Ricerca Scientifica in Rete.*

Naturalmente anche in questo caso (come sempre) abbiamo di fronte due strade: o cercare di ridurre il nuovo nelle misure del vecchio e versare il vino nuovo negli otri vecchi, oppure, al contrario, scegliere di *prendere sul serio l'innovazione* e cercare di trarne tutte le conseguenze.

Percorrere la prima strada significa, ad esempio, costringere anche i prodotti informatici a imitare (ma ora del tutto irrazionalmente) la configurazione tipica del libro guttemberghiano<sup>27</sup>, in altre parole far finta che un prodotto di ricerca informatico (per ipotesi multimediale e interattivo) sia sempre e solo ... un libro, un buon vecchio libro, che si può dunque stampare e, ciò che più conta, «depositare», magari a cura dello stesso Autore in cinque copie presso la Prefettura. Si tratta certo di una scelta assolutamente lecita, ma che vista con occhio esterno, somiglia troppo alla scelta (che pure a suo tempo fu fatta da parecchi) di continuare a copiare a mano i libri anche dopo l'invenzione della stampa.

Sono sempre possibili, beninteso, delle soluzioni di compromesso, peraltro quasi obbligatorie allo stato delle cose: per esempio è un compromesso ragionevole il sistema dei *preprint* depositati negli *open archives*, ovvero pubblicare *on line* i lavori scientifici, in siti specializzati o su riviste<sup>28</sup>, ma periodicamente anche stamparli e depositarli nei cinque fatidici esemplari previsti dalla legge, essenzialmente per fini concorsuali.

<sup>27</sup> In particolare certa produzione editoriale di CD-Rom sembra significare la volontà di seguire una tale strada: penso alla scelta di produrre grandi enciclopedie in CD-Rom, che però vengono venduti solo come accompagnamento della versione cartacea della stessa Enciclopedia, e non in *sostituzione* di questa. Tale scelta, del tutto irrazionale, è evidentemente motivata dall'esigenza di continuare a vendere un oggetto cartaceo che la stessa presenza dei CD-Rom ha reso del tutto obsoleta (per costo, ingombro, difficoltà di aggiornamento, ecc.).

<sup>28</sup> Cfr. ad esempio, in ambito scientifico, <http://jhep.sissa.it>, e in ambito umanistico «Reti medievali».

Francamente sembra più interessante e ragionevole cercare di percorrere con decisione la seconda strada, cioè cercare di prendere seriamente in considerazione l'innovazione informatica e anzi operare affinché essa si possa svolgere *iuxta propria principia*.

Questa scelta implica allora la necessità di cercare di capire (ed eventualmente di elaborare) le regole *proprie* della modalità informatica di pubblicazione, cioè di produzione e trasmissione delle conoscenze innovative; ma a partire dunque da ciò che Orlandi definisce «le caratteristiche intrinseche alle tecnologie elettroniche»<sup>29</sup>, e non a prescindere da queste. Fra tali caratteristiche, di cui tener conto prioritariamente, Orlandi elenca:

«... flessibilità, ma anche instabilità spazio-temporale, ed incertezza di attribuzione d'autore etc. ...».

Orlandi insinua, a mio parere del tutto fondatamente, che tali caratteristiche possano risultare contraddittorie con:

«... i requisiti di certezza e stabilità che si pretendono dalle pubblicazioni scientifiche»<sup>30</sup>.

Per «flessibilità» e per «instabilità spazio-temporale» si intende verosimilmente la possibilità di modificare in ogni momento dei testi o dei documenti residenti nella rete, mentre «l'incertezza di attribuzione d'autore» allude alla possibilità di non firmare i lavori, o di firmare collettivamente, o con pseudonimi, o anche (al limite) di firmare lavori non propri.

Ma a questo punto del nostro ragionamento il discorso può e deve tornare a farsi *tecnico*, e dunque del tutto *risolvibile* (e proprio grazie alle potenzialità della macchina informatica).

Non sembra infatti difficile ipotizzare la creazione di un luogo pubblico, e più precisamente di un *Centro di Certificazione della Ricerca Scientifica in Rete*, in grado di *certificare* (su richiesta degli Autori interessati) l'esistenza nella rete, ad una data X di un oggetto digitale Y<sup>31</sup>. Un tale Centro dovrebbe essere naturalmente autonomo da qualsiasi Governo, ma essere finanziato

<sup>29</sup> Dal documento che impostava la ricerca del Centro Linceo Interdisciplinare «Beniamino Segre» citata *supra* nella nota 1: Orlandi (2002).

<sup>30</sup> Orlandi (2002).

<sup>31</sup> Da ricordare che già nel 1994 è stata creata a Ginevra «Interdeposit», una federazione che riunisce le organizzazioni interessate al deposito delle opere digitalizzate; Interdeposit ha creato un sistema internazionale di identificazione (IDDN: Inter Deposit Digital Number) con il quale si può effettuare il deposito.

dallo Stato, e dovrebbe godere della stima e della fiducia incondizionata della comunità scientifica: sarebbe bello che esso potesse essere ospitato e garantito da un'istituzione prestigiosa come l'Accademia dei Lincei.

Il prodotto intellettuale certificato dal Centro (quale che sia la sua forma) dovrebbe, naturalmente, anche essere ricopiato e conservato presso il medesimo Centro, in una costituenda «Banca» delle ricerche, con riferimento alla data di esposizione nella rete Internet.

Un semplice numero di certificazione (rilasciato dal Centro al momento della certificazione-copiatura del prodotto) sarebbe sufficiente (e necessario) per poter esibire, anche in sede concorsuale, il lavoro scientifico in questione<sup>32</sup>.

L'importante dal nostro punto di vista è che tale Certificazione Scientifica non contraddirebbe affatto la mobilità e la flessibilità proprie della modalità informatica, e della rete Internet in particolare, ma anzi servirebbe a *valorizzare* proprio tali caratteristiche. Ad es. essa potrebbe (dovrebbe) essere ripetuta *molte volte*, in corrispondenza dell'evoluzione del lavoro intellettuale *in progress*<sup>33</sup>.

Ne risulterebbe così, come prezioso «prodotto secondario» della Certificazione, anche la possibilità di ricostruire, *a posteriori* o mentre si svolge, l'andamento di una linea di ricerca, la sua evoluzione, i suoi tempi, ecc. Si eviterebbe così la possibilità, oggi del tutto presente, di una dispersione irreparabile di attività di ricerca non concluse (oppure semplicemente gestite da supporti informatici ormai obsoleti, ecc.). Un altro «prodotto secondario» di straordinaria portata ed utilità sarebbe il poter disporre di uno «stato dell'arte» tendenzialmente completo e continuamente aggiornato, in grado di rispondere finalmente alla domanda: «chi attualmente sta lavorando su che cosa?», evitando così non solo i plagii e i furti ma anche le ripetizioni e gli sprechi.

L'importante è capire che anche la modalità propriamente informatica della pubblicazione scientifica (essenzialmente connessa all'uso della rete) potrebbe svolgersi secondo le proprie caratteristiche e peculiarità (peraltro straordinarie) senza essere costretta ad attardarsi a mimare, insensatamente,

<sup>32</sup> L'esistenza di tali dispositivi di identificazione comporta già di per sé la necessità per chiunque si riferisca alla ricerca in questione (a cominciare dal suo stesso Autore) di attenersi onestamente a quanto è in ogni momento verificabile da parte di chiunque.

<sup>33</sup> Nel caso di banche-dati (testuali o no), che per loro natura siano in via di accrescimento nel tempo, si potrebbe pensare ad un ritmo periodico di Certificazione dello stato di avanzamento.

le modalità di Gutenberg. Ci sembra tuttavia di aver argomentato che questo passaggio ad una nuova modalità di pubblicazione non implica affatto la rinuncia alle caratteristiche di scientificità, di responsabilità personale e di rigore su cui la comunità scientifica fonda la propria organizzazione interna e la propria deontologia.

#### BIBLIOGRAFIA

- BLISSETT L., 1999. *Q*. Einaudi, Torino.
- BONCOMPAIN J., 2002. *La Révolution des auteurs. Naissance de la propriété intellectuelle (1773-1815)*. Fayard, Paris.
- BORGHI M., 2003. *La manifattura del pensiero. Diritti d'autore e mercato delle lettere in Italia (1801-1865)*. Franco Angeli, Milano.
- CANFORA L., 2002. *Il copista come autore*. Sellerio, Palermo.
- Filosofia e Diritto nell'era di Internet* (Atti della giornata di studio, Roma, Università di "Tor Vergata", 21 novembre 2002). Aracne, Roma 2003.
- GIMENO BLAY F.M., *Écrire au Moyen Âge*. In: *Anthropologie et histoire de la mutation graphique et de ses effets. Autour d'Armando Petrucci* (École française de Rome, 30 janvier - 1<sup>er</sup> février 2003), in corso di stampa.
- HARDIN G., 1968. *The Tragedy of the Commons*. *Science*, 162: 1243-1248.
- LAIDI Z., 2003. *La propriété intellectuelle à l'âge de l'économie du savoir*. *Esprit*, 299 (novembre): 116-131.
- LEFEBVRE G., 1970. *La Rivoluzione francese*. Einaudi, Torino.
- MORDENTI R., 2002. *Informatica e critica dei testi*. Bulzoni, Roma.
- MOSCATI L., 2001. *Sul diritto d'autore fra codice e leggi speciali*. In: *Iuris Vincula*. Studi in onore di Mario Talamanca, V: 497-527.
- ORLANDI T., 2002. *Proposta di una ricerca riguardante il settore informatico da condurre presso il Centro Linceo Interdisciplinare*. Roma, 29 marzo (dattiloscritto).
- PALAZZOLO M.I., 2003. *Gli incerti destini del diritto d'autore*. *La Fabbrica del Libro*, Bollettino di storia dell'editoria in Italia, a. XIX (1): 2-4.
- PELLEN R., 1999. *Propriété intellectuelle et société de l'information*. *Le Médiéviste et l'Ordinateur*, 38 (hiver): 44-52.
- SORDI P., 1999. *Internet e la tutela del diritto d'autore*. *Economia e diritto nel terziario*, 2: 499-526.
- UBBELOHDE A.R.J.P., 1964. *Il passaggio dall'empirismo alla scienza come base della tecnologia*. In: C. SINGER, E.J. HOLMYARD, A.R. HALL, T.I. WILLIAMS (eds.), *Storia della tecnologia*. Vol. 4, Boringhieri, Torino: 677-695.



ANDREW WILSON\*

## ACCESS ACROSS TIME: HOW THE NATIONAL ARCHIVES OF AUSTRALIA PRESERVES DIGITAL RECORDS

### INTRODUCTION

The National Archives is the agency responsible for records and archives in the Australian Federal Government. We are a national organisation with offices in every state and territory, and with total records holdings of approximately 450 shelf kilometres. Much of the significant developmental work being done in the NAA at the moment is on projects focussed on preserving our collection and, in particular, the formats of records that we regard as being at risk. One such project is the so called «Agency to Researcher» project (AtoR), aimed at developing and implementing a viable approach to the problem of the long-term preservation of records that are «born digital».

But, before I discuss the National Archives preservation work I want to make sure that the context of the project I am discussing is clear. Firstly, some basic definitions are necessary, but I do only have two:

1. records: recorded information created or received and maintained by an organisation in the transaction of business; and
2. digital records: records created in digital form and processed by computers. Note that this does not include systems or working applications.

Additionally, I must make clear that the records received by the National Archives, all of which can be regarded as significant cultural heritage resources, are in a multitude of formats. We have significant holdings of photographs, posters, maps, architectural drawings, films, playscripts, musical

\* Preservation National Archives of Australia; contact: [cornel.platzer@naa.gov.au](mailto:cornel.platzer@naa.gov.au)

N.B.: This paper describes the NAA approach as it was in late 2003. In the time since then some changes have been made to the details of Implementation. However, the conceptual approach remains unchanged.

scores and sound recordings, and a small number of digital records. The collection's main strength is material created since Australian Federation in 1901, by agencies of the Australian Government, such as departments, royal commissions, statutory authorities, military units, security, intelligence and law enforcement agencies, diplomatic posts and foreign relations. These are complemented by substantial collections of 19th-century records. We also hold papers of Governors-General, Prime Ministers, and Ministers of State<sup>1</sup>.

#### PROBLEMS WITH INFORMATION TECHNOLOGY

The widespread use of computer technology, invading all aspects of life and particularly affecting the manner in which organisations carry out their business, poses special problems for the long-term retention and use of records. These issues are no doubt familiar to most people but I discuss them here to set the proper context for the work I intend to describe below.

Firstly, there is the problem of short technology cycles. By this I refer to the rapid turnover in computer chip capabilities and in the architecture of supporting hardware technologies. A consequence of this is that hardware generations are generally regarded now days as lasting about 18 months. The second issue is the rapid change in software applications design, including those of operating systems, especially in the specifications of the data formats created by applications. Together these two issues of technological obsolescence cause increasing difficulties for those of us interested in keeping records useable over time. Backward compatibility seems to have become a less and less implemented feature in new software (and it never was for hardware) resulting in an inability of new software to understand the data formats created by earlier versions of the same software. So, for example, the current version of Microsoft Word for Windows cannot read documents created in Word 3.1 on the DOS operating system.

A third problem is that of the short lifetime of digital media. Despite the hyperbole of advertisers there is no digital medium that is suitable for archival retention and storage. Copying all one's precious digitised photographs

<sup>1</sup> This paragraph is taken from the National Archives of Australia website (slightly amended), at: [http://www.naa.gov.au/the\\_collection/the\\_collection.html](http://www.naa.gov.au/the_collection/the_collection.html)

to CDs or DVDs and expecting them to survive for 20 years is a risky undertaking with no guarantee of success.

The consequence of these three major issues of digital technology is that an archival organisation, for whom permanent retention does not mean 10 or 20 years but forever, cannot retain digital records in their original data formats on any of the currently available offline storage media and expect them to be accessible in the long-term (*i.e.* in excess of 30 years).

### APPROACHING DIGITAL RECORDS

Normally, a user of records in traditional formats (such as paper) directly experiences the record through its source object, *i.e.* the physical object itself, be it a letter, map, black and white photograph, or an object such as a model of a building. In this tradition it is only necessary to preserve the object in order to preserve the record.

But this paradigm does not hold for digital records because digital records are *mediated* by technology. In the digital context we cannot directly experience the source object because it exists only as a bitstream (an ordered sequence of 1s and 0s). In order to make sense of the data we need to use a computer, an operating system and a software application, what we refer to as a «process». Digital records thus cease to be physical objects and are, instead, the result of the mediation of technology and data. The experience of the object only lasts for as long as the technology and data interact. As a result, each viewing of a record is a new «original copy» of itself – two people can view the same record on their computers at the same time and will experience equivalent «performances» of that record<sup>2</sup>.

For the National Archives then, a digital record is a performance, and this is the core of our approach to preserving digital records. In this view of digital records it is meaningless to talk about preserving the source object. Instead, we need to preserve the performance in order to preserve the record. So our approach focuses on the interaction between data and technology that

<sup>2</sup> H. HESLOP, S. DAVIS, A. WILSON, *An Approach to the Preservation of Digital Records*. National Archives of Australia, December 2002, 8; [http://www.naa.gov.au/recordkeeping/er/digital\\_preservation/summary.html](http://www.naa.gov.au/recordkeeping/er/digital_preservation/summary.html)

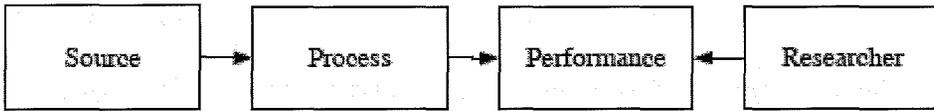


Fig. 1. – The Digital Records Performance Model.

interprets the data, determining what is archivally important about that interaction (the *essence* of the record), and then creating and maintaining the ability to repeat the essence on demand in a sustainable manner.

You might ask «Why not just keep source and process?». As I have briefly mentioned above, there are major problems with retaining the process. Processes become obsolete, so, for example, DOS was superseded by Windows 3.1 which was superseded by Windows 95 which was superseded by Windows NT; now, DOS is an obsolete operating system. This is marketplace obsolescence, driven by the commercial realities of software companies. There is the problem of incompatible operating systems, so applications and most data formats designed for Windows are not interchangeable with the Macintosh operating system or \*nix operating systems. Even with the same family of operating systems there are problems of interchangeability, such as that between Windows 3.1 and Windows 95 or Windows NT. The final issue with keeping processes is that they are generally proprietary and access is governed by licence, not purchase. Licences do not last forever and can, theoretically, be revoked or withdrawn.

To summarise, digital records are different from traditional records because they are *mediated* by technology. Consequently, preserving the object is meaningless, and it becomes necessary to focus on: the interaction between the data and the technology that interprets the data; determining what is archivally important about that interaction (the *essence* of the record); and creating and maintaining the ability to repeat the essence of the record on demand in a sustainable manner.

#### THE NAA APPROACH

How do we get past the problems associated with trying to preserve processes? We could keep a master copy of every source we accept into custody

– we might call this *passive* access. Researchers would get the bitstreams, the 1s and 0s that make up the record. But this version of access is likely to be insufficient for most people. An alternative would be to intervene actively in the preservation process to enable the recreation of the performance. In this approach to access we could replace the original source and process with new versions of both and enable *active* access to the essence of the original performance. This approach, which has been adopted by the National Archives, is based on extensive experience with preserving audio-visual records. What we learned from this experience is that preservation effort is needed at the beginning of the process; that there need to be minimal requirements for researchers; that future users need to know about how we preserved the records; and that it is necessary to control the look and feel of the original in order to retain what is essential to the record.

The National Archives approach, then, is based on the active intervention model described above. In summary, our approach has 5 components, which consist of the following:

- 1) define the «essence» of the record;
- 2) acquire our own processes to suit this essence;
- 3) normalise the source to suit our processes;
- 4) maintain the source and processes over time;
- 5) recreate the performance as we need.

#### THE 5 COMPONENTS OF THE NAA APPROACH IN DETAIL

##### 1. *Define Essence.*

Firstly, we need to define essence. To define essence let me say briefly that it is what we want to preserve out of each record's performance, *i.e.* those aspects of the original that are essential to the record's value. Thus, we omit those aspects of the original that are incidental to the record's value. Essence is not static and differs from record format to record format. The most basic question that has to be asked is «how important are the various elements of a performance?». For example, are the different views on a computer screen of the same record significant? Is it necessary to capture the colours of the user interface, or the fonts, formatting, and dimensions of objects. Next we need

to question if it is important or necessary to retain embedded information? By this I mean such things as email headers (showing time sent and received) and formulas in spreadsheets.

Sometimes just the content needs to be preserved and we can forget about or ignore «look and feel» and structural information. Importantly, I want to stress that we have different rules for different record genres and different creation contexts. So the final determination of essence results in a different outcome for every different data format that we will be dealing with.

## 2. *Acquire our own processes.*

By this we mean that we do not intend to try to preserve the original process(es) at all. Our intention is to design and acquire our own processes for capturing and reproducing performances, using an open standards approach. We favour open standards for digital preservation because:

- they are not owned by anyone;
- we can adopt multiple implementations (or make our own implementation);
- using open standards results in more practical and effective interchange of data; and
- our standards platform of choice is XML, an open standard itself.

Significantly for the National Archives, the adoption of this approach to develop or acquire our own processes means that as an organisation we commit to maintaining those processes over time (*i.e.* in excess of 20 years).

## 3. *Normalise the source.*

«Normalisation» is our own term for the one-off migration process by which records in source data formats are converted to different formats that suit our preservation platform, currently eXtensible Markup Language. So, when we «normalise» records we migrate transferred material into an XML data format that is designed and endorsed by us. This data format needs to carry the essence of the original performance, and these normalised XML documents become the Preservation Master. The National Archives, consequently, has made a commitment to always being able to process the Preservation Masters using our preservation platform. The preservation platform,

which does the conversion to XML, is a software tool currently under development by the National Archives of Australia, using the open source software development approach. More information about the tool, called Xena, is provided later in this paper.

#### 4. *Maintain the source and process.*

Having normalised the source records our next job is to maintain the source and the process, both of which are needed for ultimately recreating the performance. To maintain the normalised data objects we put everything into a digital repository, with redundant copies on media which will be regularly refreshed. Our current approach is to store two copies on mirrored RAID storage, with a third copy on digital tape. To maintain the our process we need to ensure that as technology changes we acquire a new preservation processing platform to replace existing platforms as they become obsolete. We will do this either by selecting a suitable commercial vendor product, or by building our own platform as we are doing currently. In addition, we intend to maintain the reliability and integrity of our master copies by not allowing researchers to use the «master» material. They will only be give access to copies and will have no direct access in any form to the digital repository.

#### 5. *Recreate the performance.*

The final part of the whole process is to ensure that we can recreate the performances we are trying to preserve. The test of any preservation program is in the two areas of accessibility and authenticity, and both of these are crucial to demonstrate the viability and reliability of the preservation approach. The National Archives approach has been especially concerned with incorporating mechanisms to ensure both the accessibility and authenticity of the digital records we process through our preservation process, since our mandated role is to preserve and provide access to the archival resources of the Australian Government<sup>3</sup>.

How do we ensure that our process can demonstrate both of these characteristics and appropriately recreate the performance we have been attempting to maintain and preserve? To achieve these goals we make copies of the

<sup>3</sup> Commonwealth of Australia, Archives Act 1983, s. 5 (a) and (j).

