Una proposta di *library automation*

Il progetto di biblioteca digitale dell'area biofisica del CNR di Genova

di Roberta Maggi e Roberto Di Cintio

Sin dai tempi di Tolomeo Sotere, 200 anni prima di Cristo, l'uomo ha cercato di raggruppare le fonti della conoscenza in archivi cartacei di dimensioni sempre crescenti, dovendo affrontare, conseguentemente, non pochi problemi per la loro organizzazione, conservazione, catalogazione e consultazione. Ancora oggi, dopo oltre venti secoli, nonostante l'ausilio non indifferente della tecnologia elettronica, il concetto di biblioteca universale è ben lungi dall'essere praticabile, sia per motivi squisitamente economici, sia per problemi di natura pratica, sia, infine, per i non trascurabili aspetti legali che la questione può coinvolgere.
Non è quindi nostro intento proporre una ricetta per la realizzazione, all'alba del terzo millennio, dell'idea tolemaica, è altresì nostro desiderio descrivere un'esperienza che ha consentito, ad una modesta realtà quale il Servizio di documentazione scientifica (SDS) dell'Area della ricerca di Genova del Consiglio nazionale delle ricerche, la concretizzazione e la sperimentazione di un sistema di *library automation* per un settore scientifico circoscritto all'area biofisica.

Tab. 1 - Elenco delle testate scientifiche selezionate per la sperimentazione

- Annual review of biochemistry
- Annual review of biophysics and biomolecular structure
- Annual review of neuroscience
- Annual review of physiology
- Biochimica et biophysica acta. Biomembranes
- Biochimica et biophysica acta. Reviews on biomembranes
- Biological cybernetics
- Biopolymers
- European biophysics journal
- FEBS letters
- The Journal of cell biology
- The Journal of general physiology
- The Journal of membrane biology
- Journal of neurocytology
- Journal of neurophysiology
- The Journal of neuroscience
- The Journal of physiology
- Pfl¸gers archiv: European journal of physiology
- Physiological reviews
- The Plant cell
- Plant physiology
- Proteins
- Quarterly reviews of biophysics

In altre parole si è voluta creare parallelamente ad una biblioteca tradizionale, con tutti i ben noti vincoli di accesso e fruibilità, una biblioteca digitale totalmente scevra da limiti spazio-temporali.

Analisi del progetto

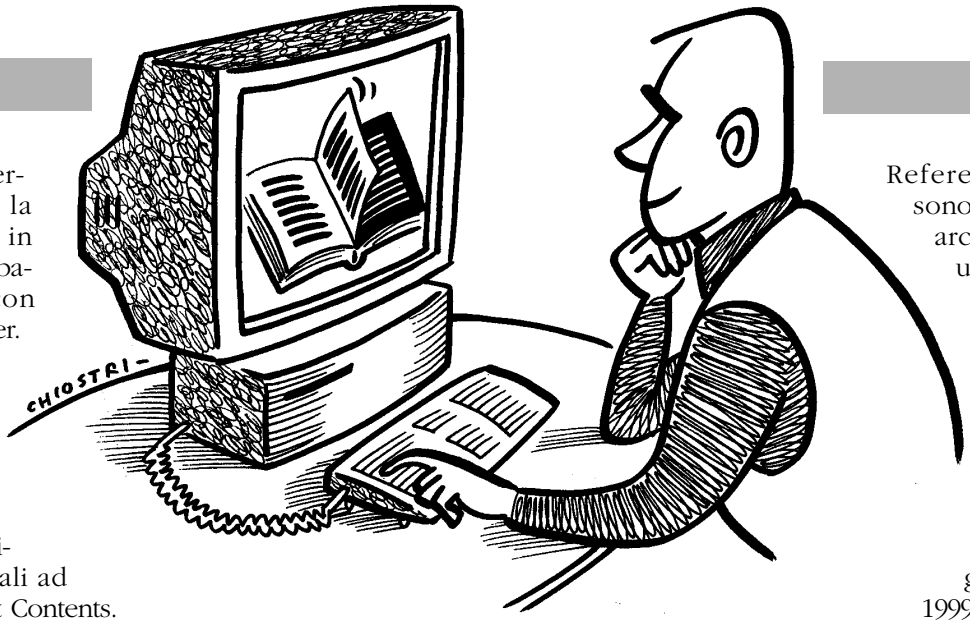
L'ipotesi progettuale si è articolata in varie fasi tra loro concatenate:
a) individuazione delle testate scientifiche da inserire nella sperimentazione;
b) validazione del sistema informativo per la caratterizzazione delle rassegne scientifiche;
c) analisi delle procedure per la digitalizzazione dei supporti cartacei;
d) progettazione dell'architettura hardware per la distribuzione dell'informazione;
e) scelta del protocollo di comunicazione per la veicolazione dei documenti digitali.

