Dimensionamento e politiche di gestione di una biblioteca digitale

Autori:
ing. Carlo Jacob (mailto:jacob@posta.cilea.it), esperto di elaborazione di immagini digitali e databases, docente nel corso CILEA "Progettare il digitale in biblioteca"

Abstract (Italiano):

Il dimensionamento dei supporti di memorizzazione di una biblioteca digitale presenta delle notevoli difficoltà, dovute in larga parte all'incertezza nel definire le classi tipologiche dei materiali di un patrimonio bibliotecario, dalle quali dipendono, fra l'altro, le efficienze di comprensione dei vari formati file. La difficoltà nell'affrontare in maniera deterministicamente esatta il problema del dimensionamento impone il ricorso a metodi statistici e di calcolo delle probabilità (metodi Monte Carlo) che non si esauriscano nel calcolo di semplici medie, ma forniscano elementi sufficienti per condurre analisi "what-if" e permettano di valutare economicamente le politiche di realizzazione e mantenimento della biblioteca digitale. La valutazione dell'entità delle memorie di massa necessarie, inoltre, non esaurisce il compito del responsabile (manageriale e/o tecnico) del progetto, poiché deve essere necessariamente completata da quello delle gerarchie dei supporti nel management system, tenendo anche conto della vita media e dell'obsolescenza tecnologica dei supporti fisici, tutti aspetti strettamente correlati alle prestazioni che l'intero sistema deve erogare, soprattutto nel caso di fruizione in rete. Viene, inoltre, presentato un esempio di dimensionamento con un software *ad hoc* (DBD.EXE), usato dall'autore nel corso CILEA "Progettare il digitale in biblioteca".

Abstract(English):

Dimensioning mass memories for a digital library is a very challenging problem. That's for the uncertainty in defining the typological classes of the library "objects" and their amounts. In fact, the size of mass memories depends not only on amounts of volumes in each class, but also on the page sizes, the type of printing and the images represented in each volumes, which affect the efficiency of the various compression methods used in recording digital pages (both lossless and lossy). As a consequence, in dimensioning a digital library every deterministic (analytical) method is useless, at least in the majority of cases. It seems more appropriate using statistical and probability calculus methods, such as Monte Carlo methods. These methods allow to simulate the amount of memory needed by each volume in library, as a limit. Of course, statistics allow to simulate only "few" volumes, extending results to all the volumes in library. Above all, simulation can allow very interesting "what-if" analyses, varying the hypotheses about library structure and policies about its management and costs. Last but non least, evaluation of needed mass memories amounts must be completed by the design of the memories hierarchy and analysis of physical supports mean life expectation, together with their technological obsolescence. All items strictly correlated to system efficiency for online digital libraries. An example of digital library dimensioning is presented using an "ad hoc" software (DBD.EXE), used by the author during the CILEA course "Design of a Digital Library".

Keywords:

C.Jacob,E.Groppo,CILEA,Biblioteca Digitale,Progettare il digitale in biblioteca,Dimensionamento,Digitalizzazione,Supporti di memorizzazione,Memorie di massa,Simulazione,Metodi Monte Carlo,Digital Library,Dimensioning,Digitization,Mass Memories,Simulation,Monte Carlo Methods, Design of a Digital Library,DBD,DBD.EXE

1. Il progetto della digitalizzazione della biblioteca

L'ampiezza e la diversificazione della biblioteca media italiana rendono estremamente difficile la valutazione esatta dell'entità delle memorie di massa che dovrebbero conservare la sua versione digitale e quella conseguente dei costi dei supporti fisici. E questo perché la valutazione dipende da un enorme numero di parametri difficilmente quantificabili con esattezza, alcuni propri della composizione dei materiali bibliotecari, tipizzabili solo su grandi linee, e altri, invece, legati ad alcune caratteristiche imprevedibili del processo di digitalizzazione, come la resa delle compressioni dei dati numerici ottenuti, intese a risparmiare spazi di memorizzazione.

Un altro aspetto che complica ogni valutazione è quello relativo ai modi diversi con cui si può condurre il progetto complessivo della digitalizzazione, in altri termini la molteplicità delle impostazioni progettuali possibili, tanto che si può ben dire che la scelta progettuale è strettamente legata ad una ben precisa visione di politica culturale (o dei beni culturali). Data una biblioteca, non esiste un unico progetto di digitalizzazione, ma forse esiste, da parte dei responsabili, una visione politico-culturale molto più precisa di quanto a priori si possa credere. Quest'ultima interviene fortemente nelle scelte progettuali e, quasi sempre, in modo assolutamente critico e determinante, basti pensare alla scelta di digitalizzare tutto il materiale o solo una parte, oppure a quella di riservare la massima precisione digitale a certi materiali piuttosto che ad altri. Discorso a parte merita la scelta di procedere per gradi, "ingrandendo" via via la biblioteca digitale memorizzando nuovi dati. Ma questa scelta implica, da una parte, una pianificazione politica delle "priorità" di digitalizzazione da assegnare ai vari materiali e, dall'altra, una seria pianificazione del budget a breve, medio e lungo periodo, che, di nuovo, non può fare a meno di una visione delle dimensioni complessive dell'intera biblioteca digitale. Inoltre, un progetto minimamente lungimirante per quanto riguarda la sicurezza dei dati e la pianificazione degli investimenti non può non tener conto di un aspetto cruciale dell'architettura di una biblioteca digitale: la vita media e l'obsolescenza tecnologica dei supporti[1].

Insomma, è necessario convincersi che il progetto sarà influenzato inevitabilmente dagli assunti politici iniziali e dagli obiettivi che si intendono raggiungere, fra i quali uno dei più importanti, a mio avviso, è quello delle modalità di fruizione generalizzata in rete della biblioteca in quanto OPAC. E questo costringe la progettazione a misurarsi anche con i problemi del disegno dell'architettura del sito di accesso[2].

Nell'incertezza che caratterizza la conoscenza dei dati di una grande biblioteca e della sua versione digitale, quindi, non ha senso valutare deterministicamente *una tantum* le dimensioni dei supporti di memorizzazione. Non rimane che far ricorso, a parità di visione complessiva, ai metodi statistici (o stocastici), che conducono a valutazioni medie, minime e massime con probabilità che, di nuovo, dipendono in larga misura dall'impostazione politica adottata più o meno consapevolmente a priori e dagli obiettivi prefissati.

La prima fase del progetto consiste nel fissare i parametri da prendere in considerazione. E' una fase tutt'altro che semplice, perché, da una parte, richiede una non superficiale conoscenza della particolare biblioteca, ma, dall'altra, presuppone anche una conoscenza non trascurabile dei processi di digitalizzazione e memorizzazione che le moderne tecnologie informatiche mettono a disposizione. Ecco allora la necessità che tutto il progetto, dal disegno alla realizzazione, sia un tipico lavoro di *equipe*, dove si alternano competenze bibliotecarie e tecnico-scientifiche.

I parametri di valutazione non sono determinati una volta per tutte. Alcuni sono estremamente "naturali", come quelli che dividono in classi o tipologie il materiale e ne fissano le caratteristiche statistiche, ma altri richiedono uno sforzo di immaginazione che non può che essere fatto

[1] Cfr. 5 Archivi per decenni

[2] Cfr. [9] per una trattazione evoluta del progetto di siti con grammatiche e semantiche specializzate alla materia trattata. In particolare il problema dei thesauri e delle relazioni fra oggetti del sito

collegialmente in seno al gruppo di progetto poiché richiedono competenze diverse, come quelli relativi alle percentuali di pagine da digitalizzare per i testi o volumi delle diverse classi o quelli relativi alla precisione (risoluzione) da adottare per le diverse tipologie. Di seguito viene riportata un lista ragionevole dei parametri più importanti[3].