Preservation Masters available to users, so that they can recreate the essence of the performance in a browser, or other familiar desktop computer application. Alternatively, because our current approach is based on XML, we could transform the Preservation Masters into other, accessible, formats such as web pages for casual access, as long as the other formats can carry the essence of the records. A second strategy is to keep a complete record of all the actions that have been applied to the source record from time of transfer, *i.e.* preservation metadata, and to preserve the original bitstreams. This allows us to not only give users a complete audit trail for each record, but also ensures that we can give users access to the bitstream of the original data objects, although we are not able to guarantee access to the bitstreams.

#### IMPLEMENTATION IN PRACTICE

I now want to explain in some detail the way the National Archives has actually implemented the approach discussed in the preceding pages.

##### *NAA Implementation concepts.*

In essence, there are three main principles which underlie the way the National Archives has implemented its approach to preserving digital records. The approach:

- follows the Open Archival Information System framework<sup>4</sup>;
- is a non-proprietary, open source solution; and
- is based on the extensible markup language (xml)<sup>5</sup>.

I would like to emphasise at this point that the National Archives has managed the implementation of this research and development project with a primary philosophy of ensuring that the organisation is not tied to any particular software or hardware vendor for its solution. Our software tools are deliberately developed as cross-platform tools using as much open source sof-

<sup>4</sup> See <http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf>

<sup>5</sup> <http://www.w3.org/TR/2004/REC-xml-20040204/>

ware as possible, and all hardware purchases are of commercial off-the-shelf products. This guarantees that the Archives is not dependent on a single supplier for any part of its implementation and can change suppliers as dictated by business or financial requirements.

### *Implementation.*

The physical implementation of the NAA approach consists of three separate components:

1. Quarantine;
2. Preservation; and
3. Storage.

Each of these components is physically separated from each of the others, as well as being isolated from all other networks managed by the National Archives, including our major business system, RecordSearch. Access to the hardware is restricted to digital preservation staff only and transfer of data between components is by what is popularly referred to as «sneaker net», *i.e.* devices are physically carried from one server to the next.

#### 1. *Quarantine.*

The quarantine facility is the first step in the preservation process. The function of this component is to carry out initial checking of the records received, and to ensure that no viruses are transferred through the preservation process to the digital repository. Digital records are received from government agencies with substantial documentation (recordkeeping metadata) on some form of media, and we are very catholic in the range of transfer media we can accept – floppy discs, CDs, DVDs, digital tape (various formats), hard drives (various protocols *e.g.* Firewire, USB, SCSI), even laptops. The data objects on the transfer media are copied to what we call a «carrying device», currently these are various size external hard disk drives. Once copied, the objects are checked for viruses, and checksums are compared to make sure the copied objects match the source objects. Providing the copied records pass all these checks the disk drives are then left in secure repository storage for a period of 28 days.

## 2. *Preservation.*

Following the 28 day quarantine, the carrying devices are rechecked for viruses and, if uninfected, carried across to the second component of our process, the preservation facility. It is in this component that digital records are «normalised», through the use of our preservation processing platform Xena, which was developed in-house. This software platform creates two new versions of each source record: a bitstream wrapped in XML (metadata); and the normalised version which is an XML equivalent, as discussed above. Each new version of the data object has a new checksum calculated for it, and it is written out to a different carrying device, currently an external hard disk drive, for transfer to the digital repository.

## 3. *Digital Repository (Storage).*

In this final component of the implementation, the data objects are copied from the hard disk to digital repository itself. Our current practice is to make a number of copies, for integrity, backup, and redundancy purposes. Each object is copied separately onto two different RAID disk stores – we make each copy separately so that errors are not duplicated across the different RAID stores. A third copy is made on digital tape and stored offsite, in another Archives facility.

When records are required for retrieval by the transferring agency or for reference use, they are copied out to a transport medium and given to the agency or the National Archives reference staff. No direct access is allowed to the repository and only digital preservation staff can physically copy records from the repository. In this way we hope to ensure the integrity and authenticity of the data objects maintained in the repository, and to keep them accessible over time.

## XENA

To finish this paper I will describe briefly the preservation software platform that the National Archives has been developing in-house since early 2002. The software application we are developing is now known, thanks to our public relations people, as Xena, for **X**ML **E**lectronic **N**ormalising of **A**rchives. Xena is a file-based application, developed in Java and Swing, and

requires a version of Java more recent than version 1.4 to function. It is a modular application but packaged as an executable «jar» file (java runtime). Although there has been no public release of the application to date, the source code and the executable for Xena are available for download from the SourceForge web site because NAA is developing the tool as an open source software project<sup>6</sup>. Xena is made available under the terms of the Gnu General Public Licence (GPL)<sup>7</sup>.

I mentioned that Xena is modular – it consists of a core module and a number of «plug-ins». The Xena core looks after the graphical user interface, management of the plug-ins, and some generic validation. The plug-in modules are created for the identified data types that are to be processed (normalised). Currently, there are plug-ins which handle processing of plaintext, HTML, office productivity formats (*e.g.* Microsoft Word, Excel and PowerPoint), datasets, various image formats (jpeg, TIFF, gif, png), and some email formats. Each plug-in consists of a guesser component (for guessing the input file format), input format type components, a normalised format type, one or more normalisation modules, one or more view components, sorting functionality, validation functionality, printing functionality, and the GUI interaction methods.

In summary, Xena is an open source preservation application developed by the National Archives of Australia. It is written in Java and will operate in all major operating system environments: Unix, Linux, Windows (98, NT, 2000, XP), and Mac OSX. Xena can be used to normalise data objects and also to view those normalised data objects.

## CONCLUSION

The Agency to Researcher Project (AtoR) represents a significant shift in the National Archives' approach to digital preservation. For most of the 1990s, the Archives followed a distributed custody model, which meant that digital records of archival (*i.e.* long-term) value remained in agency custody except in special circumstances. In March 2000, the Archives moved to a cu-

<sup>6</sup> See <http://sourceforge.net/projects/xena/>

<sup>7</sup> Available from: <http://www.opensource.org/licenses/gpl-license.php>

stodial model, which means it now takes into custody all digital records that are required to be retained as National Archives under approved disposal authorities. The Agency to Researcher Digital Preservation Project was initiated as a result of this policy change to ensure that digital records of long-term value will remain accessible for use over time.

The challenges of digital preservation affect all major public archives, both in Australia and around the world. In the same way that the National Archives of Australia has developed recordkeeping standards to address government recordkeeping issues, it is committed to providing innovative solutions to the problems of digital preservation. We expect that the solutions developed will be of value to many other archival institutions within Australia and overseas. Ultimately, this work will preserve millions of Commonwealth records that will be of significant value to future generations. The outcomes of the AtoR project provide a very real encouragement to others seeking to find a viable approach to the problem of preserving digital data<sup>8</sup>.

<sup>8</sup> The project is more fully described in the NAA Green Paper, available on the National Archives website at: [http://www.naa.gov.au/recordkeeping/er/digital\\_preservation/summary.html](http://www.naa.gov.au/recordkeeping/er/digital_preservation/summary.html). Information about the various XML schema used for normalising data objects can be found at: [http://www.naa.gov.au/recordkeeping/preservation/digital/xml\\_data\\_formats.html](http://www.naa.gov.au/recordkeeping/preservation/digital/xml_data_formats.html)

STEPHAN HEUSCHER\*

## WORKFLOW MODELING LANGUAGE EVALUATION FOR AN ARCHIVAL ENVIRONMENT

ABSTRACT. — Although archives today are more and more urged to preserve digital data, the processes in archives are still predominantly bound to the handling of traditional paper-based records and strongly depend on implicit knowledge, *i.e.*, the processes through which records are being transferred to and stored and maintained in the archives are only roughly sketched and rarely explicitly documented (a few exceptions like, *e.g.*, paper de-acidation or climate conditions in storerooms, aside). This is not a problem with paperbound records, because the goal of preserving the records' integrity, originality, and authenticity, is achieved by keeping carrier or substrate of the information (*e.g.* paper, polyester film etc.) physically and chemically as inert as possible. As digital records do not rely on a specific carrier or substrate whose intactness or inertness could guarantee integrity, originality, and authenticity, but only on the information content itself as well as the description of the formal structural aspects of its coding (like, *e.g.*, data formats, data semantics, rendering techniques etc.), which are subject to change. These changes must be unambiguously defined, documented, and included into the description of the record. To allow for this, the processes first have to be described. This paper addresses the question of how to best describe an archival process by taking into account the main archival paradigms outlined. We conclude that a language having high-level Petri net semantics best meets the requirements of an archival workflow language.

### 1. INTRODUCTION

Workflows are ubiquitous in our days. They are used to describe and formalize the business processes to make them repeatable, auditable, and automatically executable. Archives have always had processes when receiving, handling, restoring, or making available records and documents given into

\* Ikeep Ltd. - Digital Archives Services - Berner Technopark - Morgenstrasse, 129 - CH-3018 BERN (Svizzera). [stephan.heuscher@ikeep.com](mailto:stephan.heuscher@ikeep.com)

their custody. The records and documents created in these processes should be kept and conserved like other holdings of the archives to enable a future researcher to better recreate the original form and environment of the records and documents (for more information on the purpose of archives, refer to Keller, 2004).

While the archival processes were conducted using paper-based reporting, because the processes and personnel rarely changed and holdings needed little attention, there was no apparent need to formalize these processes, although there were some attempts (Ministerrat, 1984). As more and more records and documents are born-digital and archives integrate them into their holdings, change has become a constant. The processes depend more on the technology used and they change frequently. This might be due to the rapid advances in processing, storage, and communication inherent to information technology or even merely an upgrade of a system component. At the same time, the holdings have become heavily dependent on their infrastructure and they themselves need to be treated (*i.e.* checked, copied, converted, migrated etc.) more often to adapt to the aforementioned changes. Preservation in the earlier sense meant that the records or documents should be kept as original as possible on their original carrier, while the approach when archiving digital records and documents is to keep them intelligible and accessible using the current technologies. The «computer museum» approach featuring the collection of old hardware and software to preserve documents and records in their original form and environment, has been dismissed for obvious financial, space, personnel, and maintenance reasons.

Digital documents and records are fragile (Rothenberg, 1995: «digital information lasts forever – or five years, whichever comes first»). One flipped bit could render useless a document and all documents depending on it. The fragility is not only one of bits, but also lies in the need for the bits to be interpreted to make them meaningful and understandable to humans. It is impossible for a single human today, to comprehend all the transformations of digital data. *E.g.* the task of «simple» home computer displaying a single character stored on its harddisk cannot be described in this paper. To show the complexity of the task, we will describe just a few steps in the processing chain. First the representation of a character is read from the harddisk, processed by the controller on the harddisk, sent to the controller on the motherboard, then sent over the PCI (Peripheral Component Interconnect) bus into the main memory, where it is rendered (*e.g.* to a raster format) and sent to the

graphics card. Then it is transformed into analog signals, which are digitized and sent to the LCD (Liquid Crystal Display) controller so that the character is displayed on the monitor. We only described the process in low-level, overly simplified steps to show that even in these basics over-strain the technical knowledge of more than 99% of the people. In the digital world, it is not possible to reason by physical analogy, as *e.g.* on a shellac record, the depth of the groove is equivalent to the amplitude of the membrane of a loudspeaker<sup>1</sup>.

An additional complexity lies in the need of the computer to interpret or render digital information to make it accessible to humans. The specification of these interpretations of digital files and their formats and these implementations can be complex, ambiguous, undefined, or even erroneous. This is the reason why some archives impose restrictions on the type of files they accept. The archives try to keep the complexity at a minimal (manageable) level because it is up to them to keep the contents of the transferred files understandable and usable. Some file formats allow for many forms of content, *e.g.* the portable document format (PDF) can contain scripts, images, sound, and video, which makes files of this format very difficult to render in their original form, as the format of all encapsulated objects must also be known and the possibility to interpret them must be ensured into the undefined future (Adobe, 2003). Looking at the many formats the computer revolution produced and that are no longer in use and the exponential growth in formats in use that can not be made intelligible to humans any more, this is a task no archives can accomplish. So the formats the archives support to keep their holdings interpretable, need to be restricted.

For the archives to demonstrate that the facts mentioned above have been considered and the corresponding actions have been taken, the processes in archives holding digital material have to be formalized and themselves preserved. The vast amount of electronic records is only manageable with a high degree of automation, so the conclusion to be drawn is that the high degree of

<sup>1</sup> As an example of something more intuitively accessible to humans, we will use a shellac recording. To render the recording only a needle and a postcard are needed. If you stick the needle into the postcard, spin the record, and put the needle in groove, you will hear the recording because of a simple physical analogy: The depth of the groove represents the amplitude of the microphone that recorded the sound. In practice, the changing depth of the groove leads to the vibration of needle, which is then amplified by the postcard, the sound being transmitted to the air and thus to the ear of the listener. With digital data, there is always the need for digital components; there is never a direct physical path.

formalization facilitates the automation of the processes with computers; it is even a prerequisite!

The main questions discussed in this paper are: «Are there special considerations when modeling archival workflows?» and «How should archival workflows be modeled?». It is clear that the answer to the first question has an influence on the answer of the second, as the considerations identified therein serve as basis for the grounds of the result.

The remainder of this paper is structured as follows. Section 2 outlines the archival principles that must be considered in the specification of a modeling and identifies the rating principles (with focus on archival workflows). Section 3 provides three instances of workflows on which the modeling languages will be tested. Section 4 provides an overview over the possible modeling languages for workflows and a rating of these languages according to the rating principles. In Section 5 we summarize the results and draw the conclusion. Section 6 outlines future work.

## 2. ARCHIVAL PARADIGMS

In this Section we will establish a basis by describing the main archival paradigms and inspecting them from a workflow-modeling viewpoint. The paradigms are based on those presented by Gilliland-Swetland (2000).

### 2.1. *The sanctity of evidence.*

The preservation of the evidentiary value of records and their associated context is the creed of every archivist (Jenkinson, 1948). Records passively document the events that led to their creation. The evidence is intrinsically embedded in the records and their context and may not be immediately recognized by the record creators, nor the archives keeping them in their custody. For this reason, an unbroken chain of custody and a precise documentation of the changes to the records is of such paramount importance. Only then, the integrity of the evidential values can be assessed retrospectively. Because, in contradiction to the integrity of records within archives, every alteration of the record or its context may embed new or erase old evidence, and if there

is no documentation of the treatment of the record, the original evidence is impossible to ascertain.

Traditionally, archives were trusted to keep their holdings authentic mainly because they were archives. The idea that an archives must legitimate its actions is relatively new. Even the ISAD(G) standard for archival description in its first version (1996) had not deemed necessary means of describing how and by whom the changes in the description are made. This shortcoming was corrected in the second version (ICA, 1999) of the standard.

However, the constant advances in information technology eliminate the need for transferring records away from their creators as they can continue to reside in the information systems at little extra cost and free of the storage problems inherent to paper records. This has led to a new, additional model for the record life cycle, the records continuum model. Here, the records' creators guarantee the integrity of the records over a much longer period of time. The archive's role in this model is much more active. They are now involved in the record-creating process even before the records are created, they aid in designing the record-creating system. Moreover, they define requirements for that system and monitor the compliance in operation in order to guarantee the integrity of the records. For digital records, the problem is especially acute if the archives choose to implement a migration strategy for ensuring the long term intelligibility of their digital holdings. This strategy calls for a conscious and repeated violation of the integrity on the bit level and will therefore render useless any method for ensuring authenticity based on the bits. We are not aware of a technical mechanism that ensures authenticity on another level than the bit level. Therefore, the archives that performed this violation of integrity must provide proof of the authenticity of the migrated records. A result of this strategy is the shift from original to archival authenticity, meaning that the original files are no longer available for authenticity assessment, only the files and audit information generated by the archives. If an archives chooses emulation as preservation strategy, the authenticity problem is passed down to the emulator.

Much research has gone into the question of how to determine and ensure the authenticity of electronic records, *e.g.* InterPARES (2002) and Cullen *et al.* (2000), there are no definite answers to this problem. It all comes down to trust in records and processes and their documentation (Cullen *et al.*, 2000: 32-50).

From a workflow-modeling viewpoint this means that the design and enactment of the workflow has to have precise and unambiguous semantics. To enable retrospective evaluation of the intrinsic elements of the original record,

all alterations of the records or context must be recorded and preserved together with the records.

## 2.2. *Provenance and original order.*

These principles focus on preserving the context and internal structure of groups of records and form the core of archival theory and practice. The principle of provenance states that the records of one provenance (*i.e.* creating agency) should not be mixed with records from another provenance. This principle, first described in the early 19th century, was adopted by most archives until the end of that century. The principle of original order dictated that structure of the records were to be kept, in the order they were maintained in when they were active, even after transfer to the archives.

The aim of these principles is to preserve the context of and evidence within the records and their relation to each other. The search within the records is simpler for the agency, as it has not changed; the researcher will find the records and those closely related to them more easily. Additionally, these principles allow an archives to handle massive amounts of records, because their transfer to the archives consists only for transport and storage and does not include rearrangement or other intellectual decisions and therefore only allocates little resources.

We did not name the principle of *respect des fonds* as an archival principle in this section because it is not clearly distinguishable from the other principles in the literature (see Rousseau and Couture, 1994: 31 and 61-68; Papritz, 1998: 8-27; Schellenberg, 1961: 14; 1965: 41 and 90).

From a workflow-modeling viewpoint, these principles do not easily translate into requirements. The merely underline the importance and of documenting and defining the priori and posteriori contents and structure of the records, *i.e.* the input and output of the workflow (Papritz, 1998: 82-91).

## 2.3. *The life cycle of records.*

The life cycle of records models the functions of use of and responsibility for records. In the traditional archival perspective, every record's life

has three stages. First, in the active stage, the records is being created and modified by the record-creating body. Second, in the semi-active stage, it is used as a reference for the documented actions, it is stable (*i.e.* it will not be modified). Third, in the inactive stage, the record is no longer easily accessible by its creators (*i.e.* transferred to the archives or destroyed).

From a workflow-modeling viewpoint this means that the model must not be restricted to archival use, but must also be flexible enough to accommodate more generic functions needed in the workflow design outside of digital archives. The modeling language (or representation thereof) must also be simple enough to be understood by non-expert personnel. Because the IT environment of record-creators cannot be controlled, the modeling language must not be tied to a specific workflow engine or tool.

#### *2.4. The organic nature of records.*

Records are the byproduct of an activity of an organization or individual. The relationships between records, their creators and their legal, historical, and procedural context is complex and hard to preserve. The coherence of records with each other is determined by the activity leading to the creation and the body responsible for its creation. Archival practices exploit the organic nature of records by ensuring that records with a strong coherence are collectively captured and documented. This ensures that the legal, historical, and procedural contexts can be re-established much more easily than by the analysis of singular records.

This does not translate into workflow-modeling requirements directly.

#### *2.5. Hierarchy in records and their descriptions.*

Records are embedded in a hierarchy firstly defined by the position of the record-creating agency and secondly by its filing practices. Archival description reflects these hierarchies and describes the sets of records defining and providing sufficient information on each level of the hierarchy. The definition of this «sufficient» information may vary depending on the archives' per-

ceived needs for intellectual control, collection management, and the foreseen user interest.

This approach yields many advantages, namely:

- It allows an intuitive access to information, specifically information not easily located by subject or keyword (*i.e.*, most archives materials) (Papritz, 1998: 185).
- It permits economies in description. The depth to which the records are described can be adapted to the foreseen use and current workload of the archives.
- It can be used on any types of collections and materials. No specialized knowledge is needed to understand the archival descriptions of special forms of materials (*e.g.* maps, scripts, or video tapes).
- This does not translate into workflow-modeling requirements directly, however, it shows the familiarity with the concept of hierarchical structures.

## 2.6. *Rating criteria.*

The following subsections provide the criteria by which the workflow modeling languages will be rated. The criteria summarize the discussion in Section 1 and the prior subsections.

### 2.6.1. *Formal semantics.*

To enable the unambiguous design and enactment of a workflow, its semantics have to be well defined. The modeling language has to have formal semantics. The meaning of textually defined semantics will change over time. Therefore, we believe that only a language defined in mathematical terms can ensure its consistent interpretation for the long term.

This criterion is the direct interpretation of the sanctity of evidence, because a formal basis of the modeling language minimizes the need for elaborate descriptions in textual form, thereby lowering the margin of error.

### 2.6.2. *Flexibility.*

Archival processes change over time. The modeling language has to demonstrate that it can be easily extended to meet future, yet unknown needs. Preferably, it should already have demonstrated this ability in the course of its existence. Additionally, not every aspect of the workflow can be modeled, but some special cases may need an ad hoc extension (Ministerrat, 1984: 15). As an example, an archives could receive a CD-ROM full of WordPerfect files, even though the workflow expects Tiff files. The archives may opt to transform these files into Tiff files before continuing with the workflow. This decision and the conversion would have to be logged in the process, even if the conversion step does not occur in the designed workflow.

This criterion roots in the life cycle of records. Because the environment of the workflow cannot be controlled outside the archives, the modeling language must be easily adapted to new environments and uses. This criterion also ensures that the modeling language will not be subject to change on new requirements.

### 2.6.3. *Decomposition.*

It should be possible to design workflows in building blocks and then connect them vertically and horizontally. Vertical decomposition allows the definition of parts of independent workflows that can be connected at designated points. Hierarchical decomposition means that a workflow can be defined at different levels of complexity. On the top level, a rough sketch of the workflow steps, while on lower levels, the interactions of the systems and technical interfaces need to be specified in detail. Both of these decompositions are useful to reduce the clutter in the diagrams, *i.e.* to keep the workflows comprehensible to humans, to allow for the reuse of previously defined and tested building blocks, and to speed up the workflow development process (parallel development of building blocks).

This criterion is derived from the hierarchy in records and their descriptions. Archivist are used to the top-down approach in describing and finding records in a hierarchical structure because they deal with it on a day to day

basis. Hence, it is natural to begin with defining the task and then refining it on multiple levels of detail.

