Dieci anni di Emeroteca Virtuale. Una panoramica sul servizio di *Digital Library* del coordinamento CIBER

UGO CONTINO

Abstract

This paper illustrates what has been done by CASPUR in the past ten years for its Emeroteca Virtuale service. CIBER (Comitato Interuniversitario Basi dati ed Editoria in Rete) Digital Library, known as "Emeroteca Virtuale" (EV), provides access to more than 5300 scholarly e-journals, mostly on STM disciplines, to authorized users (students, researchers and professors from 26 universities of CIBER consortium). On occasion of its tenth anniversary, we want to show what changes EV has undergone, focusing on various elements (users, usage data, access types, hardware) which characterize the service on the whole, in order to have a view as global as possible of it. Future developments of the server will be also presented at the end of this article, showing how they have been suggested by users' legitimate needs and expectations.

Sintesi

Questo lavoro presenta l'esperienza maturata all'interno del CASPUR con il suo servizio di Emeroteca Virtuale, che quest'anno compie dieci anni. L'Emeroteca Virtuale del CIBER è, come noto, un servizio di accesso a testate elettroniche di tipo multidisciplinare: la sua base di utenti comprende studenti, ricercatori e docenti di una trentina tra università ed enti di ricerca in Italia. In occasione del suo decennale si vuole illustrare attraverso quali cambiamenti l'emeroteca sia passata, analizzando i vari aspetti (la base utenti, le statistiche d'uso, le modalità di accesso, le funzionalità avanzate, la piattaforma hardware*) che contribuiscono alla definizione complessiva del servizio, in modo da fornirne una panoramica quanto più possibile esaustiva. Completano l'articolo alcune considerazioni su quelli che potranno essere i futuri sviluppi dell'emeroteca, in linea con le legittime aspettative dei suoi utenti.*

> *L'universo (che altri chiama la Biblioteca) si compone d'un numero indefinito, e forse infinito, di gallerie esagonali, con vasti pozzi di ventilazione nel mezzo, bordati di basse ringhiere. Da qualsiasi esagono si vedono i piani superiori e inferiori, interminabilmente….*
>
> Jorge Luis Borges, *La Biblioteca di Babele*, 1941

Introduzione

L'uso di strumenti (comunemente noti come *Digital Libraries*) di accesso a risorse elettroniche a *testo completo,* siano esse rappresentate da articoli scientifici su riviste elettroniche, ovvero di *e-book* aventi una valenza scientifica, è un realtà ormai ampiamente ben consolidata nel panorama della comunità scientifica internazionale. Basti pensare al fatto che se si prova a cercare un testo su Internet che tratti di questo specifico argomento, il numero di risultati ottenuti è dell'ordine di grandezza delle migliaia di unità[43].

Tale scenario non era affatto così consolidato una decina di anni fa, quando il progetto dell'Emeroteca muoveva i suoi primi passi, il CIBER[1] non era una realtà così consolidata ed i vantaggi di una *soluzione consortile* per l'accesso a comuni risorse elettroniche erano ancora lungi dall'essere recepiti e pienamente compresi. In tal senso l'Emeroteca Virtuale (EV nel seguito) ha rappresentato, per lo meno nel panorama italiano, un servizio certamente innovativo, non privo quindi di una sua connotazione per certi versi pioneristica: scarsa infatti era la conoscenza del *servizio* (se lo si intende nell'àmbito degli strumenti *software* che la rete Internet poteva allora offrire) e scarse erano le conoscenze sulle problematiche (*hardware* e *software*) che a lungo andare

43 Una ricerca della stringa "Digital Library" fatta su Google Books ritorna un numero di risultati superiore a 16.000 opere monografiche (Ottobre 2009)

avrebbero dovuto esser prese in considerazione. Problematiche poste dalla necessità di mantenere, da una parte, un livello qualitativo all'altezza delle aspettative di una classe di utenza che, nell'arco di quasi un decennio, si è più che triplicata, e dall'altra di garantire una *scalabilità dell'infrastruttura* in grado di tollerare un tasso di crescita medio dei contenuti di circa un milione di nuovi articoli a testo completo l'anno [2].

Anche limitandosi a considerare unicamente questi due specifici aspetti (*qualità* del servizio e *qualità* dell'infrastruttura) si può ragionevolmente affermare che l'EV costituisca un successo per tutti coloro che vi hanno lavorato, sia all'interno del CASPUR, che nel più ampio gruppo del CIBER, se non altro per aver saputo raccogliere attorno a se l'interesse di una comunità scientifica rappresentativa della maggior parte delle strutture universitarie e di ricerca dell'Italia centro-meridionale, in un periodo di tempo sufficientemente lungo dal non doverlo considerare l'ennesimo servizio *sperimentale*.

I successivi paragrafi saranno dedicati a mostrare il percorso seguito in questi anni focalizzando l'attenzione sugli aspetti *oggettivi* del servizio ovvero: i *contenuti* e la sua *base utenti*; le *statistiche d'uso*; la comunità dell'*utenza registrata*; l'*infrastruttura hardware*. Al termine verrà fornito qualche indicazione su quella che sarà la struttura della nuova piattaforma e si delineeranno quelli che potrebbero essere gli sviluppi futuri del servizio, in linea con le sfide lanciate dalla comunità di utenti attuale e prossima ventura.

Il servizio di Emeroteca Virtuale: che cosa e per chi

Il servizio si inquadra storicamente nel contesto delle attività iniziate alla fine degli anni 90 nell'àmbito della collaborazione CIBER (Coordinamento Interuniversitario per le Basi di dati e l'Editoria in Rete), che ha visto inizialmente coinvolti i 5 atenei allora consorziati con il CASPUR (il Politecnico di Bari e le Università di Bari, Lecce, Roma "La Sapienza" e Roma Tre), e che si è evoluta, successivamente, in una struttura che conta ad oggi 26 atenei partecipanti. La tabella seguente mostra una breve storia dell'evoluzione del CIBER in funzione degli enti aderenti, che allo stato odierno raccolgono una popolazione studentesca complessiva pari al 40% di quella italiana. Dal 2008 il CIBER è anche organo del consorzio CASPUR.