La seconda fase consiste nell'attribuire ai parametri scelti dei valori statistici, intervalli di variazione, medie etc. Anche questa non è una fase semplice e anche qui le competenze richieste sono molteplici.

Al termine di queste due fasi (le più importanti) si passa al dimensionamento vero e proprio della biblioteca digitale usando appositi programmi informatici.

Vale la pena notare come l'impostazione statistica del progetto richieda l'uso di *software* di dimensionamento fortemente orientato alla simulazione delle operazioni reali di digitalizzazione (metodi **Monte Carlo**)[4], così come determinate statisticamente dai valori dei parametri fissati. In altri termini, tali programmi informatici simulano i risultati che si potrebbero ottenere digitalizzando un certo numero di volumi della biblioteca, per poi estrapolare statisticamente tali risultati parziali a quelli totali relativi all'intera biblioteca. Si potrebbe obiettare che, una volta noti i valori statistici dei parametri di valutazione, alcuni risultati globali si possono calcolare facilmente in base a formule che, al più, richiederebbero una calcolatrice tascabile, senza ricorrere alla simulazione. Il problema è che, di solito, questo è vero per i calcoli dei risultati medi, ma le formule per arrivare alla valutazione delle dimensioni minime e massime accettabili sono spesso complicatissime e inaffrontabili se non a costo di sofisticazioni eccessive.

Infine, la "precisione" dei risultati finali non potrà che essere valutata sulla base di intervalli di probabilità più o meno esigui fissati a priori.

1.1. Metadati

I metadati associati al complesso dei dati digitali di un volume occupano a loro volta spazi di memorizzazioni non sempre irrilevanti. Ma oltre ai metadati di catalogazione, descrittivi, amministrativi e gestionali del volume[5], i moderni applicativi per il trattamento delle immagini digitali consentono di associare a ciascun *file*-immagine "metadati" che descrivono l'immagine stessa, come i parametri tecnici per la decodifica del *file*, indispensabili, la data di acquisizione e il suo autore, il tipo di scanner etc. Di questi metadati, sicuramente il più impegnativo dal punto di vista dello spazio di memorizzazione è il profilo di colore dell'immagine. L'uso del profilo di colore è raccomandato da **ICC** (**I**nternational **C**olor **C**onsortium)[5] poiché consente di convertire i colori dell'immagine a quelli del medium di rappresentazione (video, stampante a colori etc.), altrimenti l'immagine verrà rappresentata come se il suo modello di colore[6] coincidesse con quello del medium, il che di solito ha come conseguenza valori cromatici che appaiono più o meno differenti da quelli dell'originale. Un profilo di colore consiste di tabelle di conversione di entità a volte notevole, se prevede conversioni di colore tabellari, molto veloci. Naturalmente, anche questi spazi di memorizzazione dovrebbero essere considerati in un dimensionamento affidabile. Di solito se ne tiene conto semplicemente aumentando la dimensione calcolata di qualche percento.

---

[3] Cfr. 2 Una scelta dei parametri di valutazione

[4] I metodi Monte Carlo richiedono conoscenze di statistica e calcolo delle probabilità non indifferenti. Per pubblicazioni in italiano relative alla simulazione cfr. il classico [4]. Per il calcolo delle probabilità un buon testo è [10]

[5] Per i metadati MG-ICCU cfr. [6] [8]

[6] Approssimativamente, il modello di colore di un'immagine fa corrispondere la tripletta R(rosso), G(verde) e Blue(blu) di ciascun pixel, o C(ciano), M(magenta), Y(giallo) e K(nero) per i punti di stampa, a tre coordinate assolute di colore XYZ, così come definite dalle regole CIE 1931 (**C**ommission **I**nternationale de l'**E**clairage)[1], riferimento unico per tutti i dispositivi che trattano il "colore"

2. Una scelta dei parametri di valutazione

Come detto, la scelta dei parametri di valutazione non è immediata, e soprattutto non è univoca. Di seguito si riporta una scelta ragionevole e il più possibile "naturale" di parametri, fra quelli che influenzano in maniera più decisa i risultati.

2.1. La divisione del materiale bibliotecario in classi

La divisione in "classi" o "tipologie" del materiale bibliotecario è di estrema importanza e risulta critica per qualunque dimensionamento.. Oltre a costituire un concetto usuale per qualunque bibliotecario, le classi differenziano le caratteristiche dei volumi[7] rispetto alla loro capacità di dar luogo a dati digitali più o meno "ingombranti" per via del formato delle pagine, la presenza di immagini colorate o grigie, la preziosità dei caratteri, la necessità di dettagliare le pagine in immagini separate per la presenza di immagini di valore (miniature et alt.), il grado di conservazione del supporto, e così via.

Ciascuna classe sarà caratterizzata da un nome (**name**[8]) per facilità di lettura dei risultati e da una percentuale (**perc**) dei volumi di classe sul totale dei volumi in biblioteca.

Come detto, a tutti i parametri di valutazione deve essere assegnato un valore statistico ovvero un intervallo (<da> <a>) entro cui può essere scelto il valore reale. Questo deve avvenire anche per la percentuale di classe, solo che, ovviamente, ad ogni scelta di una combinazione di percentuali di classe tutti i valori estratti devono dare somma 100, valore che rappresenta "l'intera biblioteca" . Naturalmente è impossibile fissare intervalli di percentuale tali che ogni scelta di combinazioni di classi sia verificata la somma 100, a meno di casi banali.

Per esempio:

- classe 1: percentuale da 20% a 30% (la percentuale media è 25%)
- classe 2: percentuale da 40% a 60% (la percentuale media è 50%)
- classe 3: percentuale da 35% a 45% (la percentuale media è 40%)

scegliendo, poniamo, 25% per la prima classe, 52% per la seconda e 36% per la terza, la somma risulterebbe 113, valore privo di senso. In questi casi le percentuali estratte vengono normalizzate cioè forzate a dare somma 100 in base a una semplice formula. I valori normalizzati possono essere anche molto diversi dagli originari, ciononostante la normalizzazione conserva i rapporti fra le medie delle percentuali originarie. In altri termini, se la percentuale media della classe X è due volte la percentuale media della classe Y, allora lo sarà anche la media delle percentuali normalizzate. Per ritornare all'esempio precedente:

- scelta originale: 25%, 52%, 36%
- scelta normalizzata:
  classe 1 : 22.124%
  classe 2: 46.018%
  classe 3: 31.858%
  22.124%+46.018%+31.858%=100%

Per evitare vistose differenze fra valori dati e valori normalizzati è consigliabile determinare gli intervalli delle percentuali di classe fissando innanzi tutto delle medie a somma 100 per poi dare gli intervalli come (piccole) variazioni relative attorno a queste medie, le quali rappresentano il grado di **incertezza** nella determinazione di ciascuna percentuale di classe. Per esempio:

---

[7] Qui si prende in considerazione esclusivamente materiale su supporto cartaceo o immagini su un qualunque supporto. Non vengono presi in considerazione eventuali prodotti multimediali, come brani sonori o video, per i quali sarà necessario prevedere adeguati spazi di memorizzazione.

[8] I termini in grassetto rimandano al programma DBD (cfr. 3 Un software per il dimensionamento della biblioteca digitale (DBD))

- classe X: percentuale da 15% a 25% (media 20%, incertezza relativa $100 \cdot \frac{\pm 5\%}{20\%} = \pm 25\%$)
- classe Y: percentuale da 40% a 60% (media 50%, incertezza relativa $100 \cdot \frac{\pm 10\%}{50\%} = \pm 20\%$)
- classe Z: percentuale da 25% a 35% (media 30%, incertezza $100 \cdot \frac{\pm 5\%}{30\%} = \pm 16.666\%$)

Un caso banale si verifica per percentuali di classe costanti per tutta la simulazione:

- classe A: percentuale da 20% a 20% (media 20%, incertezza 0%)
- classe B: percentuale da 50% a 50% (media 50%, incertezza 0%)
- classe C: percentuale da 30% a 30% (media 30%, incertezza 0%)

in cui valori dati e normalizzati coincidono sempre.