Testate scientifiche

I periodici da inserire nella sperimentazione dovevano soddisfare due precise condizioni: avere un argomento di copertura affine alla già citata area biofisica ed essere recensite dai Current Contents, condizione questa in grado di assicurarci, con cadenza settimanale, la disponibilità, in formato elettronico, degli indici, comprensivi di abstract dei fascicoli. Sulla base di queste due sole richieste sono state selezionate le testate riportate nella tab. 1.

Sistema informativo

A seguito della sperimentazione di numerosi sistemi, la scelta ritenuta più idonea alle nostre esigenze ha riguardato due prodotti della Research Information Systems: Reference Manager 8.5 e Reference web Poster 1.1. Il primo è un tipico software per la gestione di database bibliografici,[1] mentre il secondo



consiste nell'interfaccia web per la consultazione in Internet[2] dei database strutturati con Reference Manager. Altra peculiarità di quest'ultimo è la sua capacità di catturare i riferimenti bibliografici desunti da vari archivi elettronici, quali ad esempio i Current Contents.

Procedure di digitalizzazione

Volendo mantenere una esatta corrispondenza tra il documento a stampa e quello digitale ci si è rivolti al formato PDF (Portable Document Format). I documenti realizzati in tale formato sono visualizzabili, grazie ad un *viewer* gratuito, su macchine DOS, Windows, Unix e Macintosh; il documento in PDF è quindi multipiattaforma.[3]
Le strutture adottate per la digitalizzazione degli originali cartacei sono state due apparecchiature Xerox, in particolare Xerox Document Centre 220 ST e Xerox 5765 con interfaccia Fiery SI, rispettivamente per gli originali in bianco e nero e per quelli a colore; per entrambe le apparecchiature la risoluzione massima di acquisizione era di 400 dpi, ma per gli originali in bianco e nero si è sempre impiegata la risoluzione a 300 dpi. Le stazioni di lavoro per la concretizzazione del documento digitale erano due PC con processore Intel Pentium II a 266 Mhz con 64 Mb di RAM e sistema operativo Windows 95. Su dette configurazioni erano installati, oltre al necessario Adobe Acrobat Exchange 3.0, il PaperPort LE 4.02 per la gestione dei documenti digitali in bianco e nero e Adobe Photoshop 4.0 per la gestione dei documenti digitali a colori.

Architettura hardware

La configurazione hardware adottata per ospitare il sistema informativo (Reference Manager con interfaccia Reference Web Poster) era rappresentata da un Server NT 4.0 Primary Domain Controller su PC con processore Intel Pentium II a 333 Mhz con 256 Mb di RAM e con una memoria di massa pari ad oltre 80 Gbyte (9 hard disk Wide SCSI da 9.1 Gbyte ciascuno). Sul Server NT era inoltre installato Microsoft Internet Information Server 4.0 per la costruzione di un sito web il cui accesso era limitato, per ovvi motivi di copyright, ai soli utenti del dominio <ge.cnr.it>.
Per le necessarie operazioni di backup il sistema era dotato di un masterizzatore di cd.

Protocollo di comunicazione

La descrizione dell'architettura hardware utilizzata, con l'installazione di Microsoft Internet Information Server, è elemento di per sé sufficiente a far comprendere che il protocollo di comunicazione adottato è stato TCP/IP.

Descrizione delle fasi operative

Va innanzitutto precisato che con Reference Manager sono stati creati due archivi elettronici: uno tipicamente bibliografico che comprende, per tutte le suddette testate, gli indici dei fascicoli, comprensivi di abstract, per gli anni 1998 e 1999, il secondo rappresenta invece, sempre per ciascuna testata e per ciascun volume nel suddetto intervallo temporale, la classica schedatura biblioteconomica. Per semplicità di espressione i database sopracitati verranno di seguito rispettivamente denominati TOC e ARC.
Con cadenza settimanale si procedeva a catturare dalla sezione Life sciences dei Current Contents i riferimenti bibliografici relativi alle già citate testate e ad inserirli in TOC. Questa procedura non richiede alcun intervento manuale in quanto Reference Manager riconosce, come già detto, numerosi formati di output (circa 240) ed il formato NLM-medline restituito dai Current Contents è perfettamente interpretato da Reference Manager.
Al ricevimento delle copie cartacee dei fascicoli già inseriti in TOC si operava, secondo le modalità sopra descritte, la digitalizzazione dei documenti e si provvedeva, di conseguenza, a riportare nei relativi record di TOC il link al documento digitale. Anche questa operazione veniva sufficientemente automatizzata, in quanto i nomi attribuiti agli articoli digitali erano nel formato: XXX_YY.PDF, dove XXX corrisponde alla pagina di inizio dell'articolo e YY al numero del fascicolo della rivista e tali dati, già presenti in TOC, potevano essere utilizzati con estrema semplicità. In pratica si operava una ricerca in