ANNO	NUMERO MEMBRI	ISTITUZIONI PARTECIPANTI
1999	5 CASPUR	Politecnico di Bari, Università di Bari, Università di Lecce, Università di Roma La Sapienza, Università RomaTre
2000	8 membri	Ai primi 5 membri si aggiungono: Università della Basilicata, Università di Perugia, Università di Salerno
2001	13 membri	Agli 8 membri del 2000 si aggiungono: Università di Camerino, Università della Calabria, LUMSA, Università di Macerata, Università di Palermo
2002	22 membri	Ai 13 membri del 2001 si aggiungono: Università Campus Biomedico, Università di Cassino, Università dell'Aquila, Università di Messina, Università del Molise, Università Parthenope di Napoli, Seconda Università di Napoli, Università di Roma2,Università della Tuscia
2003	26 membri	Ai 22 membri del 2002 si aggiungono: Università di Chieti, Università di Foggia, Università degli Studi di Roma "Foro Italico", Università di Reggio Calabria
2004	26 membri	Come nel 2003
2005	26 membri	Come nel 2003
2006	27 membri	Ai 26 membri del 2006 si aggiunge: Università di Teramo
2007	26 membri	Esce la Seconda Università di Napoli
2008	27 membri	Entra la SISSA di Trieste

Tab. 1 – Andamento delle adesioni al CIBER nel periodo 1999-2008

Per poter meglio comprendere come questo scenario di crescita si sia tradotto in un incremento della base potenziale di utenti dell'EV, si faccia riferimento alle figire 1 e 2, dove sono mostrate le *FTE* (*full-time equivalent*) degli studenti iscritti ed i totali dei docenti e dei ricercatori delle università e degli enti di ricerca aderenti al CIBER [3] (dati aggiornati all'a. a. 2007-2008).