Ovviamente, a parità delle caratteristiche delle classe, quanto è più elevata l'incertezza tanto più i risultati saranno dispersi, vale a dire che i loro intervalli di variazione saranno ampi.

2.2. I parametri di classe

Ciascuna classe non è solo caratterizzata da un nome e da una percentuale, ma anche da tutta una serie di parametri statistici che descrivono il comportamento di un suo volume ai fini della digitalizzazione. Il valore della maggior parte di essi consiste in un intervallo di variazione statistica (<da> <a>) La media dei due estremi di intervallo è la media del parametro relativo durante la simulazione.

2.2.1. La grandezza delle pagine dei volumi (size)

Non esiste una misura standard delle pagine di un volume, soprattutto nelle nostre biblioteche, nemmeno all'interno di una classe che si presuppone omogenea. Questo parametro fornisce l'intervallo di variazione statistica delle dimensioni (p.es. in centimetri) delle pagine dei volumi simulati. Esempi:

- da (18, 22) a (21,29.7): la larghezza delle pagine varia fra 18 e 21 cm e l'altezza fra 22 e 29.7 cm,
  con larghezza media pari a 19.5 cm e altezza media pari a 25.85 cm
- da A4 a A2: la larghezza delle pagine varia fra 21 e 42 cm e l'altezza fra 29.7 e 59.4 cm,
  con larghezza media pari a 31.5 cm e altezza media pari a 44.55 cm

2.2.2. Il numero delle pagine dei volumi (npages)

E' l'intervallo di variazione del numero delle pagine per volume di classe.

Esempio:

- da 100 a 500: ogni volume ha un numero di pagine compreso fra 100 e 500, mediamente 300

2.2.3. La percentuale delle pagine di un volume da digitalizzare (poi, pages of interest)

Poiché nella stragrande maggioranza dei casi sarebbe irrealistico digitalizzare ogni volume di classe per intero, questo parametro fissa i limiti di variazione della percentuale delle pagine da digitalizzare (pagine di interesse). Esempio:

- da 10% a 35%: le pagine da digitalizzare saranno al minimo 10% del totale delle pagine di un volume e al massimo 35%, mediamente 22.5%

Tenere presente che ogni pagina che si decide di digitalizzare produce un'immagine digitale, la quale può comportare anche una mole cospicua di dati da memorizzare.

2.2.4. La percentuale delle pagine di un volume da dettagliare (detperc)

Nella maggior parte dei casi non è necessario digitalizzare le pagine di un volume a risoluzioni geometriche elevate, a meno che non si tratti di un volume particolarmente prezioso da analizzare digitalmente in seguito con tecniche sofisticate (codici antichi, incunaboli, manoscritti etc.), soprattutto in certi dettagli, o di un volume che contenga anche illustrazioni (p.es. miniature) per le quali la risoluzione usata per la parte scritta risulta insufficiente. Da qui l'esigenza di dare una valutazione statistica della percentuale delle pagine di interesse che meritano uno o più dettagli a risoluzione maggiore. Questo parametro fornisce l'intervallo di variazione statistica di tale percentuale: Esempio:

- da 5% a 7%: le pagine di dettaglio saranno al minimo il 5% delle pagine di interesse e al massimo il 7%, mediamente il 6%
- se per un volume la percentuale delle pagine di interesse selezionata è, poniamo, il 40% e il valore selezionato per questo parametro è 5.5%, allora la percentuale delle pagine del volume da dettagliare risulterà pari a 2.2% (40% per 5.5% diviso 100)

2.2.5. La risoluzione radiometrica e quella geometrica della digitalizzazione (color,spi[9], detspi)

La scelta di questi parametri è molto critica. Valori modesti implicano modesti spazi di memorizzazione, ma comportano anche un'insufficiente grado di rilevazione delle informazioni visive contenute in un documento, cosa che può compromettere non solo una corretta lettura del documento digitalizzato, ma anche, e soprattutto, l'analisi numerica delle sue caratteristiche.

2.2.5.1. La risoluzione radiometrica, ovvero volumi a colori, grigi o bianco/nero (color)

Tramite questo parametro si decide la percentuale di volumi di una classe che saranno digitalizzati a colori RGB (profondità di colore maggiore o uguale a 24 *bit*), quella dei volumi che saranno digitalizzati a livelli di grigio GL (profondità di colore maggiore o uguale a 8 *bit*) e quella dei volumi in bianco e nero B/W ((profondità di colore uguale a 1 *bit*). Anche qui le tre percentuali saranno normalizzate a somma 100. Se, per esempio, la percentuale in RGB è 25%, quella GL 40% e quella B/W 35%, allora nella simulazione ogni volume di quella classe sarà digitalizzato in RGB, in GL o in B/W con probabilità rispettivamente pari a 25%, 40% e 35%.

2.2.5.2. Risoluzione geometrica (spi,detspi)

Come è noto, ogni digitalizzatore può campionare una scena (foto, pagina etc.) ad una risoluzione geometrica massima (campioni per pollice (spi, *samples per inch*) o campioni per centimetro (spcm, *samples per centimeter*)) che viene detta risoluzione ottica del digitalizzatore. Ma il relativo *driver* (programma che pilota il digitalizzatore) di solito può anche fornire altre risoluzioni:

- inferiori a quella ottica: compattando opportunamente le informazioni ottenute a risoluzione ottica
- superiori a quella ottica: per interpolazione dei dati a risoluzione ottica

Tralasciando le risoluzioni di cui al secondo punto, l'utente ha la possibilità di scegliere fra una vasta gamma di risoluzioni in dipendenza dal grado di precisione adatto agli obiettivi scelti.

La scelta della risoluzione geometrica si effettua tramite due parametri: la risoluzione geometrica delle pagine "normali" della classe e quella delle pagine di dettaglio.

Esempio:

- pagine normali da 100 a 300 spi: una pagina normale al minimo sarà campionata a 100 campioni per pollice e al massimo a 300, mediamente a 200 spi
- pagine di dettaglio da 300 a 600 spi: una pagina di dettaglio sarà campionata al minimo a

---

[9] Si preferisce usare l'acronimo spi (samples per inch, campioni per pollice) piuttosto che il generico ppi (pixels per inch, pixel per pollice), più adatto ad essere usato nei problemi di visualizzazione e di stampa. Per i problemi di stampa, l'acronimo dpi (dots per inch, punti per pollice) indica la risoluzione della stampante, anche se commercialmente è usato in alternativa a spi

300 e al massimo a 600 spi, mediamente a 450 spi

Tenere presente che, se si raddoppia la risoluzione, la mole di dati numerici ottenuti per pagina quadruplicherà.

2.2.5.3. Risoluzioni minime

La scelta della risoluzione geometrica e di quella radiometrica per la digitalizzazione di un testo non deve essere sottovalutata.Infatti, la quantità di informazioni raccolte da una digitalizzazione è molto superiore a quella strettamente necessaria per la pura rappresentazione (e fruizione) visiva dell'originale. Si può dire che tale massa di informazioni costituisce il sostituto virtuale dell'originale, almeno per quanto riguarda le sue proprietà ottiche. Questo sostituto virtuale di un testo non solo consente una fruizione visiva del testo stesso, come detto, ma permette di eseguire su di esso tutta una serie di analisi e studi difficilmente applicabili all'originale, dallo studio di particolari caratteri tipografici, alla ricostruzione di parti distrutte, all'analisi del supporto, al confronto di uno stesso testo da fonti diverse. La scelta di un'insufficiente risoluzione geometrica e/o radiometrica potrebbe compromettere i risultati di certe analisi, come quelle cromatiche e soprattutto quelle che hanno come obiettivo "estrazioni" e confronti di forme notevoli, come singoli caratteri o brani di manoscritti per studi grafologici.