#### 2.6.4. *Simplicity.*

The concepts and the implementation of the concepts comprised in the modeling language have to be easy to learn and communicate. The basic principles should be grasped intuitively.

This criterion derives from a requirement stated in the life cycle of records. Because the modeling language will need to be understood by non-experts, simplicity is a key factor for the acceptance of the language at the record-creating agency. Additionally, a modeling language based on easily understandable principles is better suited for long-term use, as simple principles are less prone to misinterpretation due to loss of subtleties in definitions.

#### 2.6.5. *Stability.*

To be suitable for long-term use, the modeling language must demonstrate a high degree of stability. This means that it should be mature and in active development and use. Additionally, the language should have backwards-compatibility, *i.e.* the interpretation of models created using an older version of the language must remain unchanged even when using a newer version of the language.

This criterion is derived from the general requirements towards long-term preservation. Frequent changes in the syntax and semantics of the ingested materials lead to an exponential effort in the data management of the digital archives, as all archived versions need to be kept in order to preserve the original evidence within.

#### 2.6.6. *Retraceability.*

The workflows should generate audit trails to enable their retraceability. Although this is a low-level requirement (defined in ICA, 1996), it is of major

importance in the actual implementation and the archival value of the digital records and documents processed. This criterion is listed because features in modeling languages may aid or hinder in this respect.

### 3. ARCHIVAL WORKFLOWS

This Section defines two examples of hierarchically connected workflows within digital archives to test the suitability of the language. Since we have found no consensus on standardizing archival workflows besides CCSDS (2002*a*, 2002*b*), two connected workflows on different hierarchical levels were taken from those (forthcoming) standards.

In the following Subsections we will present example workflows from a digital archives. The workflows presented here should illustrate the most important requirements for archival workflows described above. The terminology and naming in this paper («Ingest», «Data Management», «Archival Storage», «Access», etc.) follows ISO 14721:2003 (Space data and information transfer systems – Open archival information system – Reference model) that is equivalent to CCSDS (2002*a*).

The workflows proposed here are one possible definition of the workflow to achieve the goal of the workflow, *i.e.* ingest an SIP (Submission Information Package), that is, the transfer and accompanying checks of archival data from the Producer to the OAIS in the form of an SIP. Other workflows are also possible and may even be better suited in a given environment, but the ones shown here should provide a solid foundation for the comparison of the workflow modeling languages.

#### 3.1. *Ingest workflow.*

The ingest workflow is defined by the Open Archival Information System (OAIS) by the Consultative Committee for Space Data Systems (2002*a*) at a conceptual level. Additionally, a separate standard for a more detailed specification of the producer-archive interaction is currently under development (CCSDS, 2002*b*). It is the result of agreements between the digital archives' management and the producers. The main task of the ingest workflow is to integrate the records provided by the producer into the archives holdings.

This is achieved by first testing the SIP against the submission agreement, transforming it into an Archival Information Package (AIP), including the metadata generation for the package descriptions, followed by the transfer of the AIP into archival storage and that of the metadata to data management. The steps are as follows:

1. Producer delivers SIP.
2. Ingest checks SIP against contract (validation).
3. Ingest creates AIP(s) and Package Description(s).
4. Ingest stores Package Description(s) (metadata) in Data Management.
5. Ingest stores AIP(s) in Archival Storage.
6. Ingest notifies the producer, that the data has been stored and that the may delete his data.
7. Workflow complete.

Figure 1 shows a graphical representation of this workflow.

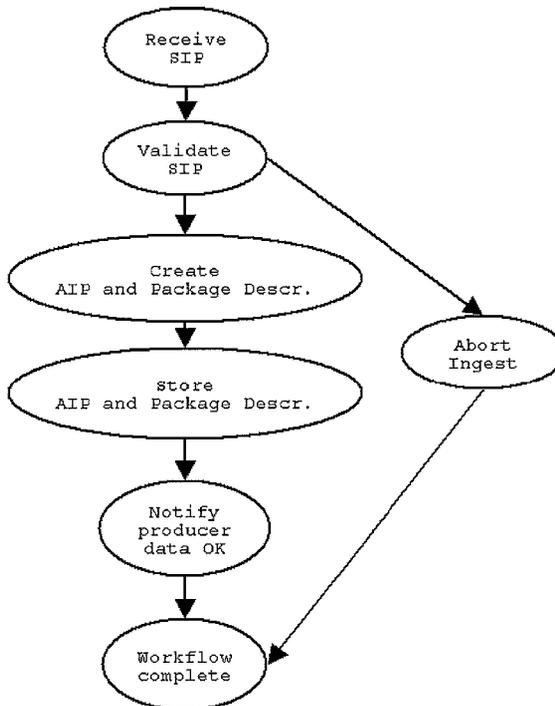


Fig. 1. – Ingest workflow.

### 3.2. *SIP validation workflow.*

The SIP validation workflow is defined by the Producer-Archive Interface Methodology Abstract Standard (CCSDS, 2002*b*) building on the OAIS Reference Model (CCSDS, 2002*a*). It is a more precise definition of the «Validate SIP» step in the submission workflow discussed in Subsection 3.2. This workflow was chosen to see how well the modeling languages tested handle hierarchical refinement, a method often used in the design of processes.

The workflow is described in (CCSDS, 2002*b*: 3-46). The steps are as follows:

1. Apply the validations (Check the conformity of the delivered objects with respect to the model of objects to be delivered and validate their contents).
  - Systematic validation (These validations are carried out after each transfer session).
- or*
- In-depth validation (These validations are only carried out if a coherent set of data objects have been delivered).
2. If the validation has failed, the producer is notified.
3. Workflow complete.

Figure 2 shows a graphical representation of this workflow.

## 4. WORKFLOW MODELING LANGUAGES

In this section the modeling languages are introduced shortly, followed by a mapping of the three example workflows, some thoughts on the mapping process, and an evaluation of the language based on the rating criteria described in Section 2.6.

### 4.1. *High-level Petri nets.*

The Petri net modeling language is widely used for the modeling, simulating, and verifying systems (ISO, 2002: 5). C.A. Petri first proposed it in 1962 for modeling interactions between systems.

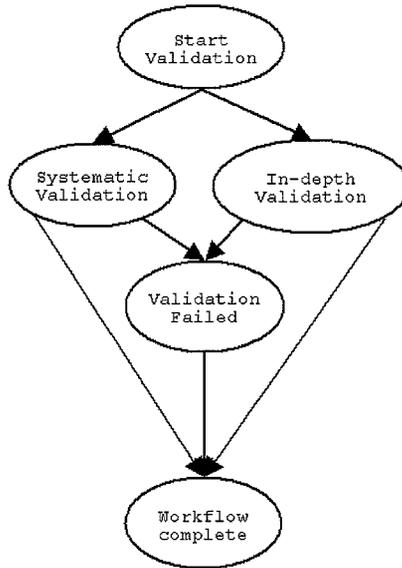


Fig. 2. – SIP validation workflow.

A Petri net is composed (in its graphical representation) as a set of places (circles) and transitions (boxes), which are connected by arcs (arrows). The arcs only connect places and transitions and they have a direction (pointing from or to a transition). The Petri net is the union of places, transitions and arcs. A Petri net also contains tokens (dots); their location symbolizes a state of the net (they can only reside in places). A transition becomes active, if all places, which have an arc pointing to it, contain at least one token. When an active transition «fires», it removes (consumes) one token in each of the aforementioned places and creates a token in each place with an arc pointing from the firing transition to the place. Whether or not an active transition fires, or which transition fires if there are multiple active transitions on the same token, has to be determined externally.

The beauty of Petri nets lies in their having a mathematically equivalent form to the graphical representation, which can be used to verify certain properties of a net, *e.g.*, safeness, boundness, reachability, deadlock. The mathematical representation of Petri nets puts the modeling language on a solid formal base, a requirement, which seldom holds for other modeling techniques.

Basic Petri nets would not suffice for the modeling of workflows, because the concepts of hierarchy and time are alien to them (van der Aalst and van de Graaf, 2003). In contrast, high-level Petri nets offer this functionality on top of the same well-researched foundation. A more detailed description of hierarchy, time, and color (value) is available in van der Aalst and van Hee (2002: 41-48). Additionally, there are ongoing efforts to standardize high-level Petri nets with the International Organization for Standardization (2002).

The following figures (3 and 4) show the transformation of the generic workflows to a Petri net representation. The transformation is straightforward, since the states of the generic workflows directly represent the Petri net transitions, because the actual work is done in the transitions. The arcs of the generic workflows represent the places.

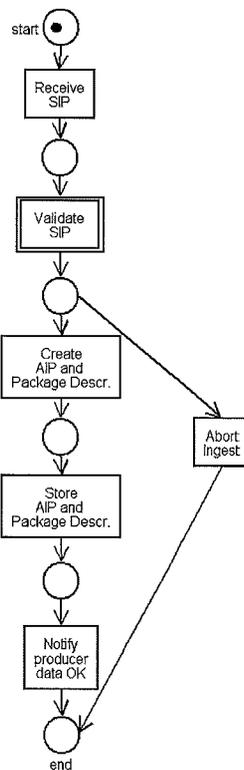


Fig. 3. – Ingest workflow (Petri net).

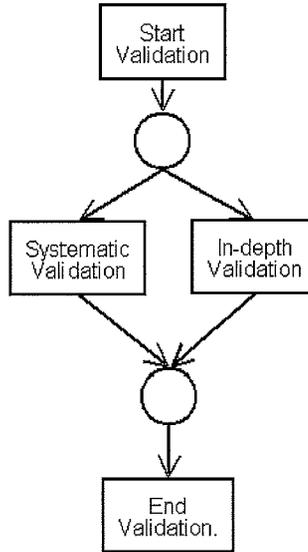


Fig. 4. – SIP validation workflow (Petri net).

The Petri net representation provides additional possibilities for workflow simulation, as it can easily be animated to check that all use cases have been thought of. This animation is not only pretty to look at, but helps non-technical users to a better understanding of the sequence of steps.

TABLE I. – *Evaluation for high-level Petri nets.*

Criteria	Compliance	Remarks
Formal Semantics	1.0	Strong mathematical backing
Flexibility	1.0	Has been extended for many modeling purposes
Simplicity	0.5	Simple core, more complex extensions for workflow modeling
Decomposition	1.0	Not provided by basic Petri nets
Stability	1.0	Very stable and ongoing research and use
Retraceability	1.0	All execution paths visible
Total	5.5	

The evaluation in table I clearly shows that high level Petri nets are very well suited to model workflows for digital archives and that the sole problem lies with its complexity.

#### 4.2. Statecharts.

Statecharts extend the conventional finite state machines with hierarchy, orthogonality, broadcasting, and history. Finite state machines are modeled using state diagrams, which are directed graphs with nodes (states) and arrows (transitions; labeled with events, conditions, and/or actions). The additions provided by the statechart formalism allow for a hierarchical structure of states and sub-states, while minimizing the number of states within the statechart.

Statecharts are widely accepted and used in the industry for specifying the behavior of complex reactive systems (Ackad, 2000: 4-6) (Kye-yune, 2000: 4- 8). There are a few proposals for applying statecharts to workflow modeling motivated by the perceived inadequacy of other modeling languages to model events, *e.g.* Wodtke and Weikum (1997) and Lüttgen *et al.* (1999). Due to its origin, the modeling of events is one of the strengths of statecharts.

The formal foundation of statecharts was initially developed by Harel (1987) and has been adapted to the various needs, which has given statecharts a reputation of highly varying semantics; von der Beeck (1994) has listed more than twenty variants.

Figure 5 shows the transformation of the two generic workflows into one statechart representation. The transformation was straightforward, since the states of the generic workflows directly represent the states of the statecharts. The arcs of the generic workflows represent the arcs of the statecharts. Since hierarchical decomposition is a central element of the statechart formalism, this was used to merge the workflows into one single workflow.

The visual statechart formalism is well suited to map states and reactions to events. Statecharts model hierarchy very intuitively, by allowing additional statecharts to be represented by a single state. However, there are no events (besides the end-event of processing) within the workflows portrayed. Together with the mixing of state and processing in this formalism, it there is little incentive to opt for this modeling language.

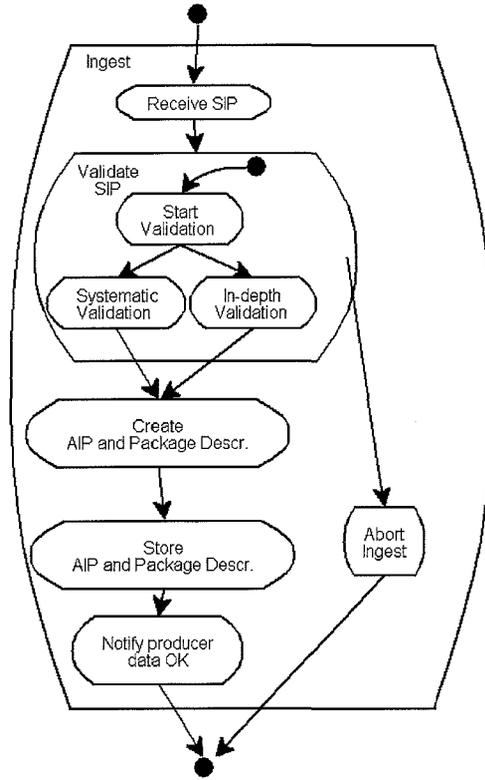


Fig. 5. – Ingest workflow (statecharts).

TABLE II. – Evaluation for statecharts.

Criteria	Compliance	Remarks
Formal Semantics	0.5	Varying semantics
Flexibility	0.5	Constrained to reactive systems
Simplicity	1.0	Intuitive
Decomposition	1.0	Built-in
Stability	1.0	Mature, but varying semantics
Total	4.5	

### 4.3. Activity diagrams.

Activity diagrams are part of the Unified Modeling Language (UML) by the Object Management Group (2003) and therefore part of the de-facto industrial standard in object-oriented modeling. It is a graphical language with textually defined semantics, thus has no formal semantics. These semantics can be defined to suit workflow modeling (Eshuis and Wieringa, 2002), but no commonly accepted semantics yet exist.

The notation of activity diagrams is based on a mixture of flow charts, statecharts, and Petri nets, having circles as start and end states, a diamond shape for decisions, rounded rectangles for action and subactivity states, and a heavy bar for synchronization and forking. They support the notion of subactivity states, *i.e.*, a hierarchical structuring of the activities. In the next version of UML the activity diagrams are supposed to obtain more Petri net-like semantics.

The following figures (6 and 7) show the transformation of the generic workflows to an activity diagram representation. The transformation was straightforward, since the states of the generic workflows directly represent the states of the activity diagrams. The arcs of the generic workflows represent the arcs of the statechart, with added diamond shapes for routing decisions.

The examples show that activity diagram are a hybrid of flow charts, statecharts and Petri nets. They do not bring anything new to the subject.

TABLE III. – *Evaluation for activity diagrams.*

Criteria	Compliance	Remarks
Formal Semantics	0.5	Not defined by UML, see Eshuis and Wieringa (2002) for semantics for workflow modeling
Flexibility	0.5	Has not been retrospectively demonstrated
Simplicity	0.5	Mix of two semantics plus special rules
Decomposition	1.0	Subactivity states supported
Stability	0.0	Semantics will change with UML 2.0
Retraceability	1.0	All execution paths visible
Total	3.5	

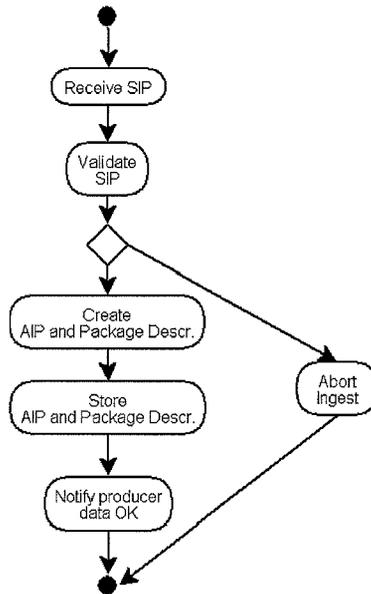


Fig. 6. – Ingest workflow (activity diagrams).

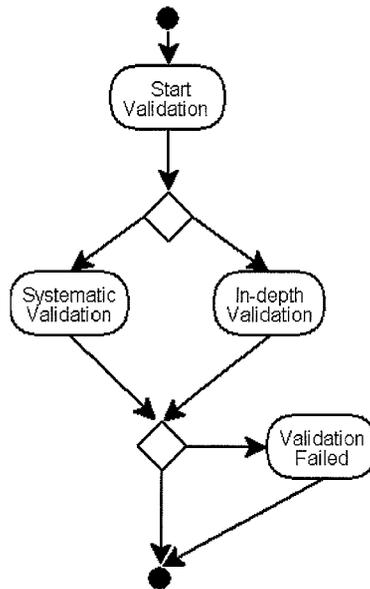


Fig. 7. – SIP validation workflow (activity diagrams).

#### 4.4. *More on modeling languages.*

Other modeling languages were initially considered, but dismissed, mainly because their semantics are not well defined, *e.g.* XML Process Definition Language (WFMC, 2002), or are not stable or well-researched, *e.g.* communicating flowcharts (Dong *et al.*, 2000) or Petri charts (Holvoet and Verbaeten, 1995).

### 5. CONCLUSION

As we have shown it is possible to model our example archival workflows in all the portrayed modeling languages. However, the requirements of an archival workflow language are best met by a language having high-level Petri net semantics (tables I to III).

Another reason to choose high-level Petri net semantics is that they represent a very well researched theory of concurrency. Petri nets are very easy to understand and they can be simulated to further aid the understanding. Petri nets are a very good at modeling workflows because the Petri net does not perform any of the tasks modeled, the workflow management system (WMS) also only manages the execution of the tasks, while specialized software or humans actually perform the work. The actual execution of the tasks is hidden to the WMS as it is hidden from the designer of the Petri net. They only see a box that consumes something and produces a result (data or token). Petri nets are ideal to involve the non-IT-personnel in the design process, because they are easily comprehensible and blend into the common model of workflows that most people have (start - work - result - work - end result).

### 6. FUTURE WORK

In future work we plan to verify and further explore the findings presented in this paper in practice, where the main focus will lie on finding suitable means for expressing workflows in machine-readable form.

Additionally, more workflows need to be designed. This should be done in cooperation with the archival community. We hope that this will raise awareness for problem of workflow description and standardization.

## ACKNOWLEDGEMENTS

We would like to thank the ARELDA project team of the Swiss Federal Archives for their support, namely Peter Keller-Marxer (discussions and review) and Thomas Zürcher (discussions on archival matters).

## REFERENCES

- ACKAD J.-C., 2000. *Optimierte automatische Statechart-Implementierung im Software- und Hardware-Entwurf eingebetteter Systeme*. PhD Thesis, Technische Universität Braunschweig.
- ADOBE SYSTEMS INCORPORATED, 2003. *PDF Reference*. 4th ed. [http://partners.adobe.com/asn/acrobat/sdk/public/docs/PDFReference15\\_v5.pdf](http://partners.adobe.com/asn/acrobat/sdk/public/docs/PDFReference15_v5.pdf) [8 October 2004].
- CONSULTATIVE COMMITTEE FOR SPACE DATA SYSTEMS (CCSDS), 2002a. *Reference Model for an Open Archival Information System (OAIS)*. Blue Book, issue 1, January 2002. <http://www.ccsds.org/documents/650x0b1.pdf> [8 October 2004].
- CONSULTATIVE COMMITTEE FOR SPACE DATA SYSTEMS (CCSDS), 2002b. *Producer-Archive Interface Methodology Abstract Standard*. Red Book (Draft). <http://ssdoo.gsfc.nasa.gov/nost/isoas/CCSDS-651.0-R-1-draft.pdf> [8 October 2004].
- COOK M., 1999. *The Management of Information from Archives*. 2nd ed., Gownner Publishing Company Limited, Aldershot.
- CULLEN C.T., HIRTLE P.B., LEVY D., LYNCH C.A., ROTHENBERG J., 2000. *Authenticity in a Digital Environment*. CLIR Reports, ISBN 1-887334-77-7. <http://www.clir.org/pubs/abstract/pub92abst.html> [9 October 2004].
- DONG G., HULL R., KUMAR B., SU J., ZHOU G., 2000. *A framework for optimizing distributed workflow executions*. In: R.C.H. CONNOR, A.O. MENDELZON (eds), *Proceedings of 7th International Workshop on Database Programming Languages*. Lecture Notes in Computer Science, 1949, Springer-Verlag: 152-167.
- ESIKS R., WIERINGA R., 2002. *Verification support for workflow design with UML activity graphs*. In: *Proceedings of the 24th International Conference on Software Engineering (ICSE 2002)*. ACM Press: 166-176. <http://tmitwww.tn.tue.nl/staff/heshuis/icse.pdf> [9 October 2004].
- GILLIAND-SWETLAND A.J., 2000. *Enduring Paradigm, New Opportunities: The Value of the Archival Perspective in the Digital Environment*. Council on Library and Information Resources, Washington, D.C. <http://www.clir.org/pubs/reports/pub89/pub89.pdf> [9 October 2004].
- HAREL D., 1987. *On the Formal Semantics of Statecharts*. In: *Proceedings Symposium on Logic in Computer Science*. Ithaca, New York: 54-64.
- HOLVOET T., VERBAETEN P., 1995. *Petri Charts: an Alternative Technique for Hierarchical Net Construction*. In: *Proceedings of the 1995 IEEE Conference on Systems, Man and Cybernetics*: 1-6.
- INTERNATIONAL COUNCIL ON ARCHIVES (ICA), 1996. *Code of Ethics*. Beijing. [http://www.ica.org/biblio/code\\_ethics\\_eng.html](http://www.ica.org/biblio/code_ethics_eng.html) [9 October 2004].
- INTERNATIONAL COUNCIL ON ARCHIVES (ICA), 1999. *ISAD(G): General International Standard Archival Description*. 2nd ed. [http://www.ica.org/biblio/isad\\_g\\_2e.pdf](http://www.ica.org/biblio/isad_g_2e.pdf) [9 October 2004].