TOC per il fascicolo relativo, si esportavano quindi i dati in formato RIS e si cancellavano i record dal database. Con una semplice macro in Word si rieditava il file in formato RIS e si reimportava in TOC. Un raffronto tra i due formati RIS, prima e dopo il link al documento digitale, viene proposto nell'Appendice.
Al completamento di un volume di una delle testate sotto osservazione si procedeva sia all'implementazione, manuale, di ARC per la rivista relativa, ovviamente con il link alla versione digitale dell'intero volume, sia alla creazione di una serie di pagine, in formato HTML, relative agli indici dei fascicoli che componevano detto volume. Questa operazione veniva automatizzata creando, in Reference Manager, un formato di output in grado di restituire appunto una pagina nel suddetto formato.

Benefici del sistema

Il progetto sin qui descritto offre all'utente telematico un doppio sistema informativo in grado di restituire allo stesso, naturalmente per il set di testate monitorate, una versione digitale degli articoli del tutto simile, per forma e contenuto, all'equivalente versione cartacea.
Data la diversa natura dei due archivi elettronici TOC e ARC è intuitivo che l'utente disporrà sia di un vero e proprio database bibliografico su cui operare ricerche mirate all'individuazione di articoli che soddisfino particolari esigenze (ad esempio: parole chiave nel titolo, anno di pubblicazione, autore ecc.), sia di un database in grado di restituire la versione digitale dell'intero volume.
A titolo di esempio vengono riportate in queste pagine le schermate web, conseguenti all'interrogazione attraverso il modulo Reference Web Poster, relative ai due sistemi informativi.

Fig. 1 - Database TOC: si ricerca un articolo di Bjorkoy, pubblicato su "Eupoean biophysics journal", che riporti nel titolo il termine "birefringence"