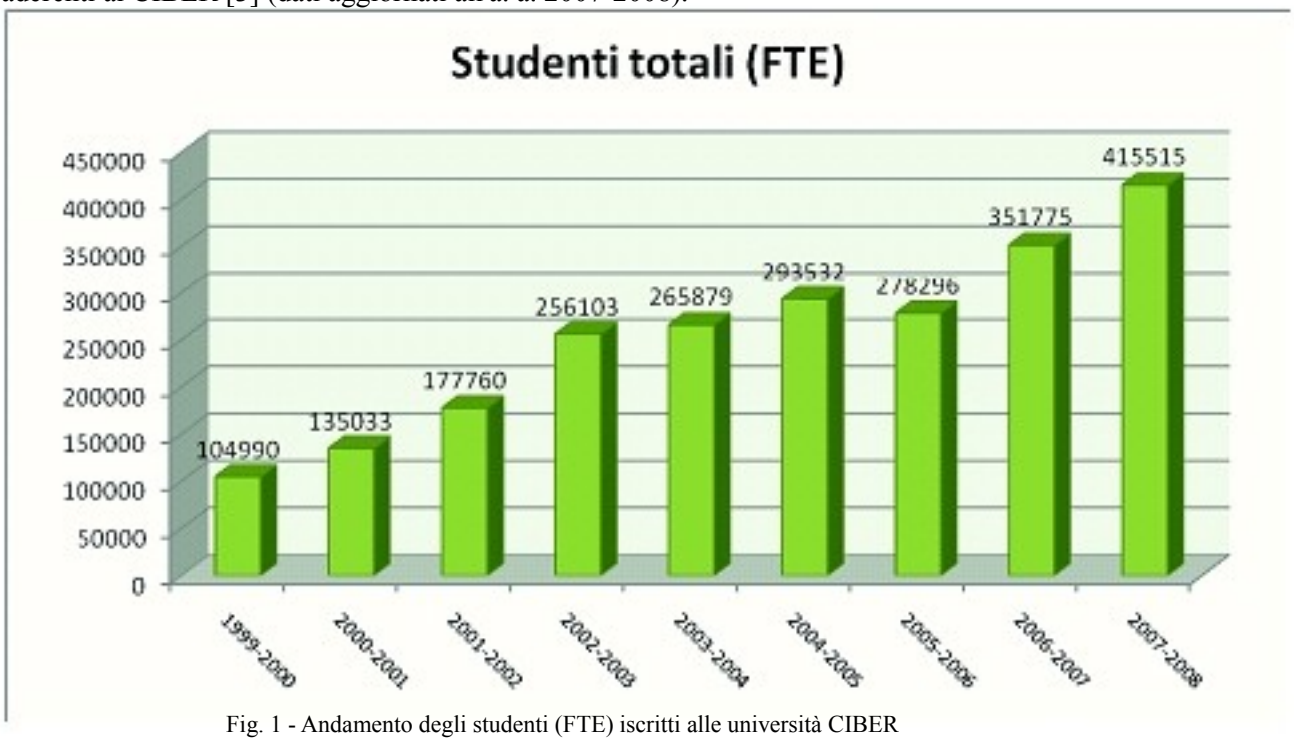


Fig. 1 - Andamento degli studenti (FTE) iscritti alle università CIBER

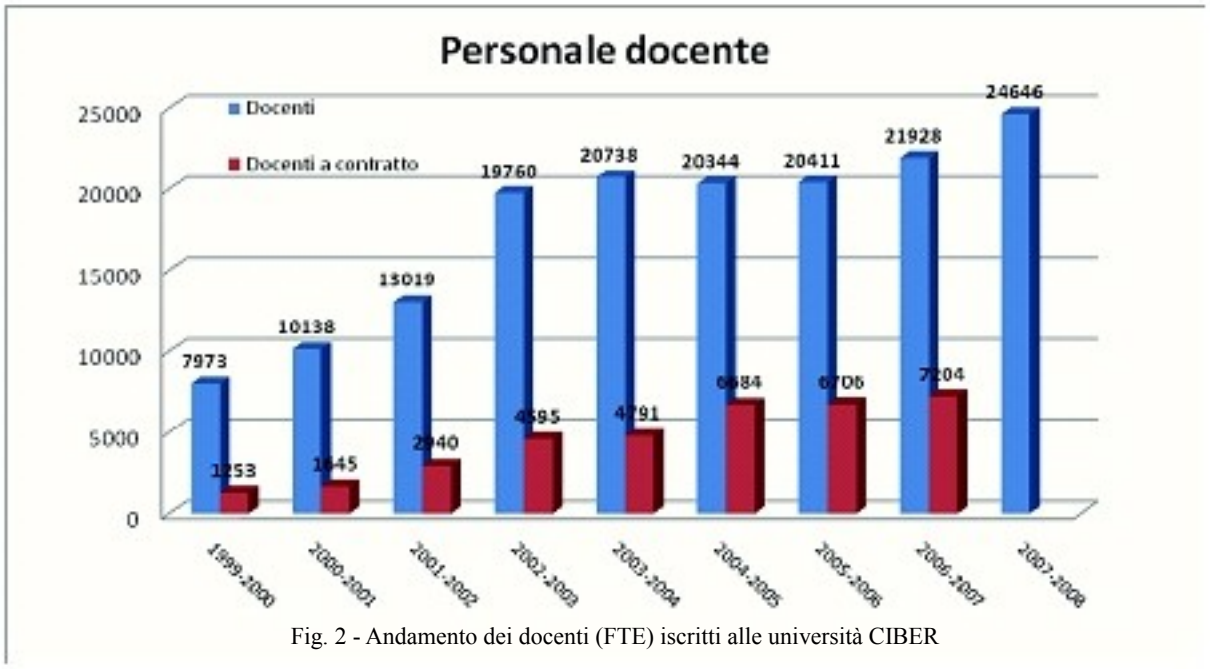


Fig. 2 - Andamento dei docenti (FTE) iscritti alle università CIBER

Com'è possibile notare nell'arco di un decennio, circa, il numero degli studenti *full-time equivalent* e del personale docente è più che triplicato.

Per avere un'idea di come si sia evoluto invece lo scenario dell'EV in relazione ai contenuti, si faccia riferimento grafico di figura 3, nella quale è mostrato il numero di riviste accessibili, inclusivo sia di quelle a testo completo (*full text*), che di quelle di cui si posseggono i soli metadati *(in tal caso sull'EV è visibile l'abstract degli articoli, mentre il full-text risiede sul sito dell'editore).*

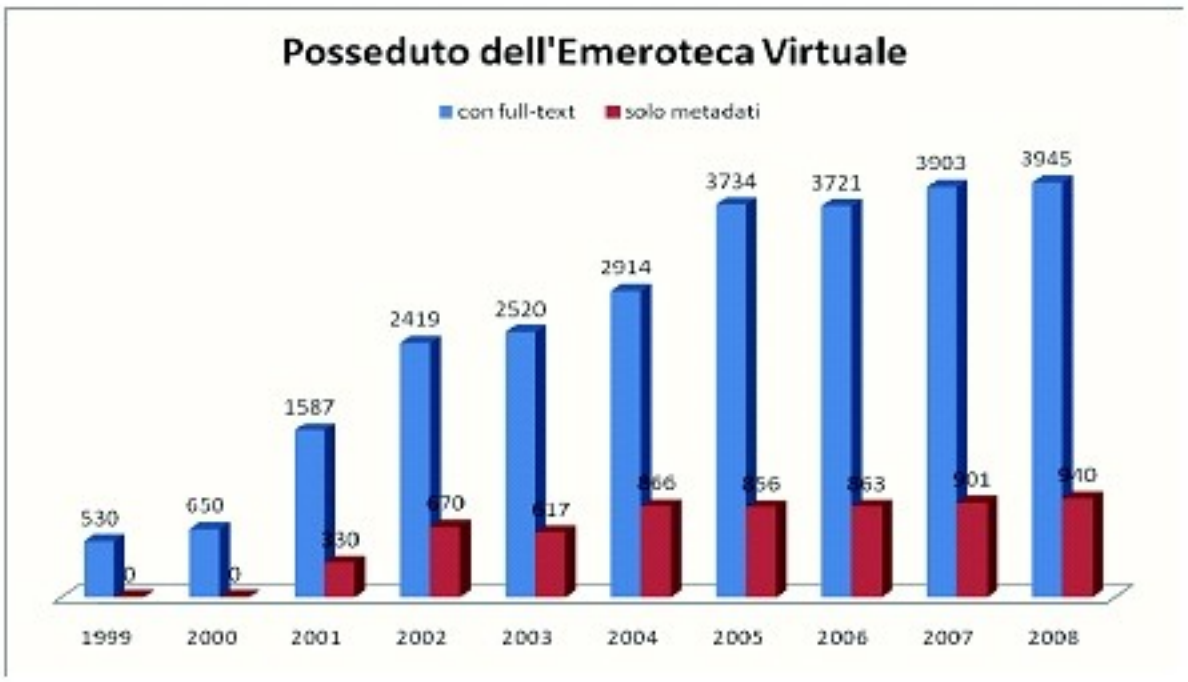


Fig. 3 - Numero di riviste accessibili in EV nell'arco degli anni 1999-2008

Nell'arco di poco più di un anno (dalla fine del '99 all'inizio del 2001) si è avuto una triplicazione delle riviste a testo completo (dall'editore Elsevier), insieme con la comparsa delle prime testate con i soli metadati (Blackwell). L'andamento crescente del grafico non solo è testimonianza dell'incremento del posseduto dell'editore Elsevier, ma anche dell'attivazione di nuovi contratti di accesso a testo pieno con altri editori (IOPP, Kluwer, Springer). Attualmente (ottobre 2009), sono accessibili in Emeroteca 5.328 riviste (di cui l'84% a testo pieno), comprendenti poco meno di 450.000 fascicoli e per un totale di quasi 7.950.000 articoli (si veda la *tabella seguente).*

Editore	Riviste a *testo pieno*	Riviste con i soli metadati
American Chemical Society	—	40
Blackwell Publishing	—	739
Coordinamento SIBA	5	—
Elsevier Science	2432	—
Insitute of Physics Publishing	52	—
Kluwer AP/Springer Business & Media	158	—
Kluwer Law International	22	—
Medknow Publications	—	69
Springer	1506	—

Tab. 2 - Editori presenti in Emeroteca Virtuale (Settembre 2009)