- INTERNATIONAL ORGANIZATION FOR STANDARDIZATION (ISO), 2002. *High-level Petri Nets - Concepts, Definitions and Graphical Notation*. Final Draft International Standard ISO/IEC 15909. <http://www.Petrinets.info/docs/pnstd-4.7.4.pdf> [9 October 2004].
- INTERPARES, 2002. *Authenticity task force final report: Establishing and maintaining trust in electronic records*. [http://www.interpares.org/book/interpares\\_bok\\_d\\_part1.pdf](http://www.interpares.org/book/interpares_bok_d_part1.pdf)
- JENKINSON H., 1948, *The English Archivist: A New Profession*. Inaugural Lecture for a New Course in Archive Administration delivered at University College London, 14 October 1947, H.K. Lewis and Co., London.
- KELLER P., 2004. *On the Purpose of Archives*. <http://heuscher.ch/PurposeOfArchives> [9 October 2004].
- KYEYUNE Y., 2000. *Developing Concepts and Methods for Module and Integration Tests of Reactive Systems*. PhD thesis, Universität Dortmund.
- LÜTTGEN G., VON DER BEECK M., CLEAVELAND R., 1999. *Statecharts via process algebra*. In J.C.M. BAETEN, S. MAUW (eds.), *Proceeding of the 10th International Conference on Concurrency Theory*. Lecture Notes in Computer Science, 1664, Springer-Verlag: 399-414.
- MINISTERRAT DER DEUTSCHEN DEMOKRATISCHEN REPUBLIK (DDR), MINISTERIUM DES INNERN, STAATLICHE ARCHIVVERWALTUNG, 1984. *Archivarbeit rationell - Arbeitsabläufe*. Staatsverlag der Deutschen Demokratischen Republik, Berlin.
- OBJECT MANAGEMENT GROUP (OMG), 2003. *Unified Modeling Language Specification*. Version 1.5. <http://www.omg.org/cgi-bin/doc?formal/03-03-01> [9 October 2004].
- PAPRITZ J., 1998. *Archivwissenschaft*. 2nd ed., vol. 3, Archivschule Marburg, Marburg.
- ROTHENBERG J., 1995. *Ensuring the Longevity of Digital Information*. *Scientific American*, 272(1): 42-47.
- ROUSSEAU J.-Y., COUTURE C., 1994. *Les Fondements de la Discipline Archivistique*. Presses de l'Université de Québec, Québec.
- SCELLENBERG T.R., 1961. *Akten- und Archivwesen in der Gegenwart*. Karl Zink Verlag, Munich.
- SCELLENBERG T.R., 1965. *The Management of Archives*. Columbia University Press, New York-London.
- VAN DER AALST W.M.P., 1998. *The Application of Petri Nets to Workflow Management*. *The Journal of Circuits, Systems and Computers*, 8(1): 21-66. <http://tmitwww.tn.tue.nl/staff/wvdaalst/Publications/p53.pdf> [9 October 2004].
- VAN DER AALST W.M.P., VAN DE GRAAF M., 2003. *Workflow Systems*. In: C. GIRAULT, R. VALK (eds.), *Petri Nets for Systems Engineering*. Springer-Verlag: 506-540.
- VAN DER AALST W.M.P., VAN HEE K., 2002. *Workflow Management: Models, Method, and Systems*. The MIT Press, Cambridge-London.
- VON DER BEECK M., 1994. *A comparison of statechart variants*. In: H. LANGMAAK, W.P. DE ROEVER, J. VYTOPIK (eds.), *Formal Techniques in Real-Time and Fault-Tolerant Systems*. Lecture Notes in Computer Science, 863, Springer-Verlag: 128-148.
- WODTKE D., WEIKUM G., 1997. *A Formal Foundation for Distributed Workflow Execution Based on State Charts*. In: F.N. AFRATI, P. KOLAITIS (eds.), *Proceedings of the International Conference on Extending Database Theory*. Lecture Notes in Computer Science, 1186, Springer-Verlag: 230-246.
- WORKFLOW MANAGEMENT COALITION (WFMC), 2002. *Workflow Process Definition Interface – XML Process Definition Language*. [http://www.wfmc.org/standards/docs/TC-1025\\_10\\_xpdl\\_102502.pdf](http://www.wfmc.org/standards/docs/TC-1025_10_xpdl_102502.pdf) [9 October 2004].



EMILIANO DEGL'INNOCENTI\*

## TRAME E FILI DI ARIANNA NELLA MASSA DIGITALE<sup>1</sup>

Il titolo di questo intervento recita «Trame e fili di Arianna nella massa digitale» ed è già di per sé significativo del nucleo concettuale intorno al quale le riflessioni proposte ruotano. Il momento storico in cui ci troviamo a vivere può a buon diritto essere definito un momento di passaggio: passaggio da quella che Marshall McLuhan definì efficacemente «galassia Gutenberg» (cfr. McLuhan, 1962) a quella che si sta imponendo come costellazione tecnologica dominante, definita «galassia Marconi» (cfr. McLuhan, 1962, trad. it., 1976: 199-201). Mario Ricciardi (1997: 179) ha scritto che lo sviluppo dei media di massa ha posto le premesse ideali per lo sviluppo di un'economia dell'immateriale: questa effervescenza non ha tuttavia implicato, almeno per ora, la sparizione delle forme consuete di gestione del sapere. Si è innescato piuttosto un processo di sovrapposizione in qualche misura ricorrente nei momenti di passaggio da un *milieu* o ciclo tecnologico ad un altro entro il quale il ciclo destinato al declino non cede di schianto sotto la pressione del ciclo antagonista, ma, al contrario, esprime il meglio della sua produzione: il perdurare della tradizione delle *artes memorandi* ben oltre l'invenzione della stampa non ne è che un esempio.

Appare tuttavia innegabile che l'impatto delle nuove tecnologie nella percezione dei fenomeni culturali (e non solo) sia sempre più determinante ed assolutamente non trascurabile: le conseguenze di questa penetrazione dell'informatica nei più diversi contesti della nostra esperienza quotidiana sono, infatti, tutt'altro che secondarie. Ad un primo livello di analisi si colloca, ba-

\* Fondazione Ezio Franceschini - Certosa del Galluzzo - 50124 FIRENZE, degli\_innocenti@hotmail.com

<sup>1</sup> Gli argomenti affrontati in questo intervento sono stati da me discussi più volte – nel contesto delle varie attività di ricerca portate avanti dalla Fondazione Ezio Franceschini e dalla SISMEI – con i Proff. Claudio Leonardi e Francesco Santi, che desidero ringraziare per la disponibilità e le preziose indicazioni fornitemi.

nalmente, la dimensione quantitativa: l'aumento smisurato della capacità individuale di produrre informazioni, che in un recente studio condotto presso l'università di Berkeley è stata stimata – per il 2002 – intorno agli 800 MB su base annua, corrispondenti a circa 5 exabytes ( $10^{18}$  bytes), in riferimento all'intera popolazione mondiale<sup>2</sup> (stimata intorno ai 6,3 miliardi di persone<sup>3</sup>). A conferire profondità ad un dato statistico altrimenti piatto, può contribuire il confronto con i risultati ottenuti, solo pochi anni prima, dalla stessa équipe: nel 1999, la quantità di «unique informations» prodotta da ciascun individuo in un anno, si attestava intorno ai 250 MB, equivalenti a circa 2 exabytes su scala planetaria (cfr. Lyman *et al.*, 2000). Ovviamente un simile aumento della massa di informazioni in circolazione – intorno al 36% su base annua (*ibid.*), per il periodo compreso fra 1999 e 2002 – rende indispensabile l'elaborazione di strategie di selezione dei contenuti e genera rilevanti ricadute anche a livello metodologico e cognitivo.

Sarebbe tuttavia riduttivo ritenere problemi simili, intendo dire di accumulazione e gestione del sapere, di esclusiva pertinenza dell'odierna società dell'informazione: le stesse preoccupazioni tormentavano già gli antichi che a partire addirittura dal IV sec. a.C. (il celebre frammento greco *dialexis*<sup>4</sup> ne è una prova) elaborarono complessi sistemi mnemonici (tanto complessi da divenire, a volte, inservibili ed autocontraddittori). A ben vedere neppure le critiche di volta in volta mosse a questi nuovi (per l'epoca) strumenti di gestione delle informazioni risultano troppo dissimili, *mutatis mutandis*, da quelle che, a più riprese, sono state rivolte ai vari protagonisti della rivoluzione mediatica

<sup>2</sup> Il dato, aggiornato al 1999, è riportato anche in Smith (2003: 26). Per lo studio originale cfr. Lyman *et al.*, 2000. Per la versione dello studio aggiornata al 2003 cfr. Lyman *et al.*, 2003.

<sup>3</sup> La fonte di questo dato è il rapporto 2003 World Population Data Sheet, stilato dal Population Reference Bureau degli Stati Uniti [<http://prb.org> (10 febbraio 2004)].

<sup>4</sup> «Grande e bella invenzione è la memoria, sempre utile sia al sapere, sia al vivere. La prima cosa è questa: se tu presti attenzione [dirigi la tua mente] il giudizio coglierà meglio le cose che ti giungono attraverso di essa [la mente]. In secondo luogo, ripeti più di una volta ciò che senti, perché udendo e ripetendo la stessa spesso la stessa cosa, ciò che hai imparato ti penetra interamente nella memoria. In terzo luogo, ciò che odi, riponilo in ciò che conosci. Per esempio: devi ricordare Χρύσιππος (Crisippo); lo collochiamo su χρυσός (oro) e ἵππος (cavallo). Un altro esempio: poniamo πυριλαμψής (lucciola) su πῦρ (fuoco) e λάμπειν (splendere). Questo per ciò che riguarda le parole. Per le cose fa' così: il coraggio ponilo su Ares e Achille, l'arte di lavorare i metalli su Efesto; la virtù su Epeo [...].» Cfr. Diels (1922: 345). La traduzione italiana è tratta da Yates (1969; trad. it., 1972: 29).

(tv, computer, ecc.). Così è descritta nel Fedro platonico l'invenzione dell'alfabeto:

«Ma quando giunse alla scrittura Theuth disse: “Questa conoscenza, o re, renderà gli Egiziani più sapienti e più capaci di ricordare, perché con essa si è ritrovato il farmaco della memoria e della sapienza”. E il re rispose: “O ingegnosissimo Theuth, c'è chi è capace di creare le arti e chi è invece capace di giudicare quale danno o quale vantaggio ne riceveranno coloro che le adopereranno. Ora tu, essendo padre della scrittura, per affetto hai detto proprio il contrario di quello che essa vale. Infatti la scoperta della scrittura avrà per effetto di produrre la dimenticanza nelle anime di coloro che la impareranno, perché fidandosi della scrittura, si abitueranno a ricordare dal di fuori mediante segni estranei e non dal di dentro e da se medesimi: dunque tu hai trovato non il farmaco della memoria, ma del richiamare alla memoria. Della sapienza, poi, tu procuri ai tuoi discepoli l'apparenza, non la verità: infatti essi, divenendo per mezzo tuo uditori di molte cose senza insegnamento, crederanno di essere conoscitori di molte cose, mentre, come accade per lo più, in realtà, non le sapranno; e sarà ben difficile discorrere con essi, perché sono diventati portatori di opinioni invece che sapienti”»<sup>5</sup>.

Né più favorevole risulta il giudizio espresso sulle pratiche mnemotecniche dallo stesso Platone nei dialoghi *Ippia maggiore* ed *Ippia minore*<sup>6</sup>, dedicati al sofista Ippia di Elide<sup>7</sup>, vissuto nel V sec. a.C. Credo che il filo rosso che unisce antichi e moderni nel tentativo di governare la massa delle informazioni sia rappresentato dal declino della memoria individuale, che ha costituito – nella fase orale della storia della cultura – il limite invalicabile della capacità umana di incremento della conoscenza: per tante che fossero, le informazioni stipate nel cervello di un solo uomo rappresentavano sempre una quantità finita e relativamente piccola, destinata alla dispersione in occasione della morte dell'individuo. Con il passaggio dalla cultura orale a quella scritta, descritto da Platone nel Fedro, il limite di accumulazione si accresce a dismisura, divenendo di fatto illimitato; parimenti, la possibilità di evitare la

<sup>5</sup> Platone, Fedro [274 E] - [275 C]. La traduzione italiana utilizzata è quella di G. Reale (Platone, 1991: 579-580).

<sup>6</sup> Platone, *Ippia maggiore* [285 D] - [286 A] ed *Ippia minore* [368 D] - [369 B].

<sup>7</sup> Nella presentazione all'*Ippia minore*, scritta da Maria Teresa Liminta (Platone, 1991: 1004), il filosofo di Elide è così descritto: «Dotato di memoria straordinaria eresse ad arte sistematica la mnemotecnica [...]».

dispersione dei contenuti informativi, fissandoli su un supporto fisico, diviene concreta.

Sarebbe troppo lungo e forse inopportuno ripercorrere in questa sede, seppur brevemente, la storia del tentativo dell'uomo di opporsi alla sempre più ingombrante massa di informazioni generata dalla società, basti solo ricordare che questo tema percorre sotterraneamente la storia della letteratura (Borges, 1956), Bioy Casares, 1953), quella della filosofia (ricorderò, fra i moltissimi, soltanto Raimondo Lullo, Giordano Bruno e Giulio Camillo<sup>8</sup>) e quella della scienza (Vannevar Bush<sup>9</sup>).

Non sembra dunque di trovarsi di fronte ad un fenomeno del tutto nuovo, almeno nella sostanza, anche se esso, per certo, ha generato ricadute del tutto particolari e degne di attenzione. Abbiamo già menzionato l'importanza di un fenomeno del tutto peculiare, conseguenza diretta della rivoluzione massmediatica: il progressivo svilupparsi di una vera e propria economia dell'immateriale. Una delle molteplici sfaccettature di questo fenomeno complesso è rappresentata dal nuovo mercato delle edizioni elettroniche. Sebbene, come ha osservato Bernard Quemada (2003: 10): «La confiance envers les méthodes informatiques succède à un intérêt tempéré d'illusions et de déconvenues. Le recours aux ordinateurs n'est plus un acte de foi comme à la fin des années 60, o un fait de curiosité, comme dans les années 70», appare fin troppo evidente come oggi non sia più sufficiente mettere insieme testi elettronici quanto piuttosto sia necessario lavorare allo sviluppo di strumenti in grado di rendere quei testi meglio analizzabili con metodologie informatiche (sulla questione delle edizioni digitali cfr. almeno Mordenti, 2001; Orlandi, 1998a, 1998b; Robinson, 2003). A questo riguardo, gli studi pionieristici di padre Busa (cfr. almeno Busa, 1951a: 479-493; 1951b; 1952: 95-97), sfociati nella pubblicazione degli *Opera Omnia* di Tommaso d'Aquino (Busa, 1992), hanno favorito la codifica di modelli metodologici innovativi per la nascente lessicografia culturale, fondati sull'analisi strutturale e funzionale e su quella quantitativa stilistica e fattoriale.

<sup>8</sup> Per un quadro completo della questione cfr. almeno: Rossi (1960); Yates (1969); Bolzoni (1995); Bolzoni (2002); Carruthers (1990); Carruthers (1998); Carruthers e Ziolkowski (2002).

<sup>9</sup> Bush (1945: 101-108; trad. it., 1992) L'articolo si può leggere sul sito web della rivista statunitense, all'indirizzo: <http://www.theatlantic.com/unbound/flashbks/computer/bushf.htm> (10 febbraio 2004).

Ripartiamo dalla sempre maggiore penetrazione dell'informatica nell'esperienza culturale di tutti i giorni: lo storico canadese Harold Innis (1950, 1951), uno degli studiosi più sensibili nei confronti di questo genere di cambiamento, ha affermato in tempi non sospetti ed in modo molto persuasivo che la natura o il carattere della tecnologia dell'informazione della quale dispone una società produce un effetto profondo sulle sue istituzioni, sul suo governo, e sulla sua vita.

Come nel caso del passaggio dall'oralità alla scrittura, anche nella nostra epoca il ruolo giocato dal supporto fisico su cui l'informazione è fissata è quantomai importante: non è un caso che le capacità dei dispositivi di archiviazione, magnetici (dischi fissi) ed ottici (CD, DVD), siano cresciute di pari passo con la proliferazione di informazioni digitali che sono il risultato di una conversione a posteriori, o di una creazione delle informazioni direttamente in forma digitale (cfr. Dollar, 1994: 161-184).

Questa situazione ha reso sempre più necessaria l'elaborazione di strategie in grado di dominare la massa digitale evocata nel titolo: da una parte si è cercato di plasmare questa formazione magmatica di informazioni, provvedendo alla creazione delle cosiddette collezioni digitali; dall'altra si è ritenuto utile, se non indispensabile, rendere questo materiale maggiormente fruibile per il ricercatore mediante lo sviluppo di strumenti di indicizzazione e ricerca. È a questo livello che si fanno pressanti le istanze di tipo qualitativo ricordate all'inizio: l'accresciuta capacità individuale di produrre informazione impone prima di tutto l'elaborazione di criteri e strategie di selezione ed ordinamento. È appena il caso di accennare come in questi processi siano implicate categorie cruciali della storia del pensiero occidentale, quali ad esempio la coppia *autore/autorità* ed il concetto di *gerarchia*<sup>10</sup>, entrate in crisi – o comunque divenute bisognose di revisione – a causa dei mutamenti strutturali generati dall'affermazione dei mass-media.

Ci si è spesso interrogati su chi debba decidere cosa conservare e cosa cancellare o gettare: è tuttavia verosimile ipotizzare che si tratti di un processo, almeno in parte, casuale, determinato in ultima analisi dalle decisioni più o meno razionali prese da singoli individui. In modo più profondo, il problema che ci viene posto riguarda le forme di intervento delle nostre società e delle nostre istituzioni nella costruzione dell'immagine di loro stesse che

<sup>10</sup> Cfr. a questo proposito Saracco (2000: 34-41) e Ortoleva (2000: 42-45).

verrà trasmessa alle generazioni future per permettere loro di capire da dove vengono e chi sono (cfr. Aymard, 1997, 3-14).

Maurice Aymard, uno storico di professione, ha ritenuto opportuno considerare il problema della selezione dei contenuti tenendo presente sia il piano dell'innovazione tecnologica che quello della metodologia della ricerca storica: muovendo dalla concezione braudeliana di storia aperta, ha evidenziato come sia particolarmente rischioso procedere all'eliminazione di tutti quegli elementi ritenuti marginali dal vaglio preventivo delle fonti, ma che potrebbero – in seguito – rivelarsi preziosi testimoni. In questa accezione del termine, *selezionare* significa, brutalmente, distruggere: tutto ciò che non è stato selezionato sarà perso per sempre. La *Weltanschauung* codificata dagli studiosi odierni, attraverso continui processi di selezione/soppressione, è destinata perciò a influenzare la percezione ed il lavoro di ricostruzione delle generazioni successive. Quello che non si è conservato non ritorna: si potrà cercare solo in mezzo a ciò che si è ritenuto, in base ai nostri parametri o alle nostre metodologie statistiche, «rilevante». La scelta, operata su basi statistiche, di un campione ritenuto rappresentativo, presuppone che le interrogazioni delle generazioni future, così come i metodi di lavoro, rimarranno gli stessi di oggi: condizione questa tutt'altro che certa. Le regole oggi in vigore corrispondono esse stesse a concezioni storicamente datate e a norme legate a un contesto mutevole, e verranno sicuramente criticate dalle generazioni future, che ci rimprovereranno di avere spese tante energie per conservare documenti che non avranno per loro alcun interesse, e di averne nello stesso tempo buttato via tanti altri, che li avrebbero invece interessati (*ibid.*). Le possibilità offerte alla ricerca storica dallo sviluppo di altre discipline che pongono nuove domande definiscono nuovi concetti e nuove metodologie, suggeriscono nuove piste di ricerca, elaborano nuovi dati che gli storici possono riutilizzare per conto proprio. Gli storici, nella loro maggioranza, si sono così adattati a una serie di trasformazioni di fondo che hanno coinvolto le società nelle quali vivono; anzi, proprio questi mutamenti li hanno stimolati a cercare nel passato sia situazioni differenti (antecedenti a queste trasformazioni) sia le loro origini. L'esercizio del mestiere di storico è cambiato. Ma è pure vero che la moltiplicazione crescente, durante l'ultimo mezzo secolo, delle registrazioni e delle riproduzioni di testi, parole, suoni e immagini (oggi combinate nelle forme più varie di prodotti multimediali) è venuta, colla massa stessa dei documenti da conservare, a impostare in termini radicalmente nuovi il problema degli archivi.

Lo sviluppo della memoria elettronica ha creato un rapporto totalmente nuovo tra il mondo contemporaneo e la massa d'informazioni da esso prodotta: le tecnologie informatiche hanno infinitamente aumentato la nostra capacità di generare, raccogliere, richiamare ed esaminare dati, fornendo agli utenti la capacità di condividere e paragonare informazioni e, persino, di raccogliere dati senza l'intervento umano, automaticamente. Ciò ha dato origine a quello che Luciana Duranti ha definito «la scoperta negativa»: «L'impatto della scoperta negativa (la scoperta di aree di ignoranza non ancora riconosciute o identificate) sullo sviluppo della conoscenza e sulla sua assimilazione nella memoria della nostra civiltà crescerà in modo esponenziale. Infatti come gli strumenti per le osservazioni diventeranno sempre più sofisticati e produttivi, le loro rivelazioni forniranno risposte a domande non ancora poste» (Duranti, 1994: 147-160).