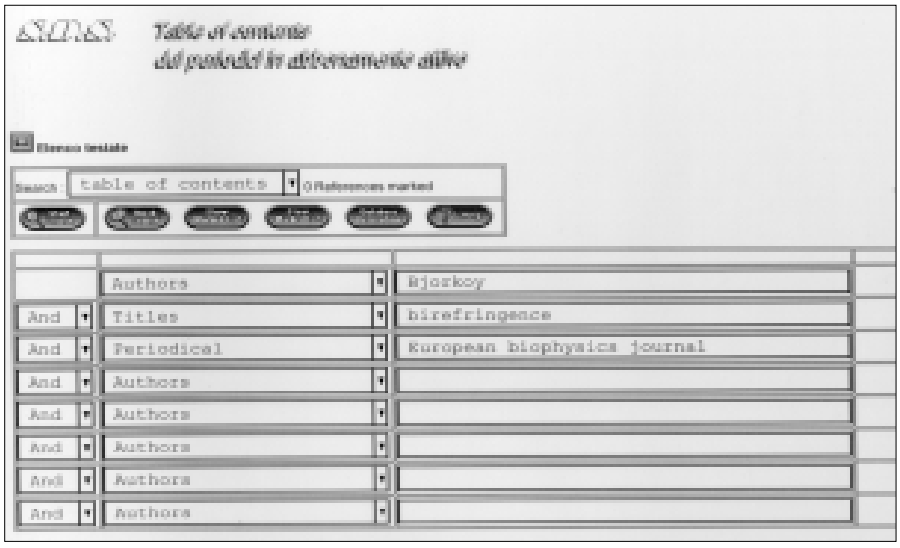
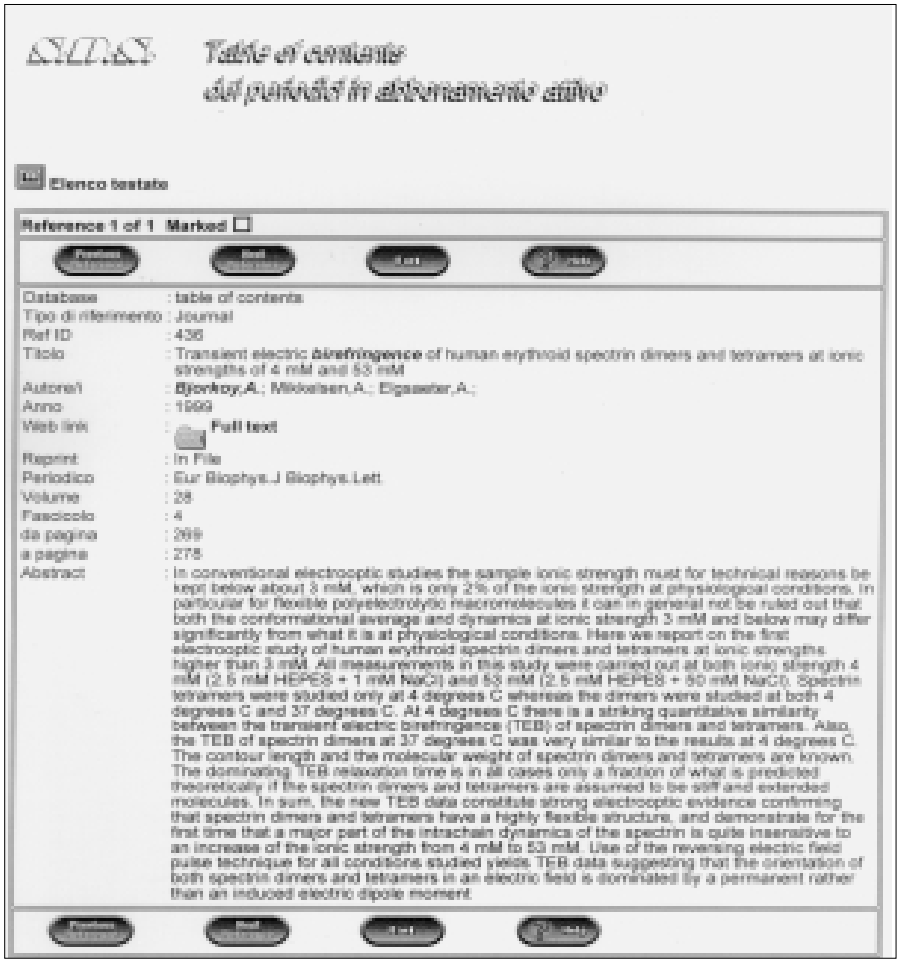


Fig. 2 - Database TOC: le condizioni di interrogazione restituiscono un solo riferimento bibliografico in cui è chiaramente visibile l'opportunità di accesso alla versione digitale dell'intero articolo



Come è desumibile da quanto sopra riportato, i due sistemi informativi, pur con filosofie differenti, consentono, in maniera distribuita ed in tempo reale, la fruibilità continua delle collezioni inserite nella sperimentazione.
Tale condizione, assolutamente non banale, tenta di risolvere tutti i limiti legati alla consultazione delle riviste su supporto tradizionale, ad esempio: orario di apertura della biblioteca, indisponibilità del materiale per operazioni di rilegatura, impossibilità di assecondare contemporanee richieste per la stessa fonte bibliografica, danneggiamento del supporto cartaceo conseguente alla fotoriproduzione dello stesso ecc.
Unica contropartita ai succitati vantaggi sono i tempi richiesti per la realizzazione dei documenti digitali, tempi peraltro assolutamente contenuti visto che la strumentazione adottata, per il bianco e nero, è in grado di digitalizzare venti pagine formato A4 in un minuto e quindi la completa trasformazione in PDF di un intero fascicolo di una rivista può essere mediamente compiuta in circa quindici minuti.
Va inoltre segnalato che i sistemi impiegati per la digitalizzazione del materiale cartaceo sono strumentazioni polifunzionali, in grado cioè di assolvere a diverse funzioni quali: a) fotocopiatura per originali sino ad A3; b) stampa anch'essa su formati sino ad A3; c) scansione per formati sino ad A3; d) fax solo per formato A4. Sono di conseguenza apparecchiature che abitualmente, almeno nella loro configurazione più limitata, trovano collocazione nelle biblioteche tradizionali. Stesso ragionamento può essere fatto per le configurazioni hardware: oggigiorno nessuna biblioteca può prescindere dalla disponibilità di un paio di PC a disposizione dell'utenza ed un Server NT è anch'esso condizione appena sufficiente per strutturare il sito web della biblioteca stessa.