Statistiche d'uso dell'Emeroteca Virtuale: il decennio trascorso

Sin dagli esordi del servizio di Emeroteca Virtuale, particolare cura è stata posta nella definizione di un ambiente di produzione delle *statistiche d'uso*, intendendo con esso la quantità di accessi registrati da parte dell'utenza accreditata e relativi a *download* specifici (di articoli, di *abstract*, di indici di riviste, etc.). È stato a tal fine definito un modulo *ad hoc*, tramite il quale gli utenti autorizzati (i *referenti* dei vari atenei partecipanti) potevano accedere ad uno specifico

portale, attraverso cui era possibile impostare i parametri di analisi dei *file* di log nei quali venivano *registrati gli accessi all'EV. Dopo aver impostato questi parametri (periodo di analisi; università di appartenenza; holdings di riferimento), un programma di elaborazione dati (basato sul software di* analisi statistiche SAS [4]) provvedeva ad effettuare le dovute operazioni e ad inoltrare, all'indirizzo di e-mail specificato dall'utente, un *report PDF* nel quale erano contenute le seguenti informazioni:

distribuzioni cumulative sugli accessi ai contenuti (riviste, fascicoli, abstract e full-text);
distribuzioni cumulative sugli accessi agli editori;
statistiche dettagliate sugli accessi alle singole testate;
distribuzioni top-n delle riviste consultate.

Quest'approccio, rimasto in auge fino al 2005, è stato abbandonato quando si è imposto, all'attenzione degli addetti ai lavori, un nuovo progetto di razionalizzazione e standardizzazione delle informazioni statistiche legate all'uso delle risorse elettroniche contenute nelle *digital library* (degli editori o consortili): il progetto Counter [5].

Il modulo precedente è stato quindi sostituito da un modulo *software*, costruito *ex-novo* e sviluppato con linguaggi *open-source*, che produce report *counter compatibili* con cadenza mensile. Tali *report* statistici, specifici per i singoli atenei e cumulativi per l'intero consorzio, vengono periodicamente depositati all'interno di una particolare area web riservata ai soli referenti CIBER [6], dove possono essere consultati e conservati.

Nel medesimo periodo è stata avviata un'attività di studio da parte dei membri del Centro di Ateneo per le Biblioteche dell'Università di Messina, che, utilizzando i dati estratti dalle statistiche di accesso dell'EV (utilizzati nel procedimento di produzione di *report Counter*), ha portato alla costruzione di un portale sulle statistiche d'uso[44], con un contenuto informativo ben più ricco di quello offerto dai *report counter* [7]. Entrambi i sistemi sono ad oggi attivi e costituiscono la base delle informazioni sulle statistiche d'uso dell'Emeroteca Virtuale disponibili per il consorzio CIBER.

Tralasciando gli aspetti relativi alle modalità tecniche con le quali sono stati implementati i moduli dedicati alle analisi statistiche (per le quali si rimanda alle note bibliografiche), si vuole invece mostrare come è evoluto nel periodo 2001-2008 il principale parametro di stima dell'*uso* dell'EV da parte dei suoi utenti: il numero di articoli *full-text* scaricati (nei due formati, PDF e *HTML). Si veda a tal fine l'istogramma di figura 4.*

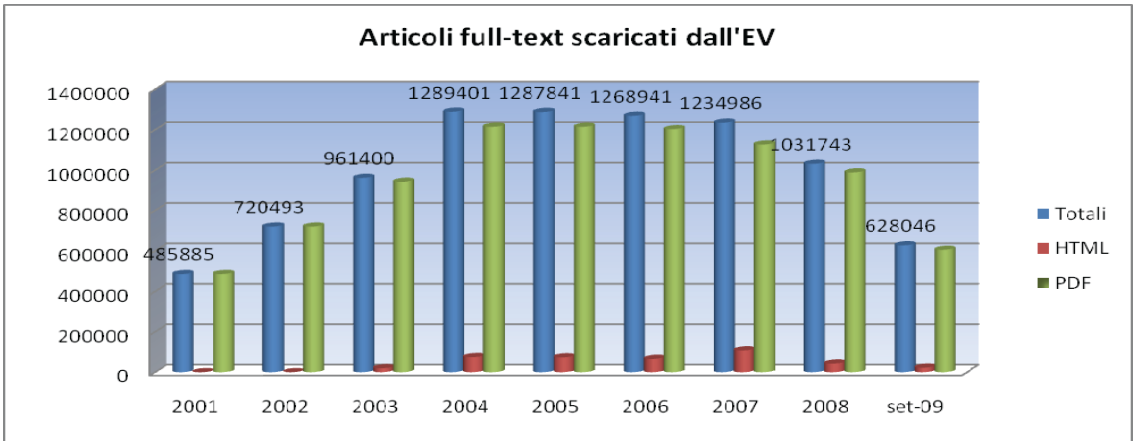


Fig. 4 - Numero di articoli *full-text* scaricati nel periodo 2001-2009 (fine settembre)

Nella figura 5 si riporta l'andamento del numero di articoli scaricati in funzione dell'editore nel periodo gennaio 2005-settembre 2009 per i tre editori Elsevier, IOPP e Kuwer, successivamente acquisita da Springer (per le annate precedenti non sono disponibili statistiche in formato counter

44 Il portale è consultabile da tutti i referenti CIBER all'indirizzo: <cab.unime.it/ciber/stat> (verifica 24 ottobre 2009)

relative agli articoli scaricati per ogni singolo editore).

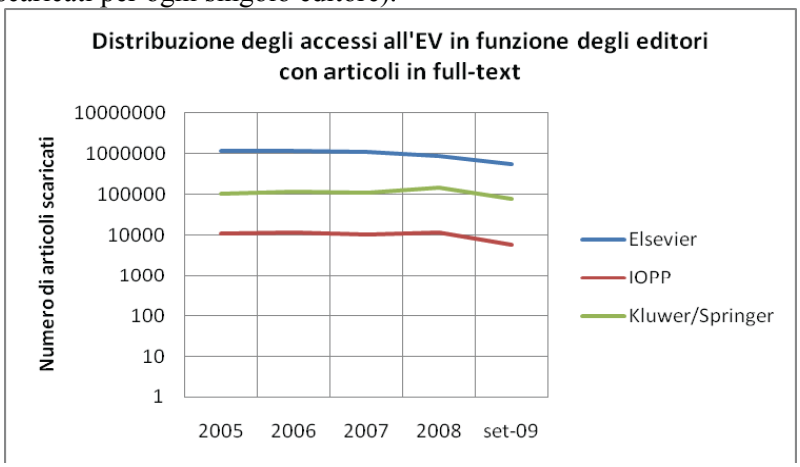


Fig. 5 - Numero di articoli full-text scaricati nel periodo 2005-2009 (fine settembre) in funzione dell'editore