Mai prima d'ora una civiltà è stata così sommersa da una tale abbondanza di dati disponibili: tale abbondanza minaccia la stessa capacità umana di controllarli, capirli ed usarli creativamente e produttivamente, così che possano alla fine generare conoscenza e diventare parte della memoria dell'umanità. L'unica speranza per la sopravvivenza della memoria elettronica dell'epoca contemporanea è la riduzione dei dati esistenti ad una quantità maneggevole. Il concetto di «scoperta negativa» sembra suggerire il metodo più appropriato per raggiungere questo scopo: le informazioni ridondanti, superflue, irrilevanti, potrebbero essere progressivamente eliminate finché non rimanga altro che un nucleo fondamentale, compatto e significativo di informazioni che potremmo capire, usare, rinnovare, integrare, interiorizzare e trasmettere alla generazione successiva. E questo nucleo sarebbe composto da ciò che appare necessario per la conservazione e la continuazione del nostro sistema giuridico e della sua rete di rapporti, della nostra cultura, e di quei germi di future scoperte che abbiamo raccolto perché le generazioni successive possano sfruttarli (*ibid.*).

Per poter eliminare l'irrilevante, bisogna saperlo distinguere dal rilevante, cioè bisogna avere una visione generale ed una comprensione di base di tutte le informazioni disponibili, un'idea chiara di ciò che costituisce il nucleo essenziale da conservare *ad perpetuam memoriam*, e l'abilità di riconoscere ciò che non gli appartiene. Se ci concentriamo specificamente sull'informazione elettronica, appare presto evidente che non è possibile ottenere una visione generale del suo universo senza conoscerne le parti e i loro collegamenti con altri mondi informativi complementari.

Sarebbe tuttavia un errore pensare alla tecnologia come mezzo per registrare tutto e per sempre, giacché grandi quantità di dati non producono ne-

cessariamente conoscenza: risulta necessario, piuttosto, far valere un principio ordinatore, mediante il quale orientarsi nella massa digitale. Il problema pertanto è duplice: da una parte individuare i dati veramente essenziali da conservare per il futuro, sfidando le inevitabili incompatibilità frutto dell'avanzamento della tecnologia; dall'altra garantire un sistema di conservazione di questi dati in riferimento alle nuove tecnologie. C'è inoltre un problema di fondo legato al processo di scelta da adottare per conservare le memorie. Chi decide? Non dovrebbe essere il tecnologo, ma l'esperto della materia, fornito però di quelle conoscenze tecnologiche in grado di fargli compiere le scelte appropriate. Di questi esperti, però, si sente la carenza: c'è pertanto il rischio di dover lasciare ai tecnologi un ruolo superiore a quello che loro competerebbe. Se la progettazione dei sistemi elettronici viene affidata completamente agli specialisti dell'ingegneria informatica, è molto probabile che i problemi di identificazione aumentino geometricamente, a causa dello sviluppo delle superstrade dell'informazione, alla conseguente proliferazione di dati, alla sempre maggiore somiglianza e incontrollabile duplicabilità dei prodotti dell'informazione e all'inevitabile sovrapposizione delle loro funzioni e scopi (cfr. Ridolfi, 1994: 247-255; Duranti, 1994: 147-160). Ciò sembra implicare, per utilizzare le parole di Charles Dollar, che «il solo fatto che la conoscenza umana possa essere rappresentata in forma digitale non implica che ogni rappresentazione digitale meriti di essere inclusa nella memoria elettronica. Nel corso degli ultimi sei millenni ci sono abbondanti prove che del materiale è stato selezionato per la preservazione mentre si è lasciato che altro materiale si deteriorasse. Non c'è ragione per la quale questo principio della selezione per la preservazione non possa e non debba influenzare la memoria elettronica» (Dollar, 1994: 163).

Un paradigma per la conservazione della memoria elettronica, fondato sull'adeguamento delle metodologie e delle tecniche della preservazione, ed orientato a facilitare il più possibile l'accesso al sapere documentato, dovrebbe tener conto delle variabili rappresentate dalla disponibilità, comprensibilità, e affidabilità del materiale digitale. Ancora Dollar offre una efficace sintesi del concetto di disponibilità: «La *disponibilità* della memoria elettronica significa accesso fisico, che comprende l'identificazione, la leggibilità, la recuperabilità. L'identificazione significa che la memoria elettronica viene organizzata e descritta in modo tale che è possibile per gli utenti distinguere gli oggetti dell'informazione [...]. La leggibilità significa che l'informazione digitale può essere elaborata su un sistema computerizzato o su un dispositivo diverso da quello che l'ha inizialmente creata o che la conserva normalmente. [...] Il terzo aspetto della dispo-

nibilità della memoria elettronica è la sua ricercabilità, che presuppone l'identificazione e la leggibilità [...], e significa che oggetti d'informazione separati o parti di essi possono venire recuperati e mostrati. La ricercabilità richiede chiavi o indici che colleghino la struttura logica degli oggetti d'informazione, alle loro posizioni fisiche di archiviazione e consentano la discriminazione tra due o più oggetti d'informazione» (Dollar, 1994: 174-180). Le altre due variabili, comprensibilità ed affidabilità, rappresentano la proiezione a lungo termine del concetto stesso di disponibilità: la prima è direttamente dipendente dal ricambio periodico delle piattaforme software, la seconda è legata alla capacità di durata nel tempo dei supporti fisici su cui le informazioni sono immagazzinate.

Nel corso dell'argomentazione sin qui svolta si sono andate delineando piuttosto chiaramente alcune linee di tendenza molto generali: ci troviamo al cospetto di un aumento della capacità individuale di produrre informazioni, il che implica che, per mantenere traccia – una traccia significativa, cioè dotata di senso, in grado di veicolare realmente un contenuto informativo – di questo sedimento, sono necessarie strategie nuove e complesse, giacché ad un mero incremento quantitativo della capacità di produrre informazioni, tipico di questo ciclo tecnologico, non necessariamente corrisponde un apprezzabile aumento della nostra capacità di comprenderne la portata.

Abbiamo precedentemente accennato alle ricadute in campo cognitivo determinate dall'affermarsi delle nuove tecnologie, argomento sul quale varrà la pena di soffermarsi ancora un poco. Appare in effetti plausibile che anche le nostre strutture cognitive siano destinate ad essere ridefinite, in qualche misura, per adeguarsi alle nuove strategie di veicolazione dell'informazione, peculiari per ciascun medium. Questa teoria, alla base della riflessione McLuhaniana, è stata ripresa ed approfondita dal suo successore alla guida del Center for Culture and Technology di Toronto: Derrick de Kerckhove (1995: 226). Questi stessi processi di adattamento furono studiati originariamente in ambito antropologico da Edmund Carpenter (1973), amico e collaboratore di McLuhan<sup>11</sup>, e godono oggi di pieno diritto di cittadinanza nell'ambito degli studi di percettologia (cfr. Sturm, 1988: 39). Quello che aveva già sostenuto Georg Simmel (1903) nel suo

<sup>11</sup> Negli anni '50 del secolo scorso, McLuhan e Carpenter, che lavoravano per l'Università di Toronto e condividevano uno spiccato interesse per il linguaggio e lo studio della comunicazione umana, decisero di unire i loro sforzi nella realizzazione di un progetto comune di ricerca, che si concretizzò nella nascita della rivista «Explorations», la cui pubblicazione fu interrotta nel 1959, dopo soli 8 numeri. Sulla questione cfr. Baragli (1980: 60-63).

*Le metropoli e la vita dello spirito*, ossia che con l'accrescersi smisurato del numero di sollecitazioni, o come lui le definiva «relazioni» (*Wechselwirkung*), la capacità di rispondere ad esse in maniera adeguata lascia spazio alla narcosi, all'intorpidimento, è oggi certificato dagli psicologi della percezione. Alla stessa maniera, infatti, i tempi rapidissimi imposti allo spettatore per la fruizione dei contenuti televisivi, rendono impossibile una risposta meditata da parte sua, o per usare una terminologia più adeguata, gli stimoli – sempre più frequenti e ravvicinati – non danno la possibilità all'utente di provvedere alla loro chiusura mediante una risposta orientata (de Kerckhove, 1995; trad. it., 1996: 21-23). Questo significa che alla successione logico-temporale del momento della riflessione e dell'azione, alla successione coordinata, logica, tra un prima-premessa e un poi-conseguenza, si stanno sostituendo movimenti mentali coordinati e simultanei, azioni dirette e immediate che devono controllare e agire su schermi e ambienti conoscibili simultaneamente e in tempo reale. Esse quindi aumentano la richiesta di attività mentale e costruttiva (Ricciardi, 1997: 179).

Un'altra conseguenza che mi sembra interessante segnalare è la necessità di escogitare nuove architetture del sapere, e codificare nuove metafore in grado di rappresentarle adeguatamente. Se fino a qualche decennio fa, infatti, la metafora prevalente era quella architettonica (cfr. Lucken, 2005: 141-149), basti pensare ai numerosissimi palazzi o castelli della conoscenza – su tutti quello escogitato dal gesuita Matteo Ricci – o ai celebri *theatri* di Giulio Camillo (1991) e Robert Fludd (cfr. Yates, 1969; trad. it., 1972: 297-341), oggi questa connotazione fisica, concreta, è andata dissolvendosi, lasciando spazio a strutture ed immagini differenti: prime fra tutte la rete e la navigazione. Questi cambiamenti nell'immaginario e nelle strutture retoriche sono indice di mutazioni più profonde, che riguardano la percezione e le teorie della conoscenza, in una parola la sfera cognitiva: fra i primi ad accorgersi di questi cambiamenti, oltre ai già citati Innis e McLuhan, ricordiamo Vannevar Bush e Ted Nelson (1981), personalità che a vario titolo hanno contribuito all'elaborazione di teorie e strumenti in grado di facilitare l'orientamento all'interno della massa delle informazioni in continua crescita – i fili ricordati dal titolo – che costituiscono altrettante possibilità di accesso analogico alle informazioni secondo dinamiche non lineari. Pare appena il caso di segnalare come l'odierno interesse per la codifica di interfacce utente sempre più efficienti e per strumenti di navigazione sofisticati possa essere considerato un'attualizzazione di problemi conosciuti anche in altre epoche, riguardanti la rappresentazione e l'indicizzazione delle informazioni: mi sembrano illuminanti a questo proposito le pagine di Lina Bol-

zioni, in cui si descrivono le dinamiche che hanno favorito il passaggio da una modalità iconica di rappresentazione, tipica della tradizione mnemotecnica di matrice erenniana<sup>12</sup>, ad una schematica di stampo ramista; motivazioni legate, ancora una volta, alla sfera cognitiva (Bolzoni, 1995). Letta in un simile contesto l'affermazione McLuhaniana secondo la quale «il *medium* è il messaggio» (McLuhan, 1964), non suona poi così insensata: i contenitori tecnologici a cui, di volta in volta, affidiamo le informazioni che siamo in grado di produrre o di raccogliere, intervengono inevitabilmente su di esse, strutturandole secondo la propria natura. Nel passaggio da analogico a digitale, pertanto, dobbiamo essere pronti ad adeguare le nostre strategie cognitive, così come le strutture retoriche, per conformarle alle nuove forme di codifica dell'informazione.

In un recente saggio, Paolo Galluzzi (2003) rende conto in maniera molto chiara di questa situazione di mutamento: la diffusione delle nuove tecnologie informatiche e le possibilità da esse offerte hanno favorito la nascita di quelle nuove entità che siamo soliti definire «biblioteche digitali». L'autore del saggio, evidenziando la consistenza «ibrida» di tali collezioni, intende riaffermare la necessità di escogitare nuovi stili cognitivi e nuove forme di rappresentazione del sapere: «in a Hybrid repository [...] information is not preserved within a rigidly compartmented space, but is like a knot in a cobweb» (*ibid.*: 100). La fedeltà ad una specifica metafora (il castello, il palazzo) implica infatti il riferirsi ad uno stile cognitivo da cui essa deriva e dipende: per questo modellare le strutture delle nuove biblioteche digitali su quelle preesistenti delle loro controparti reali significherebbe mettere il vino vecchio in botti nuove e nient'altro. Le regole d'accesso e i modi di studio elaborati nell'Ottocento per la lettura dei documenti ufficiali, poi estesi a tutte le altre forme di testi scritti, hanno raggiunto, se non superato, il punto di saturazione e non corrispondono più né alle possibilità tecniche né alle condizioni di vita di una società entrata nell'era dei mass media, cioè dell'uso simultaneo e articolato di testi, di immagini e di suoni, con sistemi di classificazione e selezione del materiale da consultare diversi da quelli ritenuti normali nel passato, come l'uso di indici e di sommari più o meno completi, o il gesto stesso di girare le pagine (Aymard, 1997: 8).

I processi di separazione e specializzazione dei saperi, delle arti e delle scienze hanno decretato la progressiva scomparsa di quelle testimonianze ete-

<sup>12</sup> Il riferimento è al testo pseudo ciceroniano conosciuto col titolo di *Rhetorica ad Herennium* ([Marco Tullio Cicerone] 1992).

rogenee dell'eredità culturale rappresentate dalle *Wunderkammern* e dai primi musei – tenuti insieme dal collante della curiosità personale del collezionista – soppiantati dalle moderne collezioni specializzate. Queste ultime presuppongono, tuttavia, lo stesso (alto) grado di specializzazione anche da parte del visitatore, altrimenti spiazzato dalla quasi totale assenza di punti di riferimento ed incapace di ricreare da solo quel contesto – quella trama – di relazioni in grado di rendere l'esperienza significativa. L'informatica può oggi aiutare a rimuovere queste barriere, e far sì che venga avviato un processo di ricostituzione del tessuto della conoscenza: a questo scopo è necessario adeguare alla nuova realtà quelle che Kant ebbe a definire le condizioni a priori della appercezione<sup>13</sup>, alla base della possibilità stessa del conoscere. Così sintetizza la questione Mario Ricciardi: «In Internet, in un sistema a ragnatela, neppure i segmenti devono necessariamente essere governati dal principio lineare e dalla regola logica della sequenzialità, poiché non sembra esistere un inizio che si protragga nel tempo fino a diventare il centro, ed essere riconosciuto quale autorità da cui tutto promana e dipende. Assumono importanza strategica altri elementi che si sottraggono al dominio “classico” tempo-spazio ma anche a quello lavoro-cultura [...]. La navigazione (altra metafora particolarmente significativa, che rinvia sempre ad acqua e aria, mai alla terra e quindi alla fisicità sensibilmente più probante) dentro questo sopra-mondo ha bisogno di bussole» (Ricciardi, 1997: 180-181).

Vorrei adesso ricapitolare brevemente alcuni degli elementi che sono emersi da questa relazione – e che ritengo, personalmente, determinanti. Ci siamo soffermati sull'accresciuta capacità umana di produrre di informazioni, quantificabile intorno al 36% su base annua (per il periodo dal 1999 al 2002) (cfr. Lyman *et al.*, 2003). Vorremmo adesso proporre una sorta di scala per valutare le esatte proporzioni del fenomeno: le nuove informazioni prodotte a livello planetario e immagazzinate su supporti tradizionali (carta, pellicola e supporti ottici e magnetici) ammontano a 5 exabytes su base annua, un numero prossimo a quello che esprime la quantità di parole pronunciate dall'uomo nel corso della sua

<sup>13</sup> «Nella estetica trascendentale, dunque, noi isoleremo dapprima la sensibilità, separandone tutto ciò che ne pensa coi suoi concetti l'intelletto, affinché non vi resti altro che l'intuizione empirica. In secondo luogo, separeremo ancora da questa ciò che appartiene alla sensazione, affinché non ne rimanga altro che la intuizione pura e la semplice forma dei fenomeni, che è ciò solo che la sensibilità può fornire a priori. In questa ricerca si troverà che vi sono due forme pure di intuizione a sensibile, come principi della conoscenza a priori, cioè spazio e tempo, del cui esame noi ci occuperemo ora». Cfr. Kant (1781; trad. it., 1963: 67).

storia (*ibid.*). Questa cifra, tuttavia, equivale a meno di un terzo della quantità di nuove informazioni veicolate attraverso canali come il telefono, la radio, la televisione ed Internet, che si aggira intorno ai 17.7 exabytes all'anno (*ibid.*). Come detto, dunque, le istanze di selezione e indicizzazione risultano di capitale importanza.

Per quanto riguarda le strategie di selezione, può essere interessante notare come – paradossalmente – in un futuro non troppo lontano, in cui il costo di un disco fisso della capacità di un terabyte ( $10^{12}$  bytes) sarà di fatto piuttosto abbordabile (al di sotto di 300 €), le migliori strategie di selezione dei contenuti (eccezion fatta per i flussi video in alta risoluzione) potrebbero consistere nella conservazione indiscriminata delle informazioni, giacché un processo di revisione dei dati, con conseguente cancellazione di una parte di essi, potrebbe risultare poco vantaggioso. Per riempire un disco fisso da un terabyte bisognerebbe infatti raccogliere ed archiviare – ogni giorno e per un intero anno – qualcosa come 9800 immagini, 26 ore di audio digitale o 4 ore di video digitale in alta risoluzione. Tuttavia abbiamo già espresso la convinzione che conservare tutto non significa necessariamente conservare nella maniera adeguata.

Anche in quest'ottica, per quanto concerne l'ottimizzazione dei processi di indicizzazione, risulta necessaria una revisione a livello categoriale che coinvolga i concetti utilizzati correntemente come metafore di riferimento per l'organizzazione dei dati: questo processo dovrebbe condurre verso un progressivo abbandono della modalità gerarchica di archiviazione, tuttora dominante, in favore di strategie meno vincolanti. Un esempio di gerarchizzazione più flessibile può essere costituito dalle cosiddette «fluid collections»<sup>14</sup>, in cui

<sup>14</sup> «In our implementation, document collections comprise three elements (each of which can be null). The first is a query term. Query terms are specified in terms of document attributes. Queries can test for the presence or absence of particular attributes on a document, can test the specific value of an attribute, or can perform type-specific value comparisons (for instance, a wide range of date specifications can be provided, such as “changed within 2 hours” and “modified last week”). Query terms in document collections are “live”, the collection contains the matching documents at any moment, so that documents may appear or disappear depending on their immediate state. Since these queries are embodied in collections, they are persistent. In addition to the query term, the document collection stores two lists of documents, called the inclusion and exclusion lists. Documents in the inclusion list are returned as members of the collection whether or not they match the query. Documents in the exclusion list are not returned as members of the collection even if they do match the query. When the query is null, the inclusion list effectively determines the collection contents. So, the contents of the collection at any moment are the documents in the inclusion list, plus those matching the query, minus those in the exclusion list. We call these three-part structures “fluid collections”». Cfr. Dourish *et al.* (2000: 140-170).

criteri di ordinamento statici (liste di inclusione o esclusione per un determinato elemento entro una specifica collezione) convivono con strategie dinamiche di organizzazione delle informazioni (queries del tipo «considerato un elemento X all'interno del database, controlla se per esso è verificabile o meno la condizione Y»): una struttura ibrida di questo tipo opera in maniera statica e gerarchica solo quando il risultato della query è nullo (in questo caso il contenuto della collezione è rappresentato dagli elementi presenti nella lista di inclusione). In tutti gli altri casi, il contenuto della collezione è il risultato della somma degli elementi contenuti nella lista di inclusione con quelli in grado di soddisfare la query, meno – ovviamente – quelli contenuti nella lista di esclusione.

Il progressivo abbandono di strategie di mera accumulazione delle informazioni, insieme all'esigenza di elaborare nuovi modelli di gestione della conoscenza appaiono, dunque, priorità difficilmente eludibili. Non dovremmo infatti mai dimenticare che il vero valore dei flussi informativi sta nella nostra capacità di utilizzarli come materiale per la costruzione della nostra stessa Storia: una rappresentazione articolata nel tempo e nello spazio della presenza dell'uomo e del suo passaggio<sup>15</sup>.

#### BIBLIOGRAFIA

- 2003 *World Population Data Sheet*, stilato dal Population Reference Bureau degli Stati Uniti [<http://prb.org> (10 febbraio 2004)].
- AYMARD M., 1997. *Quali archivi per quale storia*. In: M. MORELLI, M. RICCIARDI (a cura di), *Le carte della memoria. Archivi e nuove tecnologie*. Pref. di G. Vacca, Editori Laterza, Roma-Bari: 3-14.
- BARAGLI E., 1980. *Il caso McLuhan*. La Civiltà Cattolica, Roma.
- BIOY CASARES A., 1953. *La invención de Morel*. Praef. J.L. Borges. Emecé Editores, Buenos Aires (trad. it.: *L'invenzione di Morel*. Pref. J.L. Borges e G. Piovene, trad. L. Bacchi Wilcock, Bompiani, Milano 1966).
- BOLZONI L., 1995. *La stanza della memoria. Modelli letterari e iconografici nell'età della stampa*. Einaudi, Torino.
- BOLZONI L., 2002. *La rete delle immagini. Predicazione in volgare dalle origini a Bernardino da Siena*. Einaudi, Torino 2002.

<sup>15</sup> «The pinnacle of value is achieved when the user constructs a “story” out of media. By story, we mean a layout in time and space. [...] Stories create the highest value for two reasons: first, because the user will select the best media to include in the story; second, because the user will attempt to present the media in the most compelling manner». Cfr. Gemmill (2002).