Fig. 3 - Database ARC: si ricerca la presenza del volume 28 del 1999 del periodico "European biophysics journal"

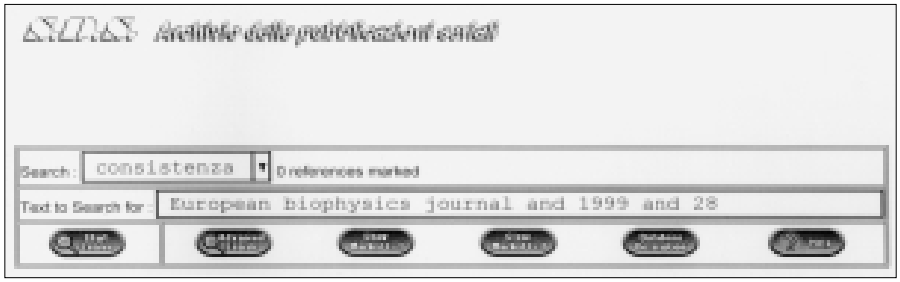


Fig. 4 - Database ARC: la ricerca restituisce un solo riferimento ove sono chiaramente indicati, oltre ai dati bibliografici della rivista, le indicazioni circa l'ubicazione della versione cartacea e la disponibilità della versione digitale

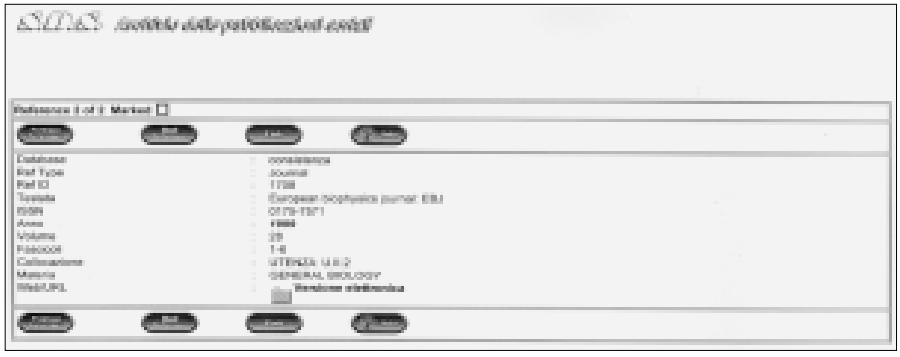


Fig. 5 - Database ARC: scegliendo l'opzione per la consultazione della versione digitale si accede a questa pagina ove possono essere selezionati i singoli fascicoli che compongono l'intero volume