Queste statistiche, oltre a mostrare un andamento decrescente a partire dal 2006 per tutti e tre gli editori, hanno un altro elemento degno di nota: il rapporto tra le tre serie di dati. Infatti per ciascun editore (partendo da Elsevier) i dati sono di un ordine di grandezza inferiore passando all'editore successivo per mole di articoli scaricati. Benchè il rapporto tra le riviste pubblicate da IOPP e Elsevier possa giustificare questo rapporto (2.430 contro 50 circa), lo stesso non si può dire per l'editore Springer (2.430 contro 1.500 circa, in questo caso): una possibile giustificazione di questo andamento potrebbe essere ricercata nel maggior *impact factor* associato alle riviste Elsevier rispetto a quelle dell'editore Springer.

Tornando alla fig.4, come detto in precedenza, l'altro elemento degno di nota, è il calo di *accessi* registrato a partire dal 2006. Quest'ultimo potrebbe essere in realtà direttamente correlato alla possibilità, a partire dal 2005 per alcuni atenei del consorzio e dal 2007 per tutti i suoi componenti, di poter accedere alle riviste anche sulla piattaforma dell'editore Elsevier, che di fatto rappresenta l'editore con il maggior numero di accessi in emeroteca (l'accesso ai siti degli altri editori è stato consentito sin dall'avvio dei relativi contratti). L'uso di banche dati bibliografiche e/o citazionali che fanno riferimento ad un articolo puntando direttamente al sito ufficiale della rivista, unito al fatto che tipicamente l'utenza dell'EV vi accede in maniera *indiretta* e quasi mai attraverso la sua *home*, possono giustificare pienamente questo andamento, soprattutto se lo si correla con quello degli articoli scaricati sul portale *Science Direct* dell'editore Elsevier. La figura 6 riporta l'andamento degli articoli scaricati dall'Emeroteca e dal sito dell'editore Elsevier per gli anni 2005-2008: analizzando i due istogrammi è possibile notare come a partire dal 2008 gli utenti del CIBER accedano più sul sito *istituzionale* della specifica rivista Elsevier, che non sul sito dell'Emeroteca.

Nella figura 7, infine, è riportato il valore del rapporto tra il numero di articoli annualmente scaricati dall'emeroteca in funzione della sua *base utenti*, ottenuta sommando all'FTE degli studenti iscritti alle università partecipanti, anche il relativo personale docente. Questi dati suggeriscono le seguenti osservazioni: primo, c'è un interesse crescente verso l'EV, testimoniato dal fatto che, nonostante l'aumento degli enti consorziati e, quindi, della sua base utenti, tale rapporto aumenti anche lui, passando da 2,51 a 4,22 articoli scaricati *in media* per ogni utente del CIBER; secondo, la flessione notata sugli scarichi totali dall'EV negli ultimi due anni, si riflette anche sul numero di articoli medi scaricati per utente (istogrammi in blu). Tuttavia se consideriamo anche gli accessi sulla piattaforma *Science Direct* [45] nel medesimo anno il valore ottenuto è pari a 4,98 e 4,94, rispettivamente, in linea con l'aumento registrato nei precedenti cinque anni.

45 *Rif.* <www.sciencedirect.com> *(verificato il 20 ottobre 2009)*

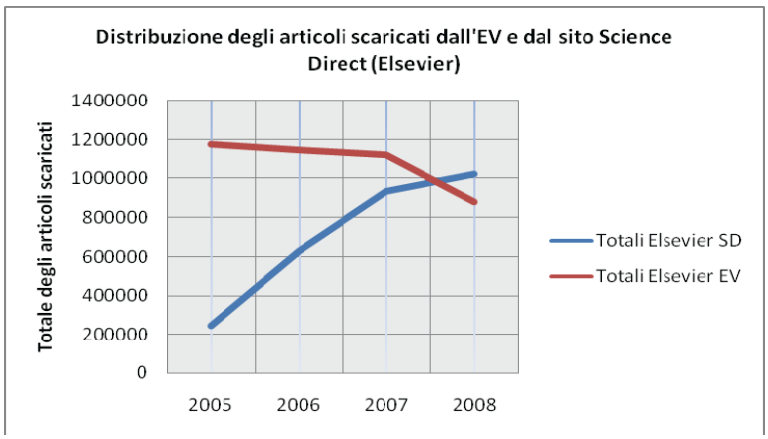


Fig. 6 – Confronto tra il numero di articoli full-text scaricati nel periodo 2005-2008 dal sito dell'EV e dal sito *Science Direct*

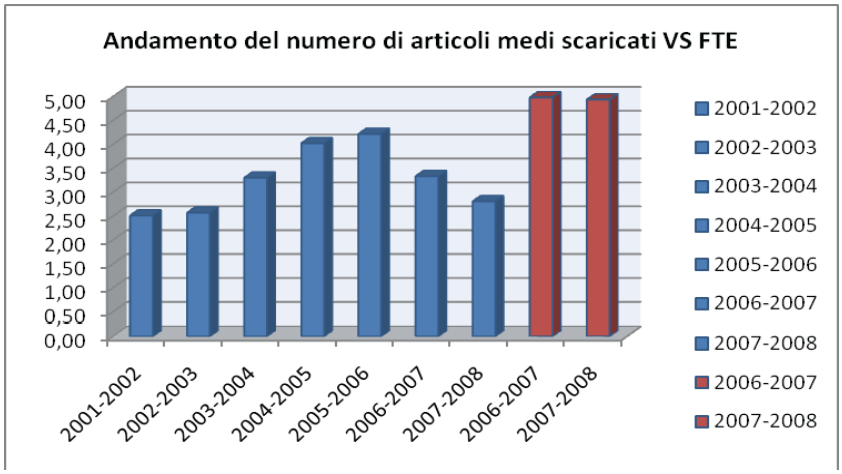


Fig. 7 – Andamento del numero medio di articoli scaricati annualmente in funzione della FTE del consorzio CIBER (i dati ripetuti per gli a. a. 2006-2007 e 2007-2008 fanno riferimento a statistiche che considerino anche gli accessi sulla piattaforma Science Direct dell'editore Elsevier)