Per quanto riguarda la risoluzione geometrica, supponendo che le dimensioni medie di un carattere di stampa siano di 1 millimetro quadrato, in base alla legge del campionamento di Shannon (o Shannon-Nyquist)[10], il passo di campionamento dX (e dY), reciproco della risoluzione orizzontale (verticale), dovrebbe risultare alquanto inferiore a mezzo millimetro, per esempio come minimo un sesto di millimetro, se si vuole distinguere chiaramente il carattere stesso:

$$dX = dY = \frac{0.1}{6} = \frac{0.1}{3 \times 2} \quad cm$$

con il che la risoluzione geometrica (frequenza di campionamento) **spcm** (campioni-*pixel* per cm) o **spi** (campioni-*pixel* per pollice) risulta:

$$spcm = 3 \times \frac{2}{0.1} = 60$$

$$spi = spcm \times 2.54 = 3 \times \frac{2}{0.1} \times 2.54 = 152.4 \cong 153$$

con 3 grado di precisione del campionamento dei caratteri. Con generici grado di precisione **Pr** e **d**imensione del **c**arattere in millimetri **dcmm**, si ha:

$$spcm = \mathbf{Pr} \cdot \frac{2}{0.1 \times dcmm}$$

$$spi = spcm \times 2.54 = \mathbf{Pr} \cdot \frac{2}{0.1 \times dcmm} \times 2.54$$

con numero totale di campioni per carattere **Ntsc** pari a:

$$Ntsc = (\mathbf{Pr} \times 2)^2$$

In realtà, considerando che in generale un carattere è un disegno a tratto, quelle che si vogliono distinguere chiaramente sono le linee che disegnano il singolo carattere, di spessore solitamente ben inferiore alla dimensione complessiva del carattere stesso. Le relazioni viste, quindi, dovrebbere comportare un grado di precisione **Pr** più o meno elevato in relazione alla natura del testo (manoscritto o stampa), al tipo di *font*, allo stile (grassetto e/o corsivo), alla granularità della

[10] Si fa riferimento al teorema di Shannon (o Shannon-Nyquist) il quale assicura che una funzione continua campionata può essere ricostruita esattamente solo se contiene armoniche di frequenza massima B molto inferiore alla metà della frequenza di campionamento S. Quanto più B si avvicina a S/2 tanto più nella ricostruzione si avranno distorsioni note come *aliasing*. Conseguenza approssimativa: il minimo dettaglio da rilevare deve contenere (molto) più di due campioni

carta etc. Con **Pr** pari a 6 e **dcmm** pari a 1, si ha:

$$spcm = 120$$

$$spi \cong 306$$

$$Ntsc = 144$$

E' evidente, però, che all'aumentare della precisione aumenta enormemente lo spazio di memorizzazione richiesto da una singola pagina di testo. Per un formato A4 a colori e a 300 campioni per pollice, lo spazio di memorizzazione risulta di circa 25 **MiB**[11]. Di conseguenza, è ragionevole ricorrere a queste risoluzioni solo per testi di particolare pregio, fermandosi, come standard, a 150 campioni per pollice nel caso di testi normali o addirittura a meno se il testo digitalizzato deve conservare solo caratteristiche di discreta leggibilità, ma non di analisi e studio.

Per la risoluzione radiometrica, usualmente la scelta è fra una digitalizzazione a colori (RGB di solito a 3 *byte* per campione-*pixel*, pari a 24 *bit* complessivi), a livelli di grigio (GL di solito a 1 *byte* per *pixel*, 8 *bit*) o bianco e nero (B/W 1 *bit* per *pixel*). Anche qui la scelta deve essere fatta accuratamente. Bisogna considerare che, in molti casi, la scelta GL (un terzo dello spazio di memorizzazione in RGB, a parità di risoluzione geometrica) non è la più corretta,anche se apparentemente la stampa presenta solo il "nero" dell'inchiostro e il "bianco" della carta. In realtà, la scelta GL conserverebbe solo le informazioni relative alla leggibilità del testo e alla forma dei caratteri, ma perderebbe tutte le informazioni circa il tipo di inchiostro usato, le tecnologie di stampa e la tipologia del supporto, tutte informazioni che meriterebbero di essere conservate, se la digitalizzazione dovesse costituire un vero e proprio sostituto dell'originale. Per esempio, sarebbe impensabile digitalizzare in GL il codice atlantico di Leonardo.

Ovviamente, la scelta RGB, GL o B/W risulta particolarmente critica nel caso di testi illustrati e, per di più, in questi casi anche la risoluzione geometrica può risultare critica se si tratta di illustrazioni con disegni a tratto, per cui una risoluzione di 150 spi sarebbe sufficiente per il testo ma non per le illustrazioni.

Infine, la scelta B/W (un ottavo dello spazio di memorizzazione in GL, a parità di risoluzione geometrica) deve essere riservata unicamente per testi di cui interessa solo una modesta leggibilità, interessando solo il contenuto del testo stesso ma non le sue proprietà tipografiche.

2.2.6. Il fattore di compressione della memorizzazione dei dati digitali (compr)

Ogni pagina digitalizzata darà luogo ad un'immagine digitale da memorizzare usando un qualunque formato *file* che preveda anche la compressione dei dati. Il fattore di compressione viene definito come percentuale:

$$100 \cdot \frac{< spazio\ di\ memorizzazione\ non\ compresso > - < \ldots compresso >}{< \ldots non\ compresso >}$$

oppure

$$100 \cdot \frac{< \ldots compresso >}{< \ldots non\ compresso >}$$

Nel primo caso il fattore di compressione sarà tanto più elevato quanto più sarà stata efficiente la compressione, mentre nel secondo si ha il caso opposto. E' impossibile conoscere in anticipo quale sarà il fattore di compressione ottenibile con un certo metodo su una certa immagine, poiché dipende fortemente dalla complessità della scena rappresentata.

Questo parametro definisce l'intervallo di variazione del fattore di compressione per una pagina normale di un volume di classe.

Esempio, con fattore di compressione del secondo tipo:

- da 50% a 70%: lo spazio di memorizzazione di ogni pagina normale sarà al minimo il 50%

[11] Viene usata la notazione standad IEEE1541: K (kilo=1000), M (mega=1000 K), G(giga=1000 M), T (tera=1000 G), P (peta=1000 T)....., Ki (kibi=kilo binario=1024), Mi (mebi=mega binario=1024 Ki), Gi (gibi=giga binario=1024 Mi), Ti (tebi=tera binario=1024 Gi) etc., con suffissi b (bit), B (byte), p (pixel), v (volume), g (pagina) etc.

dell'originale da digitalizzatore e al massimo il 70%, mediamente il 60%

2.2.7. Il fattore di compressione delle pagine di dettaglio (detcompr)

Questo parametro definisce l'intervallo di variazione del fattore di compressione per una pagina di dettaglio di un volume di classe. Esempio, con fattore di compressione del secondo tipo:

- da 80% a 90%: lo spazio di memorizzazione di ogni pagina di dettaglio sarà al minimo l'80% dell'originale da digitalizzatore e al massimo il 90%, mediamente l'85%

2.2.8. Digitalizzazioni multiple (ExtraScan)

Di un volume dovrebbero essere fatte per lo meno due digitalizzazioni. La prima,e principale, ad alta definizione e memorizzazione con compressione senza perdita di informazioni (*lossless*: tipicamente formato TIF), per copie digitali *master* che possano essere ritenute "sostituti dell'originale" e la seconda, a risoluzione medio bassa e grande compressione con perdita di informazioni (*lossy*: tipicamante formato JPEG), per la presentazione in rete del patrimonio bibliotecario. Ovviamente anche il complesso di queste ultime digitalizzazioni occupa spazio sui supporti di memorizzazione. Il parametro **ExtraScan** consente di definire classi parallele a quelle principali, dove è possibile definire nuovi parametri di classe.