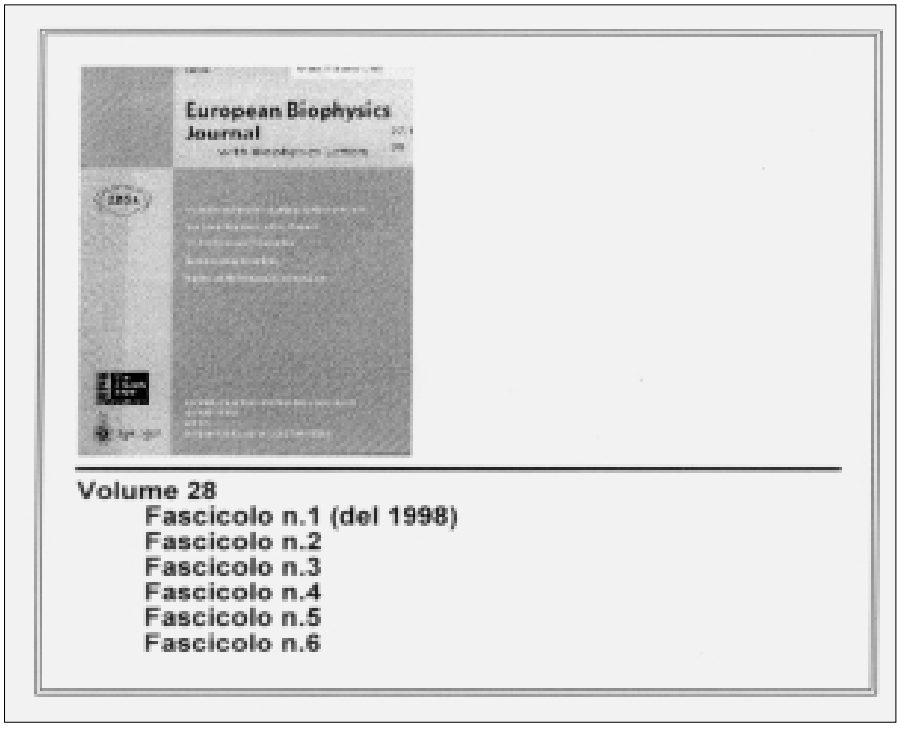
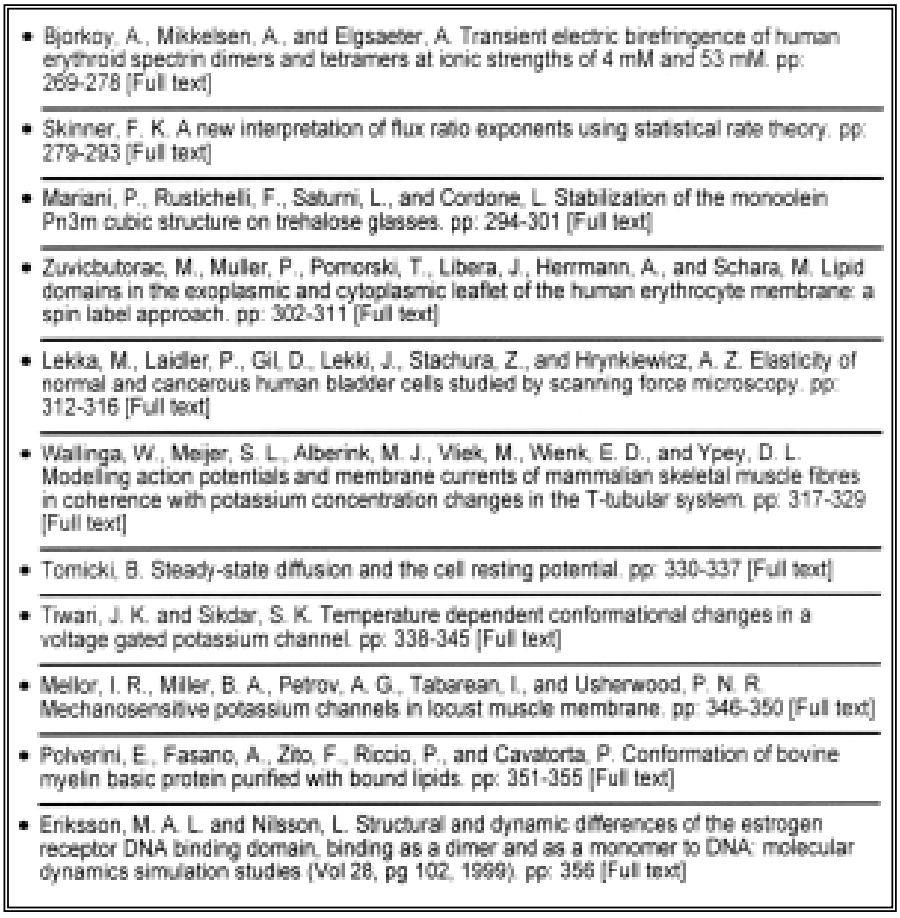


Fig. 6 - Database ARC: la selezione di uno dei fascicoli citati nella Fig. 5 (ad esempio il n.4) restituisce il contenuto dell'intero fascicolo con la conseguente possibilità di accesso alle versioni digitali di tutti gli articoli in esso presenti