Tuttavia, benché si registri una flessione annuale nel numero di *download* dalla piattaforma dell'EV, fatto che testimonia un maggiore accesso, da parte dell'utenza del CIBER, alle piattaforme istituzionali dei vari editori, dovuto probabilmente, come detto, ad una maggiore *visibilità* di questi ultimi all'interno delle più comuni banche dati citazionali o disciplinari, bisogna evidenziare comunque un elemento fondamentale: ovvero che l'emeroteca, in virtù delle particolari condizioni contrattuali sugli articoli *full-text* forniti dai vari editori, condizioni che permettono di mantenere tali dati all'interno dell'Emeroteca anche in occasione di contratti non più rinnovati, rappresenti sempre più un punto di riferimento del CIBER per ciò che concerne *l'accesso perpetuo* agli articoli scientifici, fatto che porta come ovvia conseguenza l'attenzione che, in futuro, andrà posta sulla gestione delle *politiche di preservazione* dei suoi contenuti elettronici a medio e lungo termine.

Questo risulta evidente dal grafico di figura 7, dove si sono analizzate le distribuzioni percentuali degli articoli scaricati nel quinquennio 2005-2009 in funzione dell'anno di pubblicazione del fascicolo e considerando le *distanze* in anni tra l'anno in cui è stata effettuata il *download* e l'anno di pubblicazione dell'articolo. Come risulta evidente in tutti gli anni in cui è stata effettuata l'analisi, il numero relativo di articoli scaricati appartenenti a fascicoli pubblicati almeno quattro anni prima, rappresenta di gran lunga la percentuale maggiore, non essendo mai inferiore al 40%, con una tendenza incrementale per tutte le annate considerate, raggiungendo un picco del 50% per le ultime due.

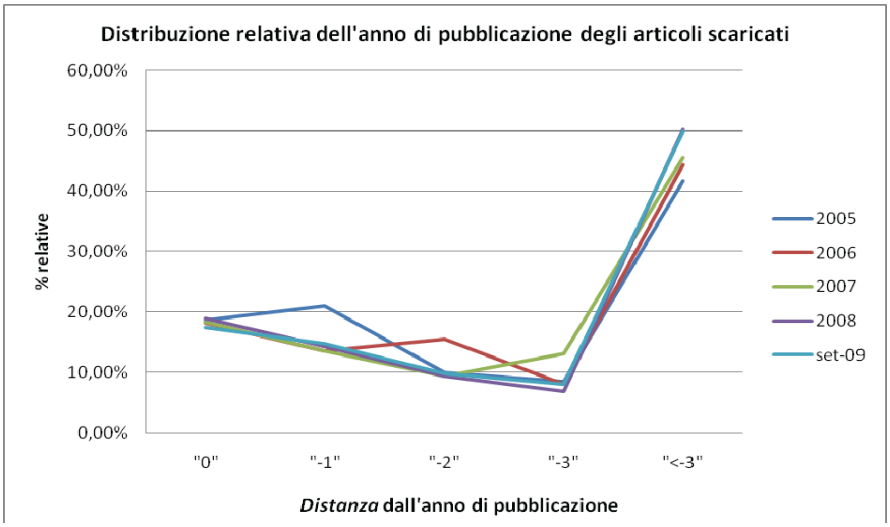


Fig. 7 – Distribuzioni percentuali relative degli anni di pubblicazione degli articoli scaricati nel periodo 2005-2009

L'utenza "registrata" dell'Emeroteca Virtuale: dal servizio di *ricerca personalizzata* a quello di *accesso remoto*

A partire dal 2001 nell'Emeroteca Virtuale è stata prevista una sezione dedicata agli utenti registrati, sfruttando a tal fine una funzionalità intrinseca del *software* Science Server, rappresentata dal *modulo di ricerca personalizzata*. Registrandosi a questo servizio l'utenza dell'Emeroteca accede ai servizi di ricerca personalizzata (sezioni *Search History* e *Saved Searches*) tramite i quali può salvare le ricerche eseguite e/o modificarle, salvare gli articoli di maggior interesse in un'area locale (sezione *My Articles*), avviare processi di ricerca automatica (sezione *Saved Searches*).

La procedura di registrazione al servizio di ricerca personalizzata avviene attraverso la compilazione di un modulo web accessibile dall'*home page* dell'Emeroteca, mentre l'associazione dell'utenza al giusto ateneo di appartenenza è fatta sulla base del dominio di posta elettronica fornito durante la registrazione.

Nel contesto dell'Emeroteca, quindi, per *utenza registrata* si è voluto intendere quell'utenza riconosciuta dal sistema grazie ad una propria *login*, piuttosto che all'indirizzo IP di provenienza. Tale approccio è stato mantenuto anche quando, nell'agosto del 2004, è stato introdotto il servizio di accesso remoto, che ha permesso una completa delocalizzazione dell'utente, garantendo a quest'ultimo la possibilità di accedere alla piattaforma indipendentemente dal suo punto di ingresso ad Internet.

Volendo ora caratterizzare da un punto di vista oggettivo il servizio di ricerca personalizzata, si può far riferimento alla sottostante figura, dove è mostrato l'andamento delle registrazioni degli utenti tra il 2001 ed il 30 Luglio 2004, data oltre la quale è stato ufficialmente introdotto il servizio di accesso remoto.