2.2.9. Parametro critico

Non tutti i parametri di classe sono responsabili in egual misura della variabilità dei risultati statistici. Il parametro che determina, a parità degli altri, la massima variabilità delle dimensioni di una classe viene detto parametro critico della classe. La determinazione intuitiva del parametro critico non è semplice poiché dipende non solo dalle singole variabilità, ma anche da come ciascun parametro interviene nel calcolo delle dimensioni della libreria digitale. In generale, candidati più probabili sono le dimensioni delle pagine e le risoluzioni di campionamento, seguiti dal numero delle pagine da digitalizzare.

2.3. Dati generici principali per il controllo dell'esecuzione

Oltre ai parametri di classe, la simulazione è guidata da una serie di dati che ne controllano l'esecuzione.

2.3.1. titolo

E' il titolo dell'esecuzione corrente.

2.3.2. volumi

E' il numero totale di volumi della biblioteca. Il dato può anche essere approssimato, poiché, con una semplice proporzione, i risultati possono essere rapportati a qualunque numero di volumi.

2.3.3. combinazioni

Numero di combinazioni di percentuali di classe (**perc**) da simulare. Il valore di questo parametro deve essere il più grande possibile compatibilmente con i tempi di calcolo, soprattutto in presenza di grandi incertezze sulle percentuali di classe o di grandi squilibri fra le stesse incertezze.

2.3.4. prove

Numero di volumi simulati all'interno di ciascuna combinazione. La simulazione completa per arrivare ai risultati finali comporterà un numero di volumi simulati pari al prodotto di **combinazioni** per **prove**, prodotto a cui sarà proporzionale anche il tempo totale di calcolo. Il valore di **prove** deve essere una frazione consistente di **volumi**. Per **prove=volumi**, ovviamente, verranno simulati, ad ogni combinazione di percentuali, tutti i volumi della biblioteca.

2.3.5. precisione

Precisione percentuale della simulazione. Quando i risultati delle successive combinazioni di classe

variano meno di questa percentuale, la simulazione si conclude.

2.3.6. sign

Percentuale dei risultati estremi da trascurare. Di solito i risultati estremi, inferiori o superiori, sono poco significativi. Il limite inferiore (superiore) dei risultati totali viene calcolato in modo tale che il complesso dei risultati inferiori (superiori) sia, in percentuale, pari a <significatività>.

2.4. La sequenza di calcolo

La simulazione, iterativa, consiste in un numero di cicli esterni pari a **combinazioni**, ciascuno composto da un ulteriore numero di cicli interni pari a **prove**.

Per ciascun ciclo esterno viene selezionata una combinazione di percentuali di classe (**perc**) normalizzate. Esempio con tre classi al generico ciclo esterno:

1. da perc=(50 60), perc=(5 15), perc=(30 40) vengono estratti tre valori 51, 10, 35
2. i tre valori vengono normalizzati: 53.125, 10.417, 36.458

Di seguito, per ciascun ciclo interno viene estratta una classe con probabilità derivata dalle percentuali di classe normalizzate e quindi viene simulato un volume appartenente a quella classe in base alle caratteristiche statistiche dei relativi parametri (da **color** a **detcompr**). Per ciascuno dei volumi simulati vengono accumulate le statistiche per la relativa classe. Il calcolo viene ripetuto per tutte le combinazioni, di numero pari a **combinazioni**. Le statistiche di ogni combinazione costituiscono un'ipotesi di biblioteca digitale. Si ha un numero di ipotesi di biblioteca digitale pari a **combinazioni**. Dall'insieme delle combinazioni vengono calcolate le medie, le deviazioni standard, i minimi e i massimi di ciascuna caratteristica della biblioteca digitale, come dimensioni della memorizzazione, pagine, volumi etc. Ogni risultato, infine, viene rapportato al numero totale di volumi della biblioteca (**volumi**).

3. Un software per il dimensionamento della biblioteca digitale (DBD)

Nell'ambito dei corsi CILEA **"Progettare il digitale in biblioteca"**, è stato messo a punto un software C++ per il dimensionamento della biblioteca digitale (**D**imensionamento **B**iblioteca **D**igitale). **DBD** segue i criteri Monte Carlo esposti nei paragrafi precedenti. Da un punto di vista architetturale, **DBD** accetta dati da un file testo di input e/o interattivamente, e registra i risultati su un file testo di output. Ambedue i file sono visibili nella console grafica di **DBD**. L'architettura "file di input/file di output" consente di conservare come documento sia il complesso dei dati che quello dei risultati. I dati sono forniti tramite un semplice linguaggio di interfaccia. L'uso del linguaggio di interfaccia facilita la parametrizzazione dei dati numerici tramite l'uso di espressioni algebriche che permettono di definire il valore di un dato basandosi sui valori di dati precedentemente definiti.

Un esempio di scrittura dei dati sul file di input è riportato in Figura 1. Il testo consiste semplicemente in una serie di definizioni di parametri del tipo:

<nome>=<valore>

senza alcuna preoccupazione di come e quando i dati verranno letti. Salvo i casi in cui non avrebbe senso (p.es.**Classe**), tutte le definizioni sono di tipo statistico e, quindi, in genere definiscono intervalli di variazione[12] (<da> <a>) piuttosto che singoli valori, anche se singoli valori sono tollerati per definire intervalli di variazione nulli e quindi la costanza del parametro corrispondente per tutta l'elaborazione(vedi size=A4).

La semplice distribuzione uniforme dei parametri del tipo (<da> <a>) a volte può non essere sufficiente. Per esempio, con questa distribuzione non è possibile esprimere la circostanza che il

[12] Un intervallo di variazione definisce una variabile aleatoria a distribuzione uniforme fra due limiti.

numero delle pagine è per il 50% pari a 200, il 10% pari a 100 e il 40% pari a 600. In DBD sono possibili altre distribuzioni più sofisticate, come la uniforme (**UNIF**), simile a quella (<da> <a>), ma con qualche differenza, la uniforme a numeri interi (**IUNIF**), la gaussiana (**NORM**), la continua a tratti (**CDIST**) e la discreta (**DDIST**).

Il parametro **size**, inoltre, è molto sensibile al tipo di distribuzione. Come per i parametri generici, anche in questo caso sono possibili distribuzioni più sofisticate della semplice uniforme del tipo, per esempio, (A4 A3). Al contrario dei parametri generici, però, qui le distribuzioni viste introducono tutte delle forti correlazioni fra larghezza e altezza di ogni formato estratto.

Per esempio, assumendo <formato> come:

<formato standard>= A4, A3, A2, A1, A0, A, B, C, D, E
<formato libero>=(cm <larghezza> <altezza>)
<formato libero>=(inch <larghezza> <altezza>)

la distribuzione **UNIF**(<da formato> <a formato>): non coincide con la semplice distribuzione uniforme del tipo (<da formato> <a formato>). Il formato della pagina varia uniformemente fra <da formato> e <a formato>, ma anche il rapporto fra altezza e larghezza (*aspect ratio*) varia uniformemente fra quello di <da formato> e quello di <a formato>, il che introduce una correlazione completa fra le altezze e le larghezze di tutti i formati estratti, al contrario della semplice distribuzione (<da formato> <a formato>) in cui non c'è nessuna correlazione fra altezze e larghezze estratte separatamente.