- Bjorkoy, A., Mikkelsen, A., and Elgsaeter, A. Transient electric birefringence of human erythroid spectrin dimers and tetramers at ionic strengths of 4 mM and 53 mM. pp: 269-278 [Full text]
- Skinner, F. K. A new interpretation of flux ratio exponents using statistical rate theory. pp: 279-293 [Full text]
- Mariani, P., Rustichelli, F., Saturni, L., and Cordone, L. Stabilization of the monoolein Pn3m cubic structure on trehalose glasses. pp: 294-301 [Full text]
- Zuvicbutorac, M., Muller, P., Pomorski, T., Libera, J., Herrmann, A., and Schara, M. Lipid domains in the exoplasmic and cytoplasmic leaflet of the human erythrocyte membrane: a spin label approach. pp: 302-311 [Full text]
- Lekka, M., Laidler, P., Gil, D., Lekki, J., Stachura, Z., and Hrynkiewicz, A. Z. Elasticity of normal and cancerous human bladder cells studied by scanning force microscopy. pp: 312-316 [Full text]
- Wallinga, W., Meijer, S. L., Alberink, M. J., Vliek, M., Wienk, E. D., and Ypey, D. L. Modeling action potentials and membrane currents of mammalian skeletal muscle fibres in coherence with potassium concentration changes in the T-tubular system. pp: 317-329 [Full text]
- Tomicki, B. Steady-state diffusion and the cell resting potential. pp: 330-337 [Full text]
- Tiwari, J. K. and Sikdar, S. K. Temperature dependent conformational changes in a voltage gated potassium channel. pp: 338-345 [Full text]
- Mellor, I. R., Miller, B. A., Petrov, A. G., Tabarean, I., and Usherwood, P. N. R. Mechanosensitive potassium channels in locust muscle membrane. pp: 346-350 [Full text]
- Polverini, E., Fasano, A., Zito, F., Riccio, P., and Cavatorta, P. Conformation of bovine myelin basic protein purified with bound lipids. pp: 351-355 [Full text]
- Eriksson, M. A. L. and Nilsson, L. Structural and dynamic differences of the estrogen receptor DNA binding domain, binding as a dimer and as a monomer to DNA: molecular dynamics simulation studies (Vol 28, pg 102, 1999). pp: 356 [Full text]

Diverso ragionamento può valere per le soluzioni software impiegate, ma in questo caso le uniche osservazioni devono limitarsi ad Adobe Acrobat Exchange, cioè al software necessario per la realizzazione dei documenti digitali in formato PDF. Infatti tanto PaperPort LE quanto Adobe Photoshop vengono forniti con le configurazioni Xerox ed un software per la gestione del patrimonio bibliografico, quale appunto Reference Manager, deve considerarsi basilare per una qualunque biblioteca.
Non riteniamo sia il caso di addentrarci in lunghe disquisizioni in merito alle motivazioni, oltre a quelle già citate, che hanno determinato la scelta del formato PDF, vogliamo solo ricordare che attualmente tale formato può considerarsi uno standard per tutta l'editoria scientifica in formato elettronico.

Considerazioni conclusive

Non deve sembrare riduttivo aver limitato la sperimentazione a partire dal 1998, in quanto l'intento del progetto era soprattutto quello di verificare la reale fattibilità dello stesso; inoltre, l'area scientifica presa come modello è particolarmente sensibile alle fonti bibliografiche recenti.
Per fornire alcuni dati che possono meglio inquadrare la reale dimensione del nostro modello sperimentale, segnaliamo che per il 1998 le testate monitorate hanno pubblicato 6.716 rassegne scientifiche, cui corrispondono altrettanti documenti digitali in formato PDF, per complessivi 4.5 Gbyte di memoria di massa occupata. Per i primi otto mesi del 1999 le stesse riviste hanno prodotto 4.382 lavori, la digitalizzazione dei quali ha impegnato uno spazio disco di circa 3.5 Gbyte. È ragionevole ipotizzare che la memoria di massa necessaria per l'archiviazione di due annate per l'intero range di riviste sarà compresa tra 10 e 11 Gbyte e, conseguentemente, l'architettura del nostro attuale sistema è in grado di ospitare gli articoli pubblicati in circa quindici anni da tutte le testate scelte. Così come la componente hardware anche le soluzioni software adottate non possono essere messe in crisi dai numeri, infatti attualmente il database TOC è caratterizzato da soli 11.098 riferimenti bibliografici, ma il software è stato testato, ed ha risposto egregiamente su database di ordine di grandezza superiori; per il database ARC, attualmente caratterizzato da un centinaio di record, eventuali problemi di saturazione non sono neppure ipotizzabili.
L'esperienza qui descritta, pur con tutti i suoi limiti, vuole testimoniare una realtà che, una volta superati gli ostacoli di natura legale con gli editori dei periodici, è potenzialmente in grado di offrire alla comunità scientifica, nazionale e non solo, un esempio di biblioteca digitale. ■

Note

[1] Jill Feldt, *Materials science and technology databases*, "Advanced materials & processes", 1994, 4.
[2] Bruce R. Schatz, *Information retrieval in digital libraries: bringing search to the net*, "Science", 1997, 5298.
[3] Giancarlo Butti, *Il documento elettronico*, "I*Ged", 1998, 1.