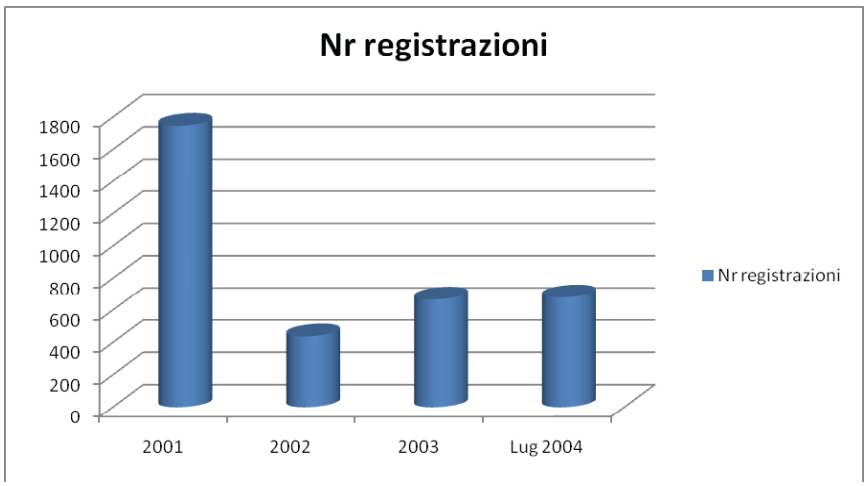


Fig. 8 – Registrazioni al servizio di ricerca personalizzata

Il dato mostrato per il 2001 è in realtà un dato cumulativo per il triennio 1999-2001; ipotizzando un tasso medio di adesione più o meno costante, il valore ottenuto è in linea con quello relativo al successivo triennio 2002-2004. Come si può notare, nonostante l'ampliamento del consorzio CIBER a nuovi atenei, con un conseguente aumento della popolazione complessiva di utenti, non si è avuto un corrispondente aumento di utenza registrata. Ciò potrebbe essere imputato: ad una scarsa diffusione a livello locale dell'informativa relativa a questo specifico servizio all'interno di ciascun ateneo del consorzio; ovvero al fatto che la pagina tipica con cui l'utenza del CIBER accede non sia la *home* dell'emeroteca (dove è presente il *link* alla sezione di registrazione al servizio di ricerca personalizzata), ma la *home* della specifica rivista, cosa che maschera di fatto l'accesso alla sezione di registrazione; ovvero, infine, ad uno scarso uso di questi strumenti da parte degli utenti, rispetto a quelli relativi alla *ricerca* (semplice ed avanza) degli articoli o di *browsing* delle riviste.

Soprattutto per fornire a tale servizio un elemento di maggior interesse per l'utenza del CIBER, è stata introdotta come detto in precedenza, nella metà del 2004, la funzionalità di *accesso remoto*. La bontà di questa scelta può esser dedotta analizzando il grafico di figura 9, dove è mostrato l'andamento delle registrazioni al servizio di accesso remoto.

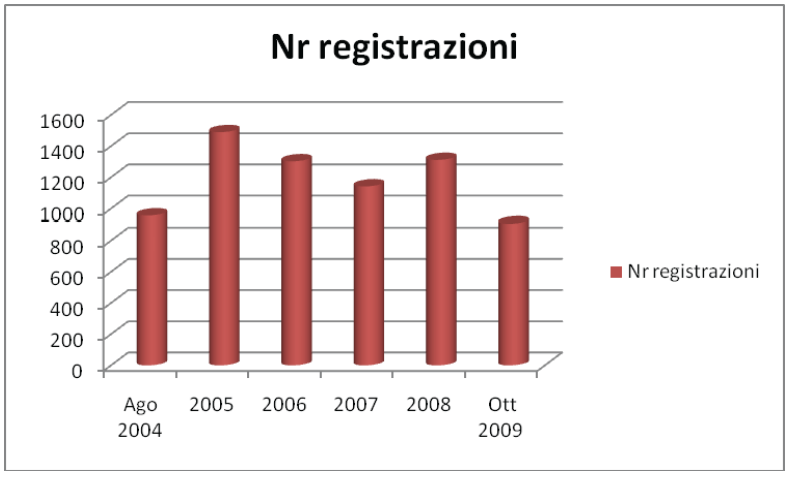


Fig. 9 - Registrazioni al servizio di accesso remoto

Rispetto al caso precedente si può notare, infatti, come il servizio di accesso remoto abbia incontrato un maggior successo rispetto al precedente servizio di ricerca personalizzata, grazie

probabilmente al concorso di due fattori: il primo, legato ad una migliore opera di pubblicizzazione di quest'ultimo all'interno del singolo ateneo; il secondo, legato ad una caratteristica intrinseca del servizio dovuta al suo saper rispondere ad un'esigenza concreta dell'utenza, ovvero quella di poter accedere ai *full-text* dell'Emeroteca anche al di fuori degli àmbiti, talvolta angusti, della rete universitaria. Questo stesso elemento positivo lo si può trovare nella figura successiva, che mostra l'andamento delle registrazioni cumulative al servizio di accesso remoto: al mese di ottobre 2009 gli utenti registrati sono più di 7100, il che da una media di circa quattro nuovi utenti al giorno.

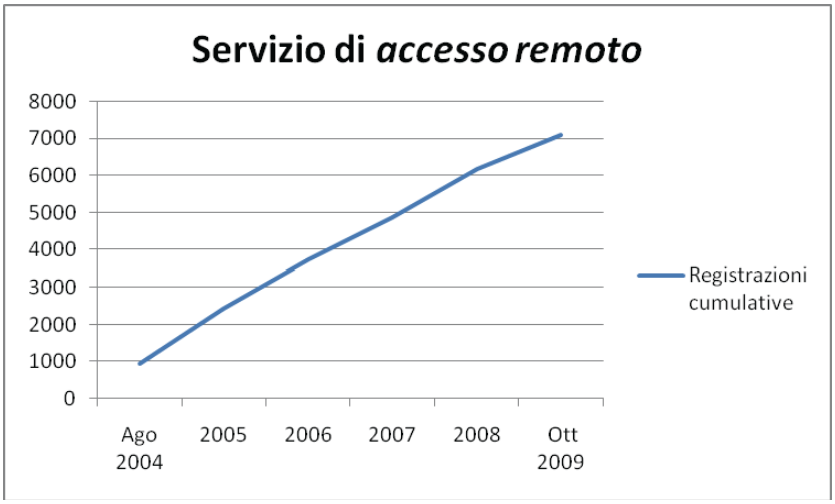
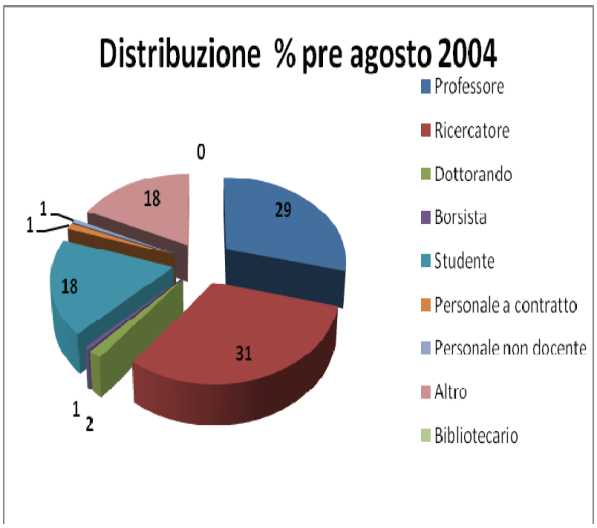