- BORGES J.L., 1956. *Ficciones*. Emecé Editores, Buenos Aires (trad. it.: *La biblioteca di Babele*. Trad. F. Lucentini, Einaudi, Torino 1955).
- BUSA R., 1951a. *Rapida e meccanica composizione e pubblicazione di indici e concordanze di parole mediante macchine elettrocontabili*. *Aevum*, 25, 6: 479-493.
- BUSA R., 1951b. *Sancti Thomae Aquinatis hymnorum ritualium varia specimina concordantium. Primo saggio di indici di parole automaticamente composti e stampati da macchine IBM a schede perforate*. *Archivum philosophicum Aloisianum*, 2, Fratelli Bocca, Milano.
- BUSA R., 1952. *Rapidissima composizione di indici e concordanze di parole mediante schede perforate*. In: *La documentazione in Italia*. Atti del Congresso mondiale di documentazione XVIII Conferenza internazionale (Roma, 15-21 settembre 1951), riferiti alla partecipazione italiana, con aggiunta di studi e contributi riguardanti l'attività documentaria in Italia, a cura di G. Avanzi e B. Balbis, Consiglio Nazionale delle Ricerche, Roma: 95-97.
- BUSA R., 1992. *Thomae Aquinatis Opera Omnia cum hypertextibus in CD-ROM*. Editoria Elettronica Editel, Milano.
- BUSH V., 1945. *As we may think*. *The Atlantic Monthly*, 176(1): 101-108 (trad. it.: *Come possiamo pensare*. In: J.M. NYCE, P. KAHN (eds.), *Da Memex a Hypertext: Vannevar Bush e la macchina della mente*. Muzzio, Padova 1992).
- CAMILLO DELMINIO G., 1991. *L'idea del teatro*. A cura di L. Bolzoni, Sellerio, Palermo.
- CARPENTER E., 1973. *Eskimo realities*. Holt - Rinehart and Winston, New York.
- CARRUTHERS M.J., 1990. *The book of memory. A study of memory in medieval culture*. Cambridge Studies in Medieval Literature, 10, Cambridge University Press, Cambridge.
- CARRUTHERS M.J., 1998. *The craft of thought. Meditation, rhetoric, and the making of images, 400-1200*. Cambridge Studies in Medieval Literature, 34, Cambridge University Press, Cambridge.
- CARRUTHERS M.J., ZIOLKOWSKI J.M. (eds.), 2002. *The medieval craft of memory. An anthology of texts and pictures*. University of Pennsylvania Press, Philadelphia PA.
- DE KERCKHOVE D., 1995. *The skin of culture: investigating the new electronic reality*. Edited by C. Dewdney, Sommerville House Publishing, Toronto (trad. it.: *La pelle della cultura: un'indagine sulla nuova realtà elettronica*. A cura di C. Dewdney, trad. M.T. Carbone, Costa & Nolan, Genova 1996).
- DIELS H., 1922. *Die Fragmente der Vorsokratiker*. Vol. 2, Berlin.
- DOLLAR C.H., 1994. *La memoria elettronica e la ridefinizione della preservazione*. In: T. GREGORY, M. MORELLI (a cura di), *L'eclisse delle memorie*. Pref. di G. Salvini, Editori Laterza, Roma-Bari: 161-184.
- DOURISH P., EDWARDS K., LAMARCA A., LAMPING J., PETERSEN K., SALISBURY M., TERRY D., THORNTON J., 2000. *Extending document management systems with user-specific active properties*. Association for Computing Machinery. *Transactions on Information Systems*, 18(2): 140-170.
- DURANTI L., 1994. *La definizione di memoria elettronica: il passo fondamentale nella sua preservazione*. In: T. GREGORY, M. MORELLI (a cura di), *L'eclisse delle memorie*. Pref. di G. Salvini, Editori Laterza, Roma-Bari: 147-160.
- GALLUZZI P., 2003. *Information technology and research in the humanities*. In: M. VENEZIANI (a cura di), *Informatica e scienze umane. Mezzo secolo di studi e ricerche*. Lessico Intellettuale Europeo, 94, Olschki, Firenze.
- GEMMELL J., BELL G., LUEDER R., DRUCKER S., WONG C., 2002. *MyLifeBits: fulfilling the memex vision*. In: *Proceedings of ACM Multimedia 2002* (Juan-les-Pins, 1-6 December 2002). ACM Press, New York.
- INNIS H.A., 1950. *Empire and communications*. Clarendon Press, Oxford (trad. it.: *Impero e comunicazioni*. A cura di A. Miconi, trad. V. Lovaglio, Meltemi, Roma 2001).

- INNIS H.A., 1951. *The bias of communication*. University of Toronto Press, Toronto (trad. it.: *Le tendenze della comunicazione*. Pref. M. McLuhan, trad. A. Lorenzini, SugarCo., Milano 1982).
- KANT I., 1781. *Kritik der reinen Vernunft*. Johann Friedrich Hartknoch, Riga (*Critica della ragion pura*. A cura di V. Mathieu, trad. G. Gentile e G. Lombardo-Radice, 8 ed., Bari 1963).
- LUCKEN C., 2005. *Entre amour et savoir. Contlics de mémoire chez Richard de Fournival*. In: A. PARAVICINI BAGLIANI (a cura di), *La memoire du temps au Moyen Age*. SISMEL Edizioni del Galluzzo, Firenze.
- LYMAN P., VARIAN H.R., CHARLES P., GOOD N., JORDAN L.L., PAL J. (adiuv.), SWEARINGEN K. (coord.), 2000. *How Much Information? 2000*.  
[<http://www.sims.berkeley.edu/how-much-info-2003> (10 febbraio 2004)].
- LYMAN P., VARIAN H.R., CHARLES P., GOOD N., JORDAN L.L., PAL J. (adiuv.), SWEARINGEN K. (coord.), 2003. *How Much Information? 2003*.  
[<http://www.sims.berkeley.edu/how-much-info-2003> (10 febbraio 2004)].
- [Marco Tullio Cicerone], 1992. *Rhetorica ad Herennium. La retorica a Gaio Erennio*. A cura di F. Cancelli, Arnoldo Mondadori Editore, Milano.
- MCLUHAN M., 1962. *The Gutenberg galaxy. The making of typographic man*. University of Toronto Press, Toronto 1962 (trad. it.: *La galassia Gutenberg. Nascita dell'uomo tipografico*. Cur. G. Gamaleri, trad. S. Rizzo, Armando, Roma 1976).
- MCLUHAN M., 1964. *Understanding Media. The Extensions of Man*. McGraw-Hill, New York (trad. it: *Gli strumenti del comunicare*. trad E. Capriolo, Il Saggiatore, Milano, 1967).
- MORDENTI R., 2001. *Informatica e critica dei testi*. Bulzoni, Roma.
- NELSON T., 1981. *Literary machines. The report on, and of, project Xanadu concerning word processing, electronic publishing, hypertext, thinkertoys, tomorrow's intellectual revolution, and certain other topics including knowledge, education and freedom*. 3 ed., Swarthmore, PA.
- ORLANDI T., 1998a. *Teoria e prassi della codifica dei manoscritti*. In: M. PICONE, C. CAZALÈ BERARD (a cura di), *Gli Zibaldoni di Boccaccio. Memoria, scrittura, riscrittura*. Franco Cesati, Firenze: 349-360.
- ORLANDI T., 1998b. *From the Book to the Electronic Edition of Literary Texts*. In: I. BUTTERWORTH (a cura di), *The Impact of Electronic Publishing in the Academic Community*, Portland Press, London: 181-182.
- ORTOLEVA P., 2000. *Il problema dell'autorità*. Technology Review, 3: 42-45.
- [Platone], 1991. *Tutti gli scritti*. A cura di G. Reale, Rusconi, Milano.
- QUEMADA B., 2003. *Lexicographie et textes numérisés*. In: M. VENEZIANI (a cura di), *Informatica e scienze umane. Mezzo secolo di studi e ricerche*. Lessico Intellettuale Europeo, 94, Olschki, Firenze.
- RICCIARDI M., 1997. *Archivi mobili/archivi intelligenti*. In: M. MORELLI, M. RICCIARDI (a cura di), *Le carte della memoria. Archivi e nuove tecnologie*. Pref. G. Vacca, Editori Laterza, Roma-Bari: 178-231.
- RIDOLFI P., 1994. *Tecnologia e memoria: l'evoluzione dei calcolatori e la conservazione delle informazioni*. In: T. GREGORY, M. MORELLI (a cura di), *L'eclisse delle memorie*. Pref. G. Salvini, Editori Laterza, Roma-Bari: 247-255.
- ROBINSON P., 2003 *Where we are with Electronic Scholarly Editions, and where we want to be*. Jahrbuch für Computer philologie, 5: 125-146.
- ROSSI P., 1960. *Clavis universalis: arti mnemoniche e logica combinatoria da Lullo a Leibniz*. R. Ricciardi, Milano-Napoli.
- SARACCO R., 2000. *Dalla informazione alla conoscenza*. Technology Review, 3: 34-41.
- SIMMEL G., 1903. *Die Grosstädte und das Geistesleben*. Jahrbuch der Gehe-Stiftung Dresden, 9: 185-206 (trad. it.: *La metropoli e la vita dello spirito*. Trad. P. Jedlowski e R. Siebert, Armando, Roma 1995).

- SMITH B., 2003. *Activities and Research for Cultural Heritage*. In: M. VENEZIANI (a cura di), *Informatica e scienze umane. Mezzo secolo di studi e ricerche*. Lessico Intellettuale Europeo, 94, Olschki, Firenze.
- STURM H., 1988. *Perception and television. The missing half second*. In: *The work of Hertha Sturm. Working papers in communications*, McGill University, Montreal.
- YATES F.A., 1969. *The art of memory*. Peregrine Books, Harmondsworth (trad. it.: *L'arte della memoria*. Trad. A. Biondi, Einaudi, Torino 1972).



MARCO VENEZIANI\*

## MAG: A METADATA SET FOR MANAGING AND PRESERVING ITALIAN DIGITAL RESOURCES

The very recent introduction of the MAG format as part of the digitalisation process promoted by the «Biblioteca Digitale Italiana» (BDI)<sup>1</sup> does not, as yet, permit any kind assessment. It is however possible to outline the features of the format itself, without even having to go into overly technical detail, and clarify the needs MAG is expected to answer.

It is well known, to the point of having become a commonplace over the years, that at the heart of information technology there lies a distinction between application programs and data, between – that is – the matter and the process of elaboration: one can often treat any given set of information with a variety of programs, and conversely one may employ a given program for the treatment of different sets of data. It is also true, however, that the dynamics which concern data as such – the gathering, updating, modelisation thereof – evolve on a time scale which is at considerable variance with the fast moving development of software.

Recent developments in the information technology area have led to stricter harmonisation, if not symbiosis, of the two separate processes. Object programming is a good example of this achievement, with the inclusion in the operating entity itself – the «object» proper in I.T. jargon – of both fun-

\* Lessico Intellettuale Europeo e Storia delle Idee - Istituto del Consiglio Nazionale delle Ricerche - Via Carlo Fea, 2 - 00161 ROMA.

<sup>1</sup> The Comitato Guida of the Biblioteca Digitale Italiana, headed by prof. Tullio Gregory, has been established within the Direzione Generale per i Beni librari e gli Istituti culturali of the Ministero per i Beni e le Attività culturali. Having taken into account the possible areas of intervention and the resources available, it has promoted firstly a digitalisation campaign which involves historical catalogues held in Italian public libraries, secondly it has promoted the development of a system for accessing musical scores, both manuscript and in print, on the Internet, and finally it has initiated the digitalisation, for Internet publication, of Italian pre-unitarian periodicals.

damental data and qualifying functions. It is now also common of multimedia environments to provide the user with messages – sounds – sights which are fully «virtual», in the sense that they do not actually reside, as such, within the application's data set, but are elaborated dynamically by program functions which in turn develop the computation from basic statistical data. The Digital Object Identifier (DOI) architecture R. Kahn gave a talk on the day before yesterday offers another valid answer to the objective of neutralising somehow the data/application divide.

From the point of view of data preservation, on the other hand, this conflation of heterogeneous entities brings with it unprecedented difficulties: alongside data it is now in fact necessary to preserve the application programs too, and all of the supporting annexes and components – from specialised modules to the operating system, through to the platforms and hardware equipment.

This state of affairs we have sketched may account for the fact that the International Standard Bibliographic Description (ISBD), after considerable hesitation, attested itself on the wording «electronic resource» as opposed to «digital resource» to designate a specific class of items to be catalogued. Indeed «digital» cannot but refer to data on account of the impropriety, within a cataloguing framework, of applying to computer programs an adjective which is semantically opposed to «analogue». By ISBD rules, therefore, multimedia CD-ROMs, or an OPAC even, are to be filed, respectively, as local and remote «electronic resources» (granted that it is otherwise clear who should be doing the cataloguing – and how).

It is also worth pointing out that in our experience it is not possible, nor should it be, to conduct a digitalisation programme (when it is realistically aimed at providing *certain* innovative services, *certain* selected items) at a pace with the informatisation of cultural and administrative bodies, and possibly not even at the rate of development and growth of the Web as a network architecture.

In this perspective the phrase «digital resource» is here strictly employed in reference to data as *information*, unrelated to any kind of software. This definition is without doubt not only plausible but advantageous. The BDI, in fact, expressly recommends that in all digitalisation processes all data be *always* acquired and stored in a format which is independent of the choice of application program and platform – thus preserving it as a proper «digital resource», obviously of the *non* native kind.

We shall see shortly how MAG, from the point of view of the Open Archival Information System (OAIS) architecture, is on this side of data management and distribution and is rather conceived for defining, through the aggregation of various elements, a non-specific Submission Information Package (SIP) appropriate to any application.

On the other hand the current variety of available «formats», proprietary or otherwise, favours the autonomy of digital data, in that the fundamental features of such formats are their being well documented, multi-platform, and widely used: XML and UNICODE are successfully replacing, in the treatment of texts, the old ASCII «plain text»; images may be processed by choosing *ad lib* from TIFF, JPEG, PNG, etc.; sounds can be stored as WAV or MP3 files, not to mention the many other audio formats. Even data which originally comes under a combination of formats may be conveniently transcribed into special formats such as PDF.

The weak spot in taking such an approach may be illustrated by means of a fairly typical example. Some digitalisation programmes include, among other objectives, the conservation of the analogue originals in spite of the awareness that a digital copy cannot replace the microfilm obtained in compliance with national and international standards of quality, as codified by, for instance, the Istituto Centrale di Patologia del Libro (National Institute for Book Pathology)<sup>2</sup>.

Once there is available an easily accessible digital version of a document it becomes less difficult to preserve rare and valuable items, as also, in general, originals consisting of highly perishable supports and inks, such as newspapers.

It is common practice to create high resolution images of the item which are stored onto a CD or DVD ISO 9660 together with some additional information (author, title, edition) and a small scanning program for use at the designated workstations. We should point out that such software, being subordinated to the computer's operating system, becomes more readily obsolete than the actual images, and that in the long run this could lead to difficulties in reading the support. It would therefore be a better solution to file the CDs

<sup>2</sup> Cfr. *Disciplinare tecnico sulla microfilmatura di giornali, periodici e materiale bibliografico di pregio*, notice n. 75/95 of the Istituto centrale per la patologia del libro, July 1995, circulated by the Ufficio centrale per i beni librari, le istituzioni culturali e l'editoria of the Ministero per i beni culturali e ambientali.

and divulge the images on an Intranet network built on more standard software.

A related problem, and not a minor one, concerns the accompanying information, which is generally text, and which is generally dependent – in terms of contents, codification, and structure – on the program used for accessing and using the information. It is instead vital that such critical data share the same technological fate as the images, rather than the fairly ephemeral fate of the viewing program. It is vital, that is, that the metadata – to call it by its proper name – be re-utilisable, improvable, increasable, without any undue restrictions.

The technical features of Extensible Markup Language (XML) have made it possible to introduce advancements on this particular aspect of digital technology. XML files, notably, contain annotated text of a standard formally described in Document Type Description (DTD) files, which are also text files. Alongside the human user, XML files can also be read by any application program enabling it to verify, by reading the DTD file, that the organisation of the text follows the specified criteria without undesired omissions, additions, or dislocations. This method is so accurate that, by means of appropriate instructions in the DTD, it is possible to obtain an annotated text as accurate in the syntax as can be obtained through Data Description Languages (DDLs) in data-base applications.

It is obvious that XML nowadays is the most apt format for collecting metadata independently of any specific program. It is however necessary to define, within an apposite DTD, the set of markup strings (tags, labels) and terminological and syntactic restraints to be adhered to: this set goes under the name of XML application profile. The application profile is in turn able to import and match tags from other profiles defined elsewhere on the Web in order to provide standardised answers to certain tasks.

The Metadata Encoding and Transmission Standard (METS), which codifies and transfers metadata, has been elaborated by the Library of Congress as a part of the «American Memory» project and is precisely an application profile for the implementation of a compact metadata container. Following criteria which have become conventional, the layout discriminates descriptive and management metadata, administrative metadata and the metadata which codifies the digital structure of the digital item. Each typology of metadata finds within a METS file its proper collocation.

MAG, for its own part<sup>3</sup>, is modelled on METS architecture, though it is in some aspects simplified; it also imports the *Dublin Core* and *NISO Data Dictionary* (*Technical metadata for digital still images*) standards – both of which are well-known XML applications devoted to the summary description of generic electronic resources the first, and to the strictly technical description of digital images the second. Thus conceived, the MAG file (which is a «metafile» or «guiding-file») is permanently associated to the digital resource it is devoted to – be it, in origin, a manuscript, periodical, monograph – and can be accessed by any application program for extracting the relevant information.

Now in its 1.0 version, after many months of testing at the Biblioteca Nazionale Centrale of Florence and other Italian libraries<sup>4</sup>, the MAG system has six distinct sub-sections, not all of which are compulsory (fig. 1):

GEN, general information regarding the acquisition of the resource;

BIB, descriptive metadata – complying with a standard explained below – regarding the analogue original for the digital resource; in its most typical form it is an SBN (National Bibliographic System) Index record for that item;

STRU, description of the internal organisation of the resource, and a further description of the images contained in each subdivision;

IMG, an element provided with each image, detailing the technical and management data relating to it;

OCR, provides the keys necessary to a contents-based search within the text; it accompanies images processed with OCR/ICR software;

DOC, devoted to multipage resources, such as WORD, RTF, or PDF files.

We shall now go into the detail of each section explaining the functions which have been supplied to each, with the aid of the diagrams and examples devised by Maria Grazia Pepe, of the Biblioteca Nazionale Centrale of Florence.

<sup>3</sup> Technical specifications of MAG version 1.0, Web resources, and accompanying documentation are available at the address <http://www.iccu.sbn.it/schemag.htm>

<sup>4</sup> A brief history of the standard, which was developed by the Gruppo di studio sugli standard e le applicazioni di metadati nei beni culturali, set up as part of the ICCU - Istituto Centrale per il Catalogo Unico in the year 2000, and to which representatives of libraries, museums and archives have contributed, is available at the address <http://www.iccu.sbn.it/grupmeta.htm>

http://www.iccu.sbn.it/MAG/metadigit.xsd

15-11-2003 11.20

```
<?xml version="1.0" encoding="UTF-8"?>
<xsd:schema targetNamespace="http://www.iccu.sbn.it/metaAG1.pdf"
  xmlns="http://www.iccu.sbn.it/metaAG1.pdf"
  xmlns:xsd="http://www.w3.org/2001/XMLSchema"
  elementFormDefault="qualified" attributeFormDefault="unqualified">
<!--
4/11/2002 Changes:
[...]
19/05/2003 Changes:
1- Added tag <collection>, and subsection <img_group> to <gen> section
2- Added complete set of DC elements to <bib> section
3- Added attribute level to <bib> section
4- Section <holdings> is repetitive
5- Added tag <library> to section <holdings>
6- Renamed <fis> to <piece>
7- Added tag <element> to section <stru>
8- Added <doc> section
9- Added imggroupID attribute to <img> and <altimg> sections
10- Added holdingsID attribute to <img> section
11- Added <image_metrics> section
12- Moved tag <dpi> to section <image_metrics> and added tag <ppi>
13- Moved tag <photometricinterpretation> from section <format> to <image_metrics>
14- Renamed <size> to <filesize>
15- Added section doc
-->
<xsd:include schemaLocation="metatype.xsd"/>
<xsd:element name="metadigit">
  <xsd:complexType>
    <xsd:sequence>
      <xsd:element name="gen" type="gen"/>
      <xsd:element name="bib" type="bib"/>
      <xsd:element name="stru" type="stru" minOccurs="0" maxOccurs="unbounded"/>
      <xsd:element name="img" type="img" minOccurs="0" maxOccurs="unbounded"/>
      <xsd:element name="ocr" type="ocr" minOccurs="0" maxOccurs="unbounded"/>
      <xsd:element name="doc" type="doc" minOccurs="0" maxOccurs="unbounded"/>
    </xsd:sequence>
    <xsd:attribute name="version" type="xsd:string" default="1.0"/>
  </xsd:complexType>
  <!-- vincolo di unicita' sul campo sequenza all'interno di ogni sezione -->
  <xsd:unique name="uniquimg">
    <xsd:selector xpath="img"/>
    <xsd:field xpath="sequence_number"/>
  </xsd:unique>
  <xsd:unique name="uniquocr">
    <xsd:selector xpath="ocr"/>
    <xsd:field xpath="sequence_number"/>
  </xsd:unique>
  <xsd:unique name="uniqudoc">
    <xsd:selector xpath="doc"/>
    <xsd:field xpath="sequence_number"/>
  </xsd:unique>
</xsd:element>
</xsd:schema>
```

Fig. 1.