Esempi:

- UNIF(A4 A3): variazione uniforme fra il formato standard A4 e il formato standard A3. E' equivalente a UNIF(A3 A4)
- UNIF((cm 20 20) (cm 40 40)): tutti i formati estratti hanno lo stesso aspect ratio pari a 40/20=2

Infine, per poter definire i valori di un parametro in base a quelli di un altro parametro precedentemente definito in ogni punto della sequenza dei dati, qualunque dato del tipo:

@<nome>=<valore>

definisce una variabile numerica di sistema di nome <nome> e valore <valore>. Il valore di questa variabile può essere usato in una qualunque espressione numerica seguente. Se <nome> è il nome di un parametro da leggere, la sua lettura avverrà normalmente dopo la definizione della variabile di sistema e avrà lo stesso valore.

Esempi:

- @**volumi**=100000 @**prove**=(volumi/4) **combinazioni**=(prove/10)
  equivale a
  volumi=100000 **prove**=25000 **combinazioni**=2500
  e alla definizione delle variabili di sistema volumi e prove
- @**dp**=.05 @**media**=500 **npages**=((**media***(1-dp)) (**media***(1+**dp**)))
  equivale a
  npages=(475 525) e alla definizione delle variabili di sistema **dp** e **media**

Il secondo esempio illustra un metodo molto comodo per assegnare una variazione statistica definita come una media più o meno una certa frazione.

4. Un esempio di dimensionamento di biblioteca digitale

4.1. I dati di input

Si suppone che la biblioteca contenga circa 100000 (centomila) volumi, divisi in quattro classi con extrascan:

1. Volumi normali, praticamente solo testo: nome “Testo normale A4”
2. Volumi con testo e illustrazioni a colori: nome “Testo+illustrazioni”
3. Volumi con prevalenza di illustrazioni a colori: nome “Illustrazioni”
4. Carte geografiche: nome “Carte geografiche”

di percentuali medie rispettivamente 50%, 30%, 8% e 2%. La scelta dei valori dei parametri di classe è riportata in Figura 1.

Per quanto riguarda la risoluzione radiometrica (RGB, GL o BW) si suppone che la prima classe possa essere digitalizzata in GL o BW, la seconda in RGB e GL, la terza e quarta esclusivamente in RGB.

La risoluzione geometrica per le pagine normali varia fra 150 e 300 campioni per pollice e quella delle pagine di dettaglio fra 300 e 600 campioni per pollice

Il parametro generale **volumi** è ovviamente 100000, **prove** 33000 mentre **combinazioni** è sufficiente che sia superiore a 2000, dal momento che per valori superiori a 2000 i risultati statistici non presentano variazioni significative. La significatività dei limiti, **sign**, viene fissata al 2%.

```
titolo="Esempio a 4 classi con Extrascan"
@volumi=100000 prove=(volumi/3)  combinazioni=100000 sign=2
Classe=(name=(TIF:Testo normale A4)  size=A4  perc=(55 65)
 npages=(100 500) poi=(10 20)
 color=(RGB=0 GL=80 BW=20)  compr=(50 60) detperc=(5 10)
 detcompr=(80 90) spi=(150 150)  detspi=(300 300)
 ExtraScan=(name=(JPG:Testo normale A4)  compr=(10 20)
            detcompr=(20 35) detperc=() spi=() detspi=() color=()
 )
)
Classe=(name=(TIF:Testo+illustrazioni A4)  size=A4  perc=(25 35)
 npages=(200 400) poi=(20 50)
 color=(RGB=80 GL=20 BW=0)  compr=(60 80) detperc=(10 20)
 detcompr=(80 90) spi=(150 150)  detspi=(300 300)
 ExtraScan=(name=(JPG:Testo+illustrazioni A4)  compr=(10 30)
            detcompr=(25 35) detperc=() spi=() detspi=() color=()
 )
)
Classe=(name=(TIF:Illustrazioni) size=((cm 30 40) (cm 50 60))
 perc=(5 11) npages=(100 200) poi=(60 70)
 color=(RGB=100 GL=0 BW=0)  compr=(80 90) detperc=(30 40)
 detcompr=(80 90) spi=(300 300)  detspi=(600 600)
 ExtraScan=(name=(JPG:Illustrazioni)  compr=(20 40)
            detcompr=(30 40) detperc=() spi=() detspi=() color=()
 )
)
Classe=(name=(TIF:Carte geografiche da A3 a A0)  size=(A0 A3)
 perc=(1 3) npages=(1 1) poi=(100 100)
 color=(RGB=100 GL=0 BW=0)  compr=(50 70) detperc=(45 55)
 detcompr=(80 90)  spi=(300 300)  detspi=(600 600)
 ExtraScan=(name=(JPG:Carte geografiche da A3 a A0) compr=(20 30)
            detcompr=(30 40)  detperc=() spi=() detspi=() color=()
 )
)
trfile=esempio.trc
```

Figura 1. I blocchi dei parametri di classe sul file di input

4.2. I risultati

4.2.1. Le tabelle e le figure riassuntive

Le percentuali di classe date, teoriche e simulate sono riportate in Tabella 1.

La dimensione totale media della biblioteca digitale(Tabella 4) risulta di circa 199 TB, in accordo con la previsione teorica.

I limiti al 2% (Tabella 5), rispettivamente circa 141 e 260 TB, risultano significativamente superiori e inferiori ai limiti minimo e massimo simulati, rispettivamente circa 131 e 280 TB.

Il maggior contributo alla dimensione totale media lo fornisce la classe "Illustrazioni" con il suo extrascan, circa 121 TB, pari a circa il 61% del totale, sebbene il suo numero medio di volumi sia l'8% del totale(Tabella 4).

L'istogramma delle dimensioni per le varie ipotesi di biblioteca digitale (Figura 2) non presenta un andamento a campana (distribuzione normale o gaussiana), ma sembra distribuito uniformemente, per lo meno nella sua parte centrale. Questo è dovuto alla elevata incertezza relativa sulle percentuali di classe, soprattutto di quelle più cospicue, e sulla elevata variabilità dei parametri critici delle varie classi:

1. Testo normale: incertezza relativa ±8.3%, parametro critico **npages**=(100 500)
2. Testo+illustrazioni: ±16.666%, **npages**=(200 400)
3. Illustrazioni: ±37.5% ,**size**=((cm 30 40) (cm 50 60))
4. Carte geografiche: ±50%, ,**size**=(A0 A3)

Le dispersioni relative (STD/media) delle dimensioni(Tabella 4) risultano:

1. Testo normale: circa 4%
   extrascan: circa 4%
2. Testo+illustrazioni: circa 8%
   extrascan: circa 8%
3. Illustrazioni: circa 21%
   extrascan: circa 21%
4. Carte geografiche: circa 29%
   extrascan: circa 29%
5. DIMENSIONE TOTALE: circa 17%

PERC. O/T/M--->	MEDIA	disp.rel.%	MINIMA	MASSIMA	inc.rel.%
TIF:Testo norma	60,0000	4,8113	55,0000	65	8,3333
TEORICA	60,0444	4,6172	52,8846	67,7083	12,3440
MONTECARLO	60,0173	3,8933	53,7489	66,5907	10,6984
TIF:Testo+illus	30	9,6225	25	35	16,6667
TEORICA	29,9732	8,8356	24,0385	36,4583	20,7183
MONTECARLO	30,0082	7,4984	24,3643	35,8141	19,0779
TIF:Illustrazio	8	21,6506	5	11	37,5000
TEORICA	7,9824	23,6577	4,6296	11,9565	45,8942
MONTECARLO	7,9734	20,5288	4,7312	11,6524	43,4023
TIF:Carte geogr	2	28,8675	1	3	50,0000
TEORICA	2,0001	37,0489	0,8929	3,4091	62,9036
MONTECARLO	2,0012	28,8269	0,9635	3,2730	57,7042

disp.rel.%: dispersione relativa=STD/MEDIA%
inc.rel.% : incertezza relativa =(MAX-MIN)/(2*MEDIA)%