Appendice

Esempio di record in formato RIS senza il link al documento digitale

```
TY  - JOUR
ID  - 130445
T1  - Structural and dynamic differences of the estrogen receptor DNA binding domain, binding as a dimer and as a monomer to DNA: molecular dynamics simulation studies
A1  - Eriksson,M.A.L.
A1  - Nilsson,L.
Y1  - 1999///
RP  - IN FILE
SP  - 102
EP  - 111
JF  - European biophysics journal
JA  - Eur Biophys.J Biophys.Lett.
J1  - EBJ
VL  - 28
IS  - 2
N2  - Molecular dynamics (MD) simulations of the estrogen receptor DNA-binding domain (ERDBD) as a dimer in complex with its DNA response element (ERE) show a significant difference in both structure and dynamics, compared to a MD simulation of monomeric ERDBD bound to its half-site response element (EREH). The C-terminal zinc binding domain (Zn-II), including a region (helix II) which is in a helical conformation in ERE- (ERDBD)(2). is considerably more flexible in EREH-ERDBD than in the dimeric complex. In EREH-ERDBD, all helical hydrogen bonds in helix II are broken and the entire Zn- II region is detached from a hydrogen bonding network that in ERE-(ERDBD)(2) connects to other parts of the protein as well as to the DNA. The regions that become flexible in EREH-ERDBD are identical to the regions where the NMR solution structure of free ERDBD is poorly ordered. This strongly suggests that dimerisation of ERDBD is required for ordering of the Zn-II region and that monomeric binding to DNA is not sufficient for the ordering. This contrasts to the glucocorticoid receptor DNA-binding domain (GRDBD) which has essentially the same mobility (uniform and limited), regardless of whether it is free as a monomer in solution, bound as a monomer to its half-site response element or in a dimeric complex with the full response element. The hydrogen bonding network that connects Zn-II with other parts of the protein and to DNA is almost identical in ERDBD and GRDBD. However, in GRDBD there is also a serine (in the N-terminal zinc coordinating region) with a central role in this network, connecting to the Zn-II region. This serine is replaced by a glycine in ERDBD and we suggest that this substitution is sufficient for destabilisation of the network, thus leading to a more flexible ZnII region, which becomes ordered first upon Forming a complex with another ERDBD and DNA ER.
```

Esempio di record in formato RIS con il link al documento digitale

```
TY  - JOUR
ID  - 130445
T1  - Structural and dynamic differences of the estrogen receptor DNA binding domain, binding as a dimer and as a monomer to DNA: molecular dynamics simulation studies
A1  - Eriksson,M.A.L.
A1  - Nilsson,L.
Y1  - 1999///
RP  - IN FILE
SP  - 102
EP  - 111
JF  - European biophysics journal
JA  - Eur Biophys.J Biophys.Lett.
J1  - EBJ
VL  - 28
IS  - 2
N2  - Molecular dynamics (MD) simulations of the estrogen receptor DNA-binding domain (ERDBD) as a dimer in complex with its DNA response element (ERE) show a significant difference in both structure and dynamics, compared to a MD simulation of monomeric ERDBD bound to its half-site response element (EREH). The C-terminal zinc binding domain (Zn-II), including a region (helix II) which is in a helical conformation in ERE- (ERDBD)(2). is considerably more flexible in EREH-ERDBD than in the dimeric complex. In EREH-ERDBD, all helical hydrogen bonds in helix II are broken and the entire Zn- II region is detached from a hydrogen bonding network that in ERE-(ERDBD)(2) connects to other parts of the protein as well as to the DNA. The regions that become flexible in EREH-ERDBD are identical to the regions where the NMR solution structure of free ERDBD is poorly ordered. This strongly suggests that dimerisation of ERDBD is required for ordering of the Zn-II region and that monomeric binding to DNA is not sufficient for the ordering. This contrasts to the glucocorticoid receptor DNA-binding domain (GRDBD) which has essentially the same mobility (uniform and limited), regardless of whether it is free as a monomer in solution, bound as a monomer to its half-site response element or in a dimeric complex with the full response element. The hydrogen bonding network that connects Zn-II with other parts of the protein and to DNA is almost identical in ERDBD and GRDBD. However, in GRDBD there is also a serine (in the N-terminal zinc coordinating region) with a central role in this network, connecting to the Zn-II region. This serine is replaced by a glycine in ERDBD and we suggest that this substitution is sufficient for destabilisation of the network, thus leading to a more flexible ZnII region, which becomes ordered first upon Forming a complex with another ERDBD and DNA UR - <A HREF=http://sds.area.ge.cnr.it/Full_text1/FTXT6/EBJ/V28/102_2.pdf><img align=top src=/Draws/file.gif border=0></A><B><BLINK> Full text</B></BLINK>
```