In the meantime we may point out that each item in a series of images (IMG) is associated to a sequence number, or protocol number, which is strictly exclusive to that item. The corresponding items in the STRU, OCR and DOC sub-sections are attributed the same sequence number.

As for describing the context of data acquisition, it is possible to specify information of different order within the GEN section (fig. 2), including a standard name for the project and a second name for the collection the resource belongs to. A typical project may in fact include more than one digital collection, or, conversely, a single collection may be dealt with in the course of more than one project.

We may also note that the repeatable element IMG\_GROUP enables to state the set of technical information which pertains to more than one image just once; to avoid repeating the same information for each image, it is sufficient to indicate the group identification number contained in the ID attribute, within the IMG element. We shall discuss some of this information further on.

The BIB section (fig. 3) is for descriptive metadata relating to the analogue original of the digital resource. In its most characteristic form this information coincides wholly, or at least in part, with the SBN Index bibliographic record, or with a record from any other standardised cataloguing archive: it follows that this layout is conceived for items which already have been catalogued.

All descriptive data is coded as repeatable elements within the namespace *Dublin Core*, but only *dc:identifier* is compulsory and contains, most relevantly, a univocal pointer for the source catalogue and to the record contained therein. The MAG standard includes a public register of these pointers and of their typology, paired with an accurate description of the underlying format and of the referenced data. All the other *Dublin Core* elements are primarily documentary, as is also the localising element HOLDINGS, which is repeatable in the event of constituting digital resources from more than one original.

The PIECE element, finally, collects all the identifying data relating to serial and periodical items, and alternatively permits multi-level cataloguing.

The third section, STRU, is the briefest of XML descriptions (fig. 3) but it absolves the fundamental role of coupling the digital resource with the internal organisation of the original it reproduces. An identification name is pro-

### Section GEN

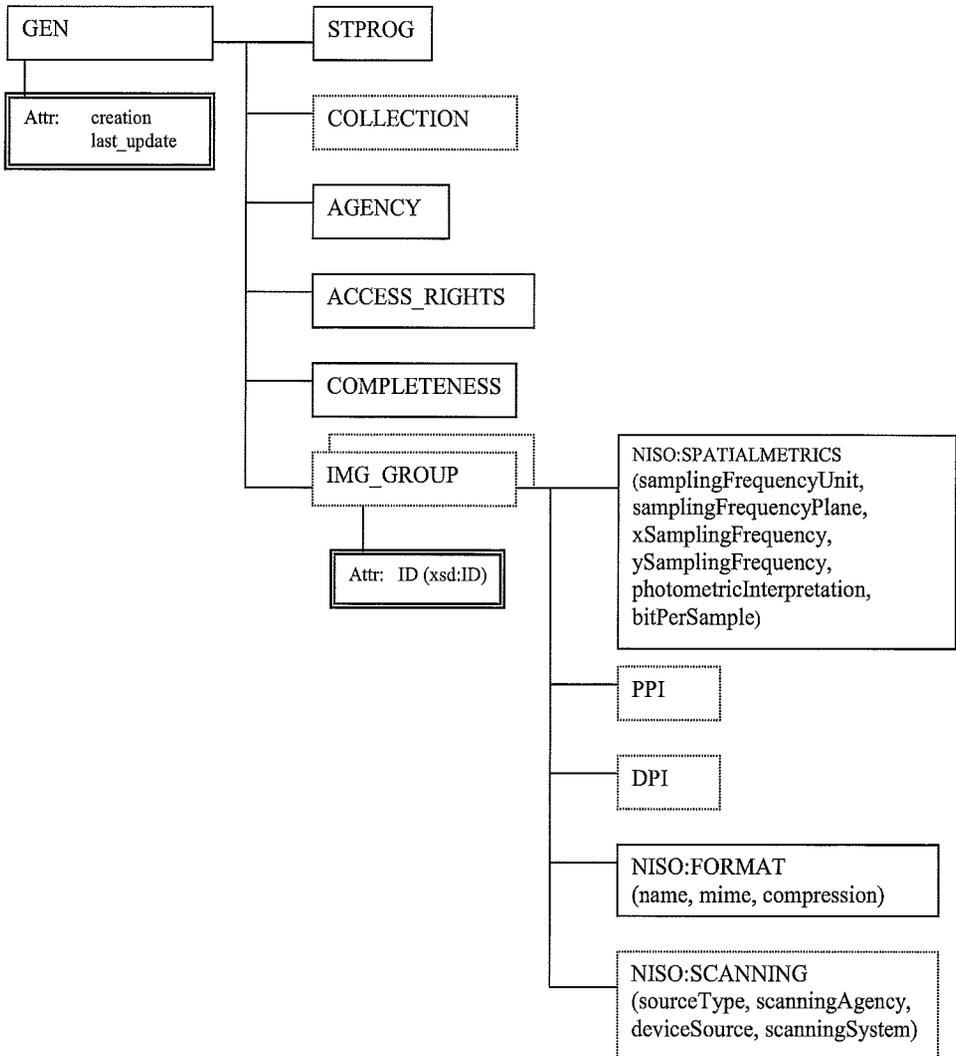
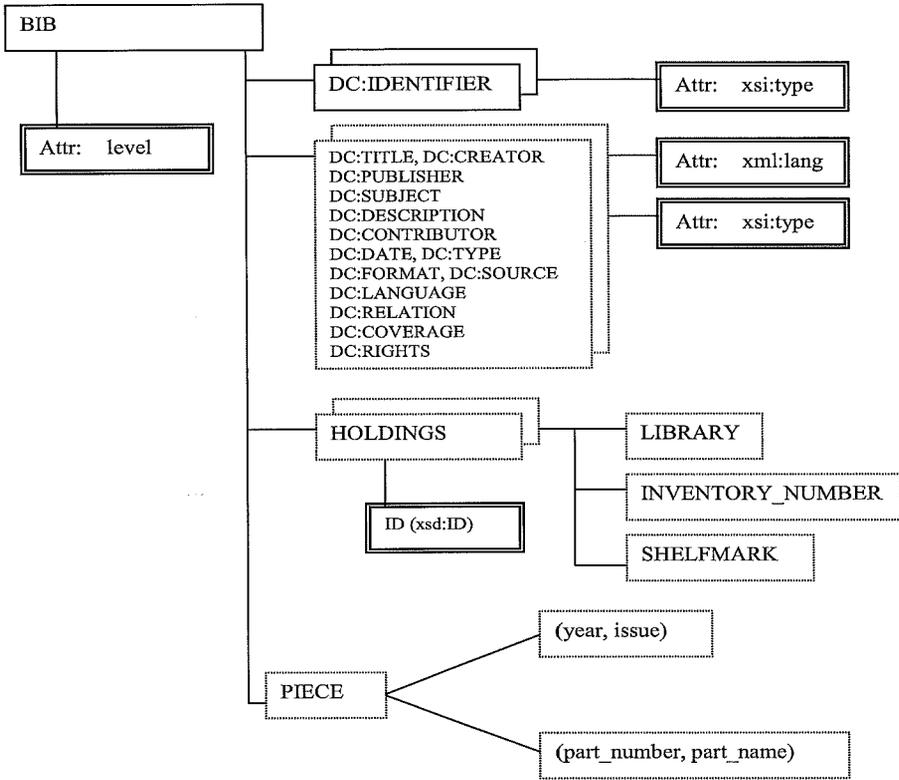


Fig. 2.

### Section BIB



### Section STRU

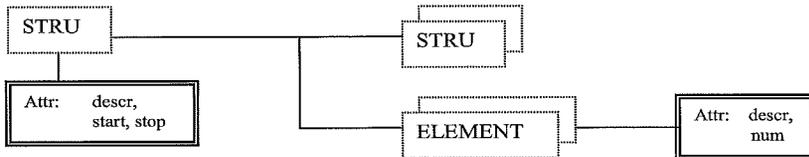


Fig. 3.

vided for each subdivision, and the images belonging to each subdivision are traceable by means of their sequence number.

On account of the fact that, at it is known, the most common structures are hierarchically ordered, the STRU section is the only recursive section of the MAG format; any STRU element may therefore incorporate one or more other STRU elements.

We have paid special attention to make the structural description entirely autonomous from the actual typology of the images: digital resources which, for instance, are available in alternative sets of images (medium and low resolution, *e.g.*) do not therefore require a duplication of the STRU section. It is on the other hand possible to implement different MAG descriptions of the same resource, for instance permitting it to be viewed sequentially or thorough direct access to a component part.

This layout is very flexible and it allows among other things to collect and store within a MAG file all the information relating to the analytical description of particular cataloguing items, such as the articles included in a periodical series, or the papers included in conference proceedings (fig. 4).

The IMG element, to be repeated with each image (fig. 5), specifies the technical and management data of the image, such as disk storage, resolution, colour, etc. The tag formats largely draw on the NISO namespace; through its two attributes, namely *holdingsID* and *imgGroupID*, the element may be related to a specific exemplary in the original reported in the BIB section or it may refer to a set of technical information previously defined in the GEN section. When the same information is specified more than once, the most recent is taken into consideration.

It is also important to note that alternative images of one same original page may be recorded within the sub-element ALTIMG, for instance images of varying resolution or black and white vs. colour images.

The OCR element (fig. 6), where necessary, relates an image to the keys necessary for a contents-based search. The keys are generally obtained by means of optic scanning of the text with OCR/ICR software, but the OCR element is also adaptable to transcriptions of the original document, as well as in the case of freely allotted terms, such as in subject-indexing.

Of the final DOC section (fig. 7), also repeatable, it suffices to say it is mainly appropriate for multi-page digital resources, such as WORD, RTF, or PDF files.

[http://www.iccu.sbn.it/MAG/Med\\_506.xml](http://www.iccu.sbn.it/MAG/Med_506.xml) - Excerpta

15-11-2003 11.20

```
<?xml version="1.0" encoding="UTF-8" ?>
<!-- L'esempio è stato realizzato grazie alla collaborazione dell'Istituto e
Museo di Storia della Scienza (http://www.imss.firenze.it) -->
<metadigit xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:niso="http://www.niso.org/pdfs/DataDict.pdf"
xmlns:xlink="http://www.w3.org/TR/xlink"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns="http://www.iccu.sbn.it/metaAG1.pdf"
xsi:schemaLocation="http://www.iccu.sbn.it/metaAG1.pdf metadigit.xsd"
version="1.0">
<gen creation="2003-03-21T11:25:00" last_update="2003-03-21T11:25:00">
<stprog>Galileo@Techa</stprog>
<agency>IMSS</agency>
<access_rights>0</access_rights>
<completeness>0</completeness>
</gen>
<bib level="m">
<dc:identifier>CCF300678</dc:identifier>
<dc:title>Di Gio. Battista Baliano Opere diuerse</dc:title>
<dc:creator>Baliani, Giovanni Battista (1582-1666)</dc:creator>
<dc:publisher>In Genoua : per Pietro Giouanni Calenzani..., 1666</dc:publisher>
<holdings>
<inventory_number />
<shelfmark>MED 506</shelfmark>
</holdings>
</bib>
<stru descr="Copertine" start="1" stop="2" />
<stru descr="Carte di guardie" start="3" stop="4" />
<stru descr="Frontespizio" start="5" stop="6" />
<stru descr="Indice" start="7" stop="8" />
<stru descr="Avviso ai lettori" start="8" stop="8" />
<stru descr="Lettere" start="9" stop="18" />
<stru descr="Dialogo Primo" start="19" stop="56" />
<stru descr="Dialogo Secondo" start="57" stop="75" />
[...]
<stru descr="Dialogo Quinto" start="120" stop="138" />
<stru descr="Dell'amicitia" start="139" stop="157" />
<stru descr="Delle onde del mare" start="158" stop="200" />
<stru descr="Trattatello della Lettera di Cambio" start="201" stop="244" >
<stru descr="Della Lettera di Cambio. Cap. I" start="203" stop="203" />
<stru descr="Del Cambio. Cap. II" start="204" stop="204" />
<stru descr="Della Fiera. Cap. III" start="204" stop="206" />
<stru descr="Dello scudo di Marca. Cap. IV" start="206" stop="206" />
<stru descr="De cambio che si faceva in Lione, e per Lione. Cap. V"
start="207" stop="207" />
<stru descr="De' Prezzi con i quali si pagavano i cambij fatti in Lione.
Cap. VI" start="208" stop="208" />
<stru descr="Dei pagamenti in Lione. Cap. VII" start="208" stop="209" />
[...]
</stru>
[...]
<stru descr="Postfazione" start="349" stop="350" />
<stru descr="Carte di guardia" start="351" stop="352" />
<stru descr="Copertine" start="353" stop="354" />
<img>
<sequence_number>1</sequence_number>
<nomenclature>Copertina: 1</nomenclature>
[...]
```

Fig. 4.

### Section img

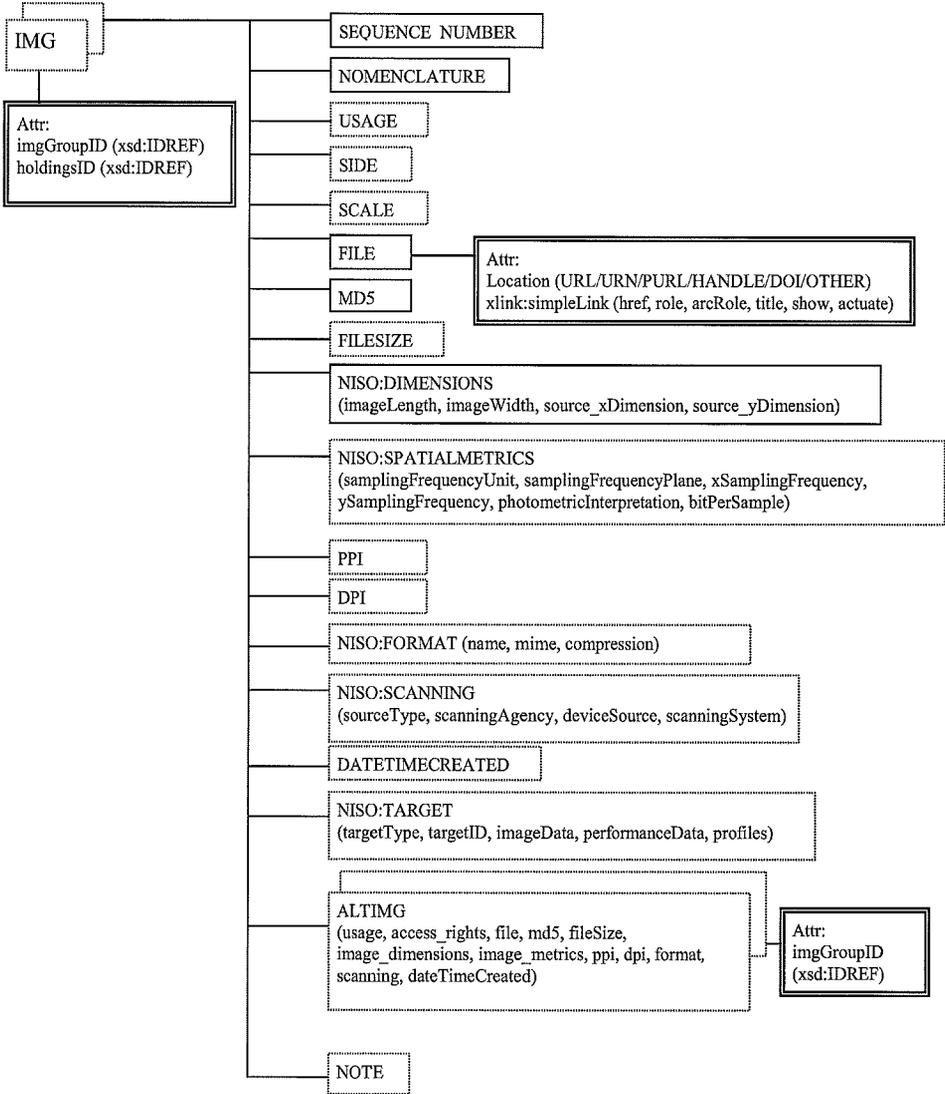


Fig. 5.

## Section OCR

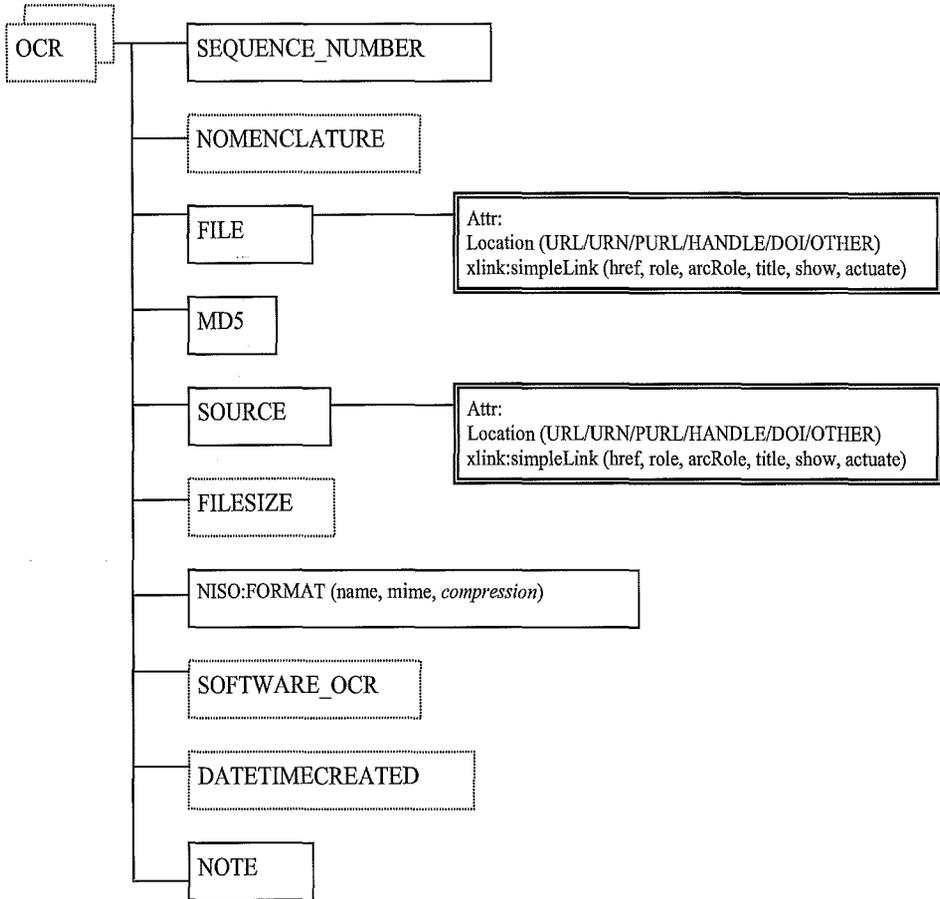


Fig. 6.

## Section DOC

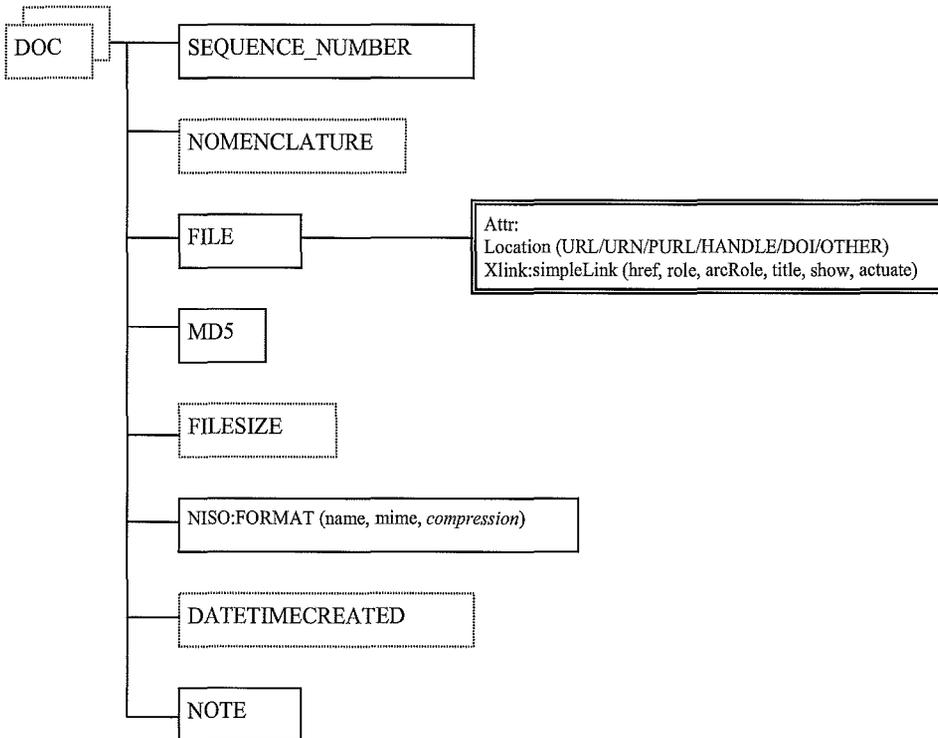


Fig. 7.

It is evident that so far the MAG scheme does not cover all typologies of digital resources. There are still no apposite sections for photographs, audio files, geographical maps and other iconographic items. They will be included when the need occurs, making sure that later versions of the standard remain compatible with the earlier ones. Open source software components would be auspicious, if they were enabled to convert a MAG file into the main corresponding formats; alternatively it would be feasible to offer digitalisation projects an application expressly dedicated to the conversion of metadata into

MAG files. The accompanying literature too ought to provide additional explanations and examples, whereas, on the other hand, counselling and technical support to MAG users cannot but be a permanent duty.