Tabella 1. Percentuali di classe originali, teoriche normalizzate e simulate (Monte Carlo) normalizzate (virgola decimale)

VOLUMI/OGGETTI>	MEDI(*)	disp.rel.%	
TOT TEORICI	100,000K	0	
TOT MONTECARLO	100,000K	3,698	
TIF:Testo norma	60,017K	3,916	
JPG:Testo norma	60,017K	3,916	ES
TIF:Testo+illus	30,005K	7,529	
JPG:Testo+illus	30,005K	7,529	ES
TIF:Illustrazio	7,978K	20,629	
JPG:Illustrazio	7,978K	20,629	ES
TIF:Carte geogr	2,000K	29,086	
JPG:Carte geogr	2,000K	29,086	ES

(*)K=1000(kilo) ….[13]
virgola (,) decimale
disp.rel.%: dispersione relativa STD/MEDIA%

Tabella 2. Risultati statistici simulati del numero dei volumi

PAGINE(*)----->	MEDIE	disp.rel.%	xOGGETTO	MINIME	MASSIME
TOT TEORICHE	7,578M	21,485	75,776	2,221M	16,282M
TOT MONTECARLO	7,579M	2,444	75,789	7,050M	8,136M
TIF:Testo norma	2,903M	3,921	48,369	2,593M	3,216M
JPG:Testo norma	2,903M	3,921	48,369	2,593M	3,216M ES
TIF:Testo+illus	3,623M	7,537	120,746	2,953M	4,310M
JPG:Testo+illus	3,623M	7,537	120,746	2,953M	4,310M ES
TIF:Illustrazio	1,050M	20,640	131,629	604,215K	1,563M
JPG:Illustrazio	1,050M	20,640	131,629	604,215K	1,563M ES
TIF:Carte geogr	3,000K	29,086	1,500	1,282K	4,968K
JPG:Carte geogr	3,000K	29,086	1,500	1,282K	4,968K ES

(*)K=1000(kilo)….[14]
virgola (,) decimale
disp.rel.%: dispersione relativa STD/MEDIA%

Tabella 3. Risultati statistici simulati del numero delle pagine digitalizzate

[13] Vedi nota 11

[14] Vedi nota 11

DIMENSIONE(*)->	MEDIA	disp.rel.%	xOGGETTO	MINIMA	MASSIMA
TOT TEORICA	199,114T	23,219	2,039G	25,043T	508,756T
TOT MONTECARLO	199,058T	17,061	2,037G	130,560T	279,891T
TIF:Testo norma	3,548T	3,930	61,986M	3,155T	3,947T
JPG:Testo norma	1,025T	3,929	17,903M	932,961G	1,140T ES
TIF:Testo+illus	19,608T	7,539	685,237M	15,975T	23,298T
JPG:Testo+illus	6,158T	7,539	215,197M	5,016T	7,319T ES
TIF:Illustrazio	120,805T	20,648	15,505G	69,859T	180,288T
JPG:Illustrazio	46,783T	20,648	6,004G	27,055T	69,833T ES
TIF:Carte geogr	820,643G	29,164	420,173M	322,865G	1,326T
JPG:Carte geogr	338,974G	29,159	173,559M	133,258G	561,138G ES

(*)K=1024 bytes(KiB),M=1024 KiB(MiB)[15]
virgola (,) decimale
disp.rel.%: dispersione relativa STD/MEDIA%

Tabella 4. Risultati statistici simulati della dimensione della biblioteca digitale

LIMITI-------->	INFERIORE	SUPERIORE
DIMENSIONI(1)	141,383T	260,027T
PAGINE(2)	7,209M	7,958M

(1)K=1024 bytes(KiB)…..[16]
virgola (,) decimale
(2)K=1000(kilo)….[17]
virgola (,) decimale

Tabella 5. Limiti delle dimensioni e del numero delle pagine della biblioteca digitale al 2%

[15] Vedi nota 11

[16] Vedi nota 11

[17] Vedi nota 11

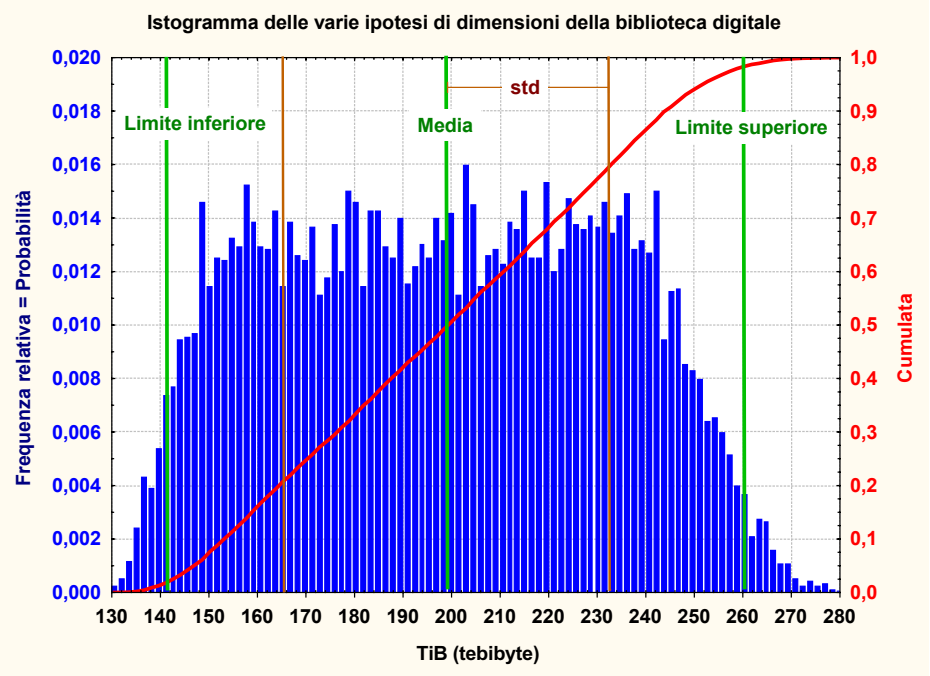

Figura 2. Istogramma delle varie ipotesi simulate delle dimensioni della biblioteca digitale

5. Archivi per decenni

La biblioteca digitale non deve necessariamente contenere soltanto immagini di testi, ma è inevitabile che contenga anche dati multimediali in senso lato (immagini, suoni, testi, filmati, oggetti complessi etc.), che possano offrire ulteriori informazioni, sotto forme più sofisticate, sulla biblioteca stessa e sui suoi materiali, soprattutto in vista della sua fruizione in rete.. Sarebbe più appropriato, pertanto, parlare dell'archivio digitale di una biblioteca come di un **archivio (o DB) multimediale**. Del resto, i più evoluti sistemi di catalogazione, come il **MAG ICCU**, già prevedono appositi *tag* per oggetti complessi.

La memorizzazione di grandi quantità di dati multimediali pone delle vere e proprie sfide non solo per quanto riguarda gli spazi di memorizzazione, ma anche, e soprattutto, per quanto attiene alla loro organizzazione razionale in *database* (DB) di massa, dove i problemi di gestione, accesso e fruizione sono enormemente condizionati da costi e velocità di movimentazione.

A tal fine, con l'avanzamento tecnologico, la gerarchia logica dei dati complessi si e' tradotta nella gerarchia dei supporti di memoria.

La moderna tecnologia delle memorie di massa ha stabilito un rapporto direttamente proporzionale fra costo di memorizzazione per **MiB** e velocità di accesso, quindi fra questi due parametri e frequenza di movimentazione.