For these latter reasons especially there is a national MAG Group<sup>5</sup> dedicated to developing the version 1.0, on the basis of the improvements and changes which occur in the course of actual digitalisation projects. There is a lot to do yet, but we seem to have set the work on the right track.

<sup>5</sup> Members of the MAG committee are: Matilde Amaturò, Marco Lattanzi (Istituto Centrale per il Catalogo e la Documentazione); Francesco Baldi (Discoteca di Stato); Giovanni Bergamin, Maria Grazia Pepe (Biblioteca Nazionale Centrale di Firenze); Gianfranco Crupi (Università degli Studi «La Sapienza» di Roma); Gloria Cirocchi, Simona Gatta (Biblioteca della Camera dei Deputati); Claudio Leombroni (Rete Bibliotecaria di Romagna - Provincia di Ravenna); Cristina Magliano, Patrizia Martini (ICCU); Maurizio Messina (Biblioteca Marciana di Venezia); Antonella Mulè (Direzione Generale per gli archivi); Antonio Scolari (Centro Servizio Bibliotecario della Facoltà di Ingegneria - Università degli Studi di Genova); Marco Veneziani (ILIESI-CNR Roma); Paul Gabriele Weston (Università degli Studi di Pavia). Information and contacts at: <http://www.iccu.sbn.it/comimag.htm>



## INDICE

COMITATO ORDINATORE .....	Pag.	4
M. BRUNORI – Parole di saluto .....	»	5
M. GUERCIO – Depositi digitali per la conservazione del patrimonio culturale. Un workshop di ERPANET e del Centro Linceo interdisciplinare .....	»	7
T. ORLANDI – Archivi e informazione .....	»	25
D. BUZZETTI – Biblioteche digitali e oggetti digitali complessi. Esaustività e funzionalità nella conservazione .....	»	41
P. MOSCATI – Linguaggi di marcatura per la conservazione e la valorizzazione dell’informazione archeologica .....	»	77
R. MORDENTI – Pubblicazione delle ricerche umanistiche in ambiente digitale .....	»	95
A. WILSON – Access across time: how the National Archives of Australia preserves digital records .....	»	119
S. HEUSCHER – Workflow modeling language evaluation for an archival environment .....	»	131
E. DEGL’INNOCENTI – Trame e fili di Arianna nella massa digitale .....	»	155
M. VENEZIANI – MAG: a metadata set for managing and preserving Italian digital resources .....	»	173



## AVVERTENZA

*Nella collezione dei “Contributi del Centro Linceo Interdisciplinare Beniamino Segre” sono apparse le seguenti pubblicazioni:*

1. M. AGENO, *Punti di contatto tra fisica e biologia* (con una Prefazione di Beniamino Segre. Corso di dieci lezioni tenute dal 22 al 26 maggio 1972), 1974.
2. B. ROSSI, *Astronomia in raggi X* (Lezioni tenute nel febbraio e marzo 1972, raccolte da Bianca Maria Belli), 1974.
3. B. TOUSCHEK, *Sull'insegnamento della teoria dei quanti* (Lezioni tenute nell'aprile 1972), 1975.
4. P.A.M. DIRAC, *The Development of Quantum Mechanics* (Conferenza tenuta il 14 aprile 1972), 1974.
5. V.C.A. FERRARO, *Il vento solare ed il campo magnetico interplanetario* (Conferenza tenuta il 17 aprile 1972), 1974.
6. Seminari su: «*La Scienza dei Sistemi*» (con una Prefazione di Beniamino Segre). Parte Prima (I Seminario: 30 novembre-4 dicembre 1970; II Seminario: 11-15 gennaio 1971), 1975.  
Parte Seconda (III Seminario: 8-12 marzo 1971; IV Seminario: 5-9 aprile 1971; V Seminario: 3-7 maggio 1971; VI Seminario: 24-28 maggio 1971), 1975.
7. Seminario sulla: «*Evoluzione Biologica*» (Roma, 10-11 gennaio, 17-19 aprile 1974), 1975.
8. Y. NE'EMAN, *Patterns and Symmetry in the Structure of Matter* (Conferenza tenuta il 15 dicembre 1973), 1975.
9. A. SEIDENBERG, *Constructions in Algebra* (Riassunto delle lezioni tenute nell'ottobre e novembre 1972), 1975.
10. Tavola rotonda sul tema: «*Problemi matematici ed economici odierni sulle assicurazioni*» (Roma, 24-25 novembre 1972), 1975.
11. R. CAMPA, *La guerra e il processo di trasformazione tecnologica* (Conferenza tenuta il 26 maggio 1975), 1975.
12. M. MEDICI, *Indirizzi verso motori automobilistici meno inquinanti* (Conferenze tenute nel marzo 1973), 1975.
13. Colloquio sul tema: «*Le tecniche di classificazione e loro applicazione linguistica*» (Firenze, 13 dicembre 1972), 1975.
14. R.R. GATTO, *Interazioni elettromagnetiche, invarianza di scala e sue possibili estensioni* (Lezioni tenute nel settembre 1972), 1976.
15. II Seminario sulla: «*Evoluzione Biologica*» (Roma, 19-22 febbraio 1975), 1976.

16. E. DE GIORGI, *Convergenza in energia di operatori ellittici* (Conferenza tenuta nel febbraio 1974), 1976.
17. G.C. MOISIL, *Sur l'emploi des Mathématiques dans les Sciences de l'homme* (Conferenza tenuta il 5 giugno 1972), 1976.
18. A. ANDREOTTI, *Lewy Problem for Cauchy-Riemann Equations* (Lezioni tenute nel febbraio 1973), 1976.
19. D. ALFONSI, M.I. BALLA, F. DE SANTIS, G. GIORGI, M. SCHAERF, *Struttura di un sistema informativo per un servizio di documentazione scientifica* (Da una manifestazione tenuta nel febbraio 1976 per iniziativa del Centro Linceo e dell'Università di Roma), 1976.
20. C.A. TRUESDELL, *Termodinamica razionale* (Corso di lezioni tenute nel gennaio 1973), 1976.
21. A. TOGNOLI, *Introduzione alla teoria degli spazi analitici reali* (Lezioni tenute nel febbraio 1973, raccolte da Dina Smit Ghinelli), 1976.
22. A. HANSON, T. REGGE, C. TEITELBOIM, *Constrained Hamiltonian Systems* (Ciclo di lezioni tenute dal 29 aprile al 7 maggio 1974), 1976.
23. H. CHESTNUT, *Influence of Technology on Modern World Evolution and Use of Dynamic Models of Macro-Economic Systems in Development Planning* (Conferenza tenuta il 21 novembre 1972), 1976.
24. A. ANDREOTTI, *Introduzione all'analisi complessa* (Lezioni tenute nel febbraio 1972), 1976.
25. T. REGGE, M. RASETTI, *Vortices and Current Algebra* (Conferenze tenute nel giugno 1975), 1976.
26. G. SANSONE, *Studi sulle equazioni differenziali ordinarie nell'ultimo cinquantennio* (Lezione tenuta il 12 dicembre 1975), 1976.
27. E. SEGRÉ, *Personaggi e scoperte nella Fisica contemporanea* (Ciclo di lezioni tenute dal novembre 1972 fino al marzo 1973), 1976.
28. Seminario sui: «*Sistemi di reperimento e selezione automatica dell'informazione*» (Roma, 17-21 aprile 1972), 1976.
29. Seminario sulle: «*Applicazioni della Scienza dei Sistemi alla Medicina e alla Chirurgia*» (Roma, 22-26 maggio 1972), 1976.
30. Convegno Internazionale sul tema: «*Trends in the Physics and Engineering of Technological Materials*» (Roma, 17-19 ottobre 1973), 1976.
31. Gruppo di studio sui: «*Fenomeni di alta energia nelle ultime fasi dell'evoluzione stellare*» (Roma-Frascati, 29 maggio-16 giugno 1972), 1976.
32. III Seminario sulla: «*Evoluzione Biologica. Il codice genetico*» (Roma, 26-28 febbraio 1976), 1977.
33. Seminario sul tema: «*Una nuova via italiana alla fisica delle alte energie: Ada, Adone ...* » (Roma, 24-25 maggio 1974), 1977.
34. G. DIRAC, *Cardinal-determining Subgraphs of Infinite Graphs* (Lezione tenuta il 16 aprile 1975), 1977.

35. H. LEWY, *On the Boundary Behavior of Holomorphic Mappings* (Lezione tenuta il 3 maggio 1976), 1977.
36. R. DUBOS, *The Resilience of Ecosystems* (Lezione tenuta il 17 dicembre 1976), 1977.
37. Seminario sul tema: «*Rapporti tra Biologia e Statistica*» (Roma, 19-20 dicembre 1975), 1977.
38. R. BAER, *Finite Metanilpotent Groups and Finite Sylow Tower Groups* (Corso di lezioni tenute nell'aprile 1976), 1977.
39. L. CESARI, *Nonlinear Analysis and Alternative Methods* (Ciclo di lezioni tenute nell'aprile 1974), 1977.
40. Convegno Internazionale: «*Problemi connessi con l'utilizzazione pacifica dell'energia nucleare in Italia*» (Roma, 12-14 aprile 1976), 1977.
41. IV Seminario sulla: «*Evoluzione biologica*» (Roma, 17-19 febbraio 1977), 1978.
42. VI. ISTRATESCU, *Topics in Linear Operator Theory* (Corso di lezioni tenute nell'aprile 1976), 1978.
43. Convegno sul tema: «*Applicazioni del teorema del punto fisso all'analisi economica*» (Roma, 9-11 marzo 1977), 1978.
44. Congresso Internazionale su: «*L'insegnamento integrato delle Scienze nella scuola primaria*» (Roma, 7-15 gennaio 1976), 1979.
45. E. MARTINELLI, *Introduzione alla teoria delle classi caratteristiche: uno sguardo panoramico* (Corso di lezioni tenute nel febbraio e marzo 1978. Redatte da Guido Lupaccioli e Paolo Piccinni), 1979.
46. A.M. ANGELINI, *Linee di sviluppo nella utilizzazione della energia solare* (Conferenza tenuta il 9 marzo 1979), 1979.
47. A. BIETTI, *Modelli matematici e statistici applicati all'Archeologia e alla Paletnologia* (Conferenza tenuta il 16 giugno 1978), 1979.
48. V Seminario sulla: «*Evoluzione Biologica. Evoluzione della sessualità ed evoluzione umana*» (Roma, 23-25 febbraio 1978), 1979.
49. A. GOLINI, *Attuali tendenze della popolazione in Italia: problemi e prospettive* (Conferenza tenuta il 9 febbraio 1979), 1979.
50. S. DE BENEDETTI, *Dall'universo di Newton a quello di Einstein* (Conferenza tenuta il 28 maggio 1979), 1979.
51. VI Seminario sulla: «*Evoluzione Biologica. Ecologia ed etologia*» (Roma, 22-24 febbraio 1979), 1980.
52. D. LAUGWITZ, *The Theory of Infinitesimals. An Introduction to Nonstandard Analysis* (Ciclo di lezioni tenute nel marzo 1977), 1980.
53. International Meeting on: «*Astrophysics and Elementary Particles, Common Problems*» (Rome, 21st-23rd February 1980), 1980.
54. G. CARERI, *Ordine e disordine nella materia. Tre lezioni sugli aspetti interdisciplinari* (Roma, 26, 28 e 30 novembre 1979), 1981.

55. W.A. ADKINS, A. ANDREOTTI, J.V. LEAHY, *Weakly Normal Complex Spaces*, 1981.
56. C. SAPORETTI, *Risultati e prospettive dell'analisi dei testi accadici mediante il calcolatore elettronico* (Conferenza tenuta il 14 dicembre 1979), 1981.
57. VII Seminario sulla: «*Evoluzione Biologica e i grandi problemi della Biologia*» (Roma, 28-29 febbraio - 1 marzo 1980), 1981.
58. S. RANZI, *L'embriologia: recenti studi a livello molecolare* (Conferenza tenuta il 7 marzo 1980), 1981.
59. Convegno sul tema: «*Problemi di popolazione: realtà attuali e prospettive*» (Roma, 13 giugno 1980), 1981.
60. VIII Seminario sulla: «*Evoluzione Biologica e i grandi problemi della Biologia. Aspetti biologici e sociali: parassitismo e simbiosi*» (Roma, 25-27 febbraio 1981), 1982.
61. A. BIETTI, *Tecniche matematiche nell'analisi dei dati archeologici* (Ciclo di tre conferenze tenuto nel dicembre 1980), 1982.
62. T. ORLANDI, *La filologia al calcolatore. Nuove prospettive per la letteratura copta* (Conferenza tenuta il 12 marzo 1982), 1982.
63. A. DE LUCA, *La teoria generale dei codici* (Conferenza tenuta il 12 febbraio 1982), 1982.
64. IX Seminario sulla: «*Evoluzione Biologica e i grandi problemi della Biologia*» (Roma, 24-26 febbraio 1982), 1983.
65. Convegno sul tema: «*Il miglioramento genetico dei cereali*» (Celebrazione del 40° anniversario della morte di Nazareno Strampelli - Roma, 10 dicembre 1982), 1983.
66. P. MOSCATI, *Ricerche matematico-statistiche sugli specchi etruschi*, 1984.
67. E. MARTINELLI, *Introduzione elementare alla teoria delle funzioni di variabili complesse con particolare riguardo alle rappresentazioni integrali*, 1984.
68. X Seminario sulla: «*Evoluzione Biologica e i grandi problemi della Biologia. L'addomesticazione degli animali e delle piante*» (Roma, 24-26 febbraio 1983), 1984.
69. Giornata di Studio sul tema: «*Archeometria. Scienze esatte per lo studio dei Beni Culturali*» (Roma, 31 maggio 1983), 1985.
70. P.E. ARIAS, V.C. DI BARI, G. ORSOLINI RONZITTI, *La ceramica attica a figure nere e rosse del Corpus Vasorum Antiquorum. L'analisi computerizzata dei dati*, 1985.
71. XI Seminario sulla: «*Evoluzione Biologica e i grandi problemi della Biologia. L'evoluzione del comportamento e del sistema nervoso*» (Roma, 1-3 marzo 1984), 1985.
72. Giornate di studio introduttive ai Seminari sulla: «*Scienza e Ingegneria dei Sistemi nelle sue più rilevanti applicazioni*» (Roma, 3-4 maggio 1983), 1985.

73. XII Seminario sulla: «*Evoluzione Biologica e i grandi problemi della Biologia. Lo svolgimento della Genetica e dell'Evoluzione dopo la riscoperta delle leggi di Mendel*» (Roma, 13-15 febbraio 1985), 1986.
74. P. MOSCATI, *Analisi statistiche multivariate sugli specchi etruschi*, 1986.
75. Convegno sul tema: «*Nuove frontiere dell'informatica: i sistemi esperti*», in collaborazione con la FINSIEL (Roma, 13-14 dicembre 1984), 1986.
76. Meeting on: «*Finite Thermoelasticity*» (Rome, 30th-31st May-1st June 1985), 1986.
77. Seminario su: «*La Scienza e l'Ingegneria dei Sistemi nella gestione delle acque*», in collaborazione con l'Istituto di Idraulica, Idrologia e Gestione delle Acque della Facoltà di Ingegneria dell'Università di Catania e con il FORMEZ (Roma, 20-22 novembre 1984), 1986.
78. XIII Seminario sulla: «*Evoluzione Biologica e i grandi problemi della Biologia. Evoluzione degli organuli cellulari*» (Roma, 26-28 febbraio 1986), 1987.
79. XIV Seminario sulla: «*Evoluzione Biologica e i grandi problemi della Biologia. Dalla biologia dello sviluppo alle biotecnologie*» (Roma, 25-27 febbraio 1987), 1989.
80. XV Seminario sulla: «*Evoluzione Biologica e i grandi problemi della Biologia. Le difese umorali e cellulari degli organismi*» (Roma, 24-26 febbraio 1988), 1990.
81. Tavola rotonda sul tema: «*Continui con memoria*» (Roma, 9 maggio 1986), 1990.
82. XVI Seminario sulla: «*Evoluzione Biologica e i grandi problemi della Biologia. La tettonica delle placche e la distribuzione dei viventi*» (Roma, 23-25 febbraio 1989), 1990.
83. XVII Seminario sulla: «*Evoluzione Biologica e i grandi problemi della Biologia. Origine ed evoluzione dell'uomo*» (Roma, 21-23 febbraio 1990), 1991.
84. *Conservazione del patrimonio culturale. Ricerche interdisciplinari, I.* 1992.
85. XVIII Seminario sulla: «*Evoluzione Biologica e i grandi problemi della Biologia. Sistematica ed evoluzione dei viventi*» (Roma, 26-28 febbraio 1991), 1992.
86. XIX Seminario sulla: «*Evoluzione Biologica e i grandi problemi della Biologia. Faune attuali e faune fossili*» (Roma, 26-28 febbraio 1992), 1993.
87. Seminario su: «*Discipline umanistiche e informatica. Il problema dell'integrazione*» (Roma, 8 ottobre 1991). A cura di Tito Orlandi. 1993.
88. *Conservazione del patrimonio culturale. Ricerche interdisciplinari, II.* 1994.
89. XX Seminario sulla: «*Evoluzione Biologica e i grandi problemi della Biologia. Dai Procarioti agli Eucarioti*» (Roma, 25-27 febbraio 1993), 1994.
90. XXI Seminario sulla: «*Evoluzione Biologica e i grandi problemi della Biologia. Evoluzione degli ecosistemi*» (Roma, 24-26 febbraio 1994), 1995.

91. *La diffusione in Italia delle metodologie scientifiche per lo studio e la conservazione delle opere d'arte*. 1994. Conservazione del patrimonio culturale. Ricerche interdisciplinari, III. 1995.
92. M. MARABELLI, *Conservazione e restauro dei metalli d'arte*. Conservazione del patrimonio culturale. Ricerche interdisciplinari, IV. 1995.
93. L. PERILLI, *Filologia computazionale*, 1995.
94. XXII Seminario sulla: «*Evoluzione Biologica e i grandi problemi della Biologia. Molecole ed evoluzione*» (Roma, 23-25 febbraio 1995), 1996.
95. XXIII Seminario sulla: «*Evoluzione Biologica e i grandi problemi della Biologia. Coevoluzione e coadattamento*» (Roma, 22-24 febbraio 1996), 1997.
96. *Il problema della formalizzazione*. Discipline umanistiche e informatica. Ciclo di Seminari (Roma, febbraio-giugno 1994). A cura di Tito Orlandi. 1997.
97. XXIV Seminario sulla: «*Evoluzione Biologica e i grandi problemi della Biologia. Le società animali*» (Roma, 26-28 febbraio 1997), 1998.
98. Simposio: «*Centenario della scoperta dell'elettrone*» (Roma, 13 novembre 1997), 1998.
99. XXV Seminario sulla: «*Evoluzione Biologica e i grandi problemi della Biologia. Ritmi biologici*» (Roma, 25-27 febbraio 1998), 1999.
100. *Il ruolo del modello nella scienza e nel sapere*. Convegno (Roma, 27-28 ottobre 1998), 1999.
101. XXVI Seminario sulla: «*Evoluzione Biologica e i grandi problemi della Biologia. Le biotecnologie*» (Roma, 25-27 febbraio 1999), 2000.
102. XXVII Seminario sulla: «*Evoluzione Biologica e i grandi problemi della Biologia. L'infertilità umana*» (Roma, 17-19 febbraio 2000), 2000.
103. Convegno: «*Valore sociale della ricerca scientifica: problemi italiani ed europei*» (Roma, 3 aprile 2000), 2001.
104. XXVIII Seminario sulla: «*Evoluzione Biologica e i grandi problemi della Biologia. Sessualità ed evoluzione*» (Roma, 21-23 febbraio 2001), 2002.
105. Convegno Internazionale: «*Archaeometry in Europe in the third millennium*» (Roma, 29-30 marzo 2001), 2002.
106. XXIX Seminario sulla: «*Evoluzione Biologica e i grandi problemi della Biologia. Fenomeni di auto-organizzazione nei sistemi biologici*» (Roma, 20-22 febbraio 2002), 2003.
107. Convegno Internazionale sul tema: «*I modelli nella ricerca archeologica. Il ruolo dell'informatica*» (Roma, 23-24 novembre 2000), 2003.
108. XXX Seminario sulla: «*Evoluzione Biologica e i grandi problemi della Biologia. Molecole e malattie*» (Roma, 20-22 febbraio 2003), 2004.
109. Convegno Internazionale: «*I culti primordiali della grecità alla luce delle scoperte di Tebe*» (Roma, 24-25 febbraio 2000), 2004.

110. Convegno Internazionale: «*Recenti tendenze nella ricostruzione della storia antica d'Israele*» (Roma, 6-7 marzo 2003), 2005.
111. Convegno Internazionale: «*L'economia palaziale e la nascita della moneta: dalla Mesopotamia all'Egeo*» (Roma, 12-13 febbraio 2002), 2005.
112. XXXI Seminario sulla: «*Evoluzione Biologica e i grandi problemi della Biologia. Genetica, Epigenetica ed Evoluzione*» (Roma, 26-28 febbraio 2004), 2005.
113. XXXII Seminario sulla: «*Evoluzione Biologica e i grandi problemi della Biologia. Neurobiologia della coscienza*» (Roma, 24-26 febbraio 2005), 2006.

**Per ordini d'acquisto rivolgersi a:**

**BARDI EDITORE S.R.L.**

**EDITRICE E COMMISSIONARIA**

**Via Piave, 7**

**00187 ROMA RM**

**Tel. +39 064817656 - Telefax: + 39 0648912574**

**e-mail: [info@bardieditore.com](mailto:info@bardieditore.com)**

**[www.bardieditore.com](http://www.bardieditore.com)**