I prodotti disponibili sul mercato possono essere ordinati secondo una scala ideale che va da bassi costi, basse velocità di accesso e usabili per basse frequenze di movimentazione (**sistemi *near-line* di archiviazione a lungo termine, tipicamente librerie di nastri LTO**), a prestazioni via via crescenti relativamente a questi parametri (**tipicamente librerie di WORM CD, DVD o MO magneto ottici, *e* librerie *on-line* , tipicamente *Hard Disk*, soprattutto in architettura RAID**).

Al di là dell'organizzazione logica di un grosso *database* multimediale[18], per risolvere i problemi di gestione, accesso e fruizione lo stato dell'arte attribuisce grande importanza alla gestione efficiente dei supporti di memoria. In un DB multimediale complesso è molto probabile che siano presenti vari tipi di supporti, anzi una loro diversificazione e specializzazione a seconda delle destinazioni d'uso all'interno di un sistema è auspicabile, non solo per ragioni di costo, ma anche di efficienza. Gli esempi sono numerosi. L'archivio di un ospedale[19], di una grande biblioteca, il DB di un museo o semplicemente l'archivio delle immagini di una casa editrice, soprattutto se in rete, devono essere progettati in modo tale da ridurre i costi, consentire ricerche logiche agevoli e permettere accessi veloci e affidabili. In tali contesti, è inevitabile che il DB debba prevedere delle "fasi di lavorazione" delle informazioni che tengano conto del loro progressivo "invecchiamento" in relazione alla gestione a brevissimo, breve, medio e lungo termine. Per esempio, in un archivio fotografico sarebbe inutile e costoso continuare a tenere su supporti veloci ma costosi immagini molto particolari che vengono richieste solo in casi eccezionali (bassa frequenza di movimentazione), sfavorendo altre tipologie di uso più frequente[20].

A parità di progettazione logica, l'efficienza di un DB multimediale, quindi, deve prevedere un'organizzazione e una gestione gerarchiche dei supporti di memoria (**HSM**, *Hierarchical Storage Management*)[21]. Poiché i costi sono proporzionali alle velocità di accesso permesse dai vari supporti, appare logico stabilire la gerarchia in base a queste ultime.

Ma nei grandi archivi digitali, l'affidabilità dei supporti di memorizzazione di qualunque tipo è ovviamente un fattore estremamente critico. Bisogna tener conto di due aspetti dei supporti:

- **vita media**
- **obsolescenza tecnologica**

Per quanto riguarda la **vita media** non ci sono eccessivi problemi, poiché per tutti i supporti si misura ormai in molte decine di anni (se non centinaia), durate che sono molto inferiori ad un'altra durata: quella della loro vita media "logica" o, in altri termini, quella dell' **obsolescenza tecnologica dei sistemi di scrittura e lettura**.

Con i ritmi attuali dell'innovazione tecnologica (6 mesi) la durata della vita media di qualunque supporto e' molto superiore all'intervallo di tempo in cui ogni supporto sarà sostituito fisicamente e logicamente da un altro più evoluto.

La conservazione dei "vecchi" metodi di memorizzazione per i "vecchi" supporti non garantisce minimamente la sopravvivenza dell'archivio. Infatti, se una certa unità A di memorizzazione, in grado di registrare e leggere il supporto B, ma non più prodotta dall'industria, si guasta, nessuno garantisce che il mercato sia in grado di offrire un unità diversa da A, ma pur sempre in grado di leggere il supporto B.

Questa criticità all'innovazione tecnologica[22] è proporzionale alla complessità dell'organizzazione dei dati sul supporto: fra qualche anno sarà molto più difficile trovare lettori in grado di leggere un disco Magneto Ottico o un CD-ROM piuttosto che un nastro, registrato sequenzialmente con un codice molto più semplice. Ecco che allora si pone il problema di progettare soprattutto l'archivio

---

[18] I tre più importanti tipi di organizzazione logica dei dati in un *database* sono il relazionale (RDBMS, *Relational Data Base Management Systems*), quello orientato agli oggetti (OODBMS, *Object Oriented Data Base Management Systems*) e quello misto (ORDBMS, *Object Relational Data Base Management Systems*). La maggior parte dei prodotti disponibili sul mercato per la creazione e la gestione di un *database* sono ORDBMS

[19] L'archivio ospedaliero è un classico esempio di DB multimediale, in quanto non deve contenere solo testi, ma anche immagini (RX, TC, NMR etc.), diagrammi, sequenze sonore etc.

[20] E' ovvio che l'archivio fotografico di un quotidiano debba privilegiare le immagini di cronaca piuttosto, mettiamo, che quelle di carattere scientifico, mentre il rapporto può essere invertito per una rivista mensile con finalità culturali

[21] cfr. [7]

[22] Per una delle più lucide analisi del problema, cfr. [3]. Per una trattazione divulgativa cfr. [11]. Per i rapporti fra tecnologie informatiche e beni culturali cfr. [2]

delle acquisizioni digitali originali delle immagini tenendo conto di questi nuovi elementi, i quali portano in primo piano una strategia di mantenimento finora sottovalutata: il **rinfrescamento dei dati con il loro riversamento periodico su supporti a nuove tecnologie**.

Per ovviare in primo luogo all'obsolescenza tecnologica, piuttosto che alla limitazione della vita media, un buon gestore di un archivio digitale deve prevedere un **piano periodico di riversamento delle informazioni** non solo su supporti più freschi (nuovi nastri, nuovi WORM e MO, nuovi HD), ma anche su supporti a nuova tecnologia, in modo tale da porsi al riparo da eventuali mutamenti di rotta del mercato, e questo anche se le "vecchie" macchine sono pienamente efficienti (o sembrano esserlo). Tutto ciò, naturalmente, comporta problemi economici che debbono essere attentamente valutati, problemi che debbono essere tenuti presenti anche nella scelta iniziale dei supporti di memorizzazione, dal momento che la memorizzazione su nastro magnetico è molto più a buon mercato (sia in termini di ammortamento che in quelli dei costi unitari per unità di supporto) di quella su WORM,MO e HD.

Ma i termini economici non esauriscono il problema del mantenimento di un archivio digitale. Forse di gran lunga più critico è l' **aspetto politico-decisionale**, per cui i curatori si dovranno assumere la responsabilità delle priorità con cui le varie sezioni di una collezione dovranno di volta in volta essere rinfrescate, a parità di *budget*.

Bibliografia e Riferimenti

[1] Commission Internationale de l'Eclairage,International Commission on Illumination, URL: http://www.cie.co.at/

[2] P.Galluzzi ed.,P.A.Valentino ed.,*I formati della memoria,*Giunti,Firenze,1997,ISBN 88-09-21190-1

[3] T.Gregory ed.,M.Moretti ed.,*L'eclisse delle memorie*, Editori Laterza,Bari,1994,ISBN 88-420-4540-3

[4] G.Iazeolla,*Introduzine alla simulazione discreta,*Boringhieri,Torino,1978,7-8761-4

[5] International Color Consortium,URL: http://www.color.org/

[6] Istituto Centrale per il Catalogo Unico, URL: http://opac.sbn.it/

[7] S.Khoshafian,A.B.Baker,*Multimedia and Imaging Databases,*Morgan Kaufmann Publishers Inc.,San Francisco,1996,ISBN 1-55860-312-3

[8] MAG: Metadati Amministrativi Gestioneli, URL: http://193.206.221.20/comimag.htm

[9] L.Rosenfeld,P.Morville,*Architettura dell'informazione per il World Wide Web*,II ed.,O'Reilly-Hops,Milano,2002,ISBN 88-8378-062-0

[10]S.M.Ross,*Calcolo delle probabilità,*Apogeo,VI ed.,Apogeo,2004,ISBN 88-503-2231-3

[11],Rothenberg,"Ensuring the Longevity of Digital Documents",*Scientific American*,Gennaio 1995