Fig. 10 - Andamento delle registrazioni cumulative al servizio di accesso remoto

Per quanto riguarda la distribuzione delle classi di utenza *pre-* e *post-* Agosto 2004, valgono le due distribuzioni nelle figure 11 e 12.



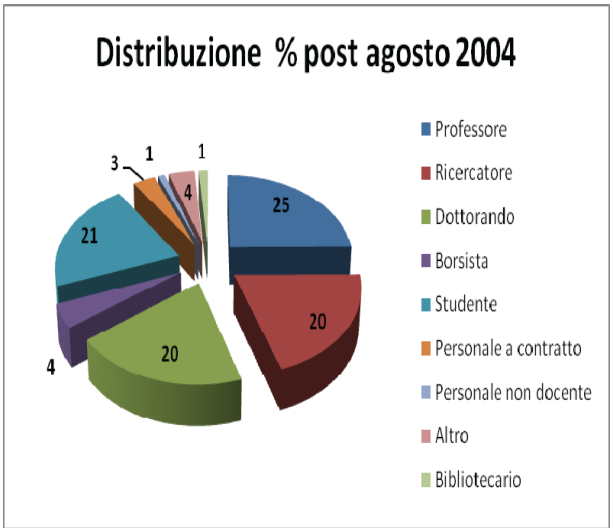


Fig. 11 – Distribuzione pre-Agosto 2004

Fig. 12 – Distribuzione post-Agosto 2004

Come si può notare confrontando le due distribuzioni, il servizio di ricerca personalizzata si caratterizza per una presenza maggiore di ricercati e docenti, rispetto alle altre classi di utenza, mentre quello di accesso remoto mostra un notevole apprezzamento da parte della classe dei dottorandi e dei borsisti ed un costante interesse da parte degli studenti. Interessante notare inoltre come la percentuale dei professori superi in questo caso quella dei ricercatori divenendo di fatto la prima classe di utenza del servizio.

Nei due successivi istogrammi vengono invece mostrate le distribuzioni degli iscritti al servizio di accesso remoto (quindi degli utenti iscritti dopo l'agosto del 2004) per tutti gli enti che partecipano al CIBER; tali istogrammi sono forniti sia in funzione dell'anno, che come totali degli iscritti

nel periodo 2004-2009. È possibile osservare come il numero degli iscritti al servizio rifletta la dimensione dell'ateneo in termini di popolazione universitaria servita. Non mancano casi di realtà molto attive, quali l'Università Campus Bio-Medico, dove a fronte di un modesto numero di docenti e studenti iscritti, il numero di utenti registrati è percentualmente elevato (si confronti, nell'istogramma di figura14 il numero totale degli iscritti di questo ateneo con quello degli iscritti dell'Università La Sapienza di Roma).

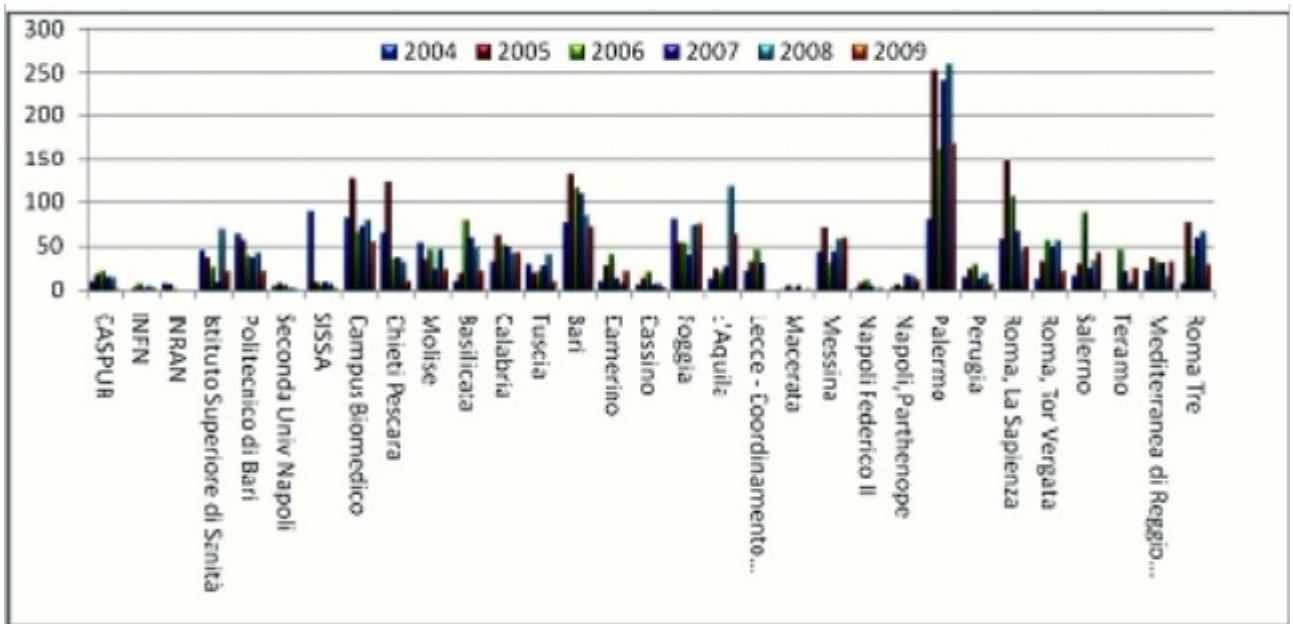


Fig. 13 - Distribuzione dell'iscrizione al servizio di accesso remoto nel periodo 2004-2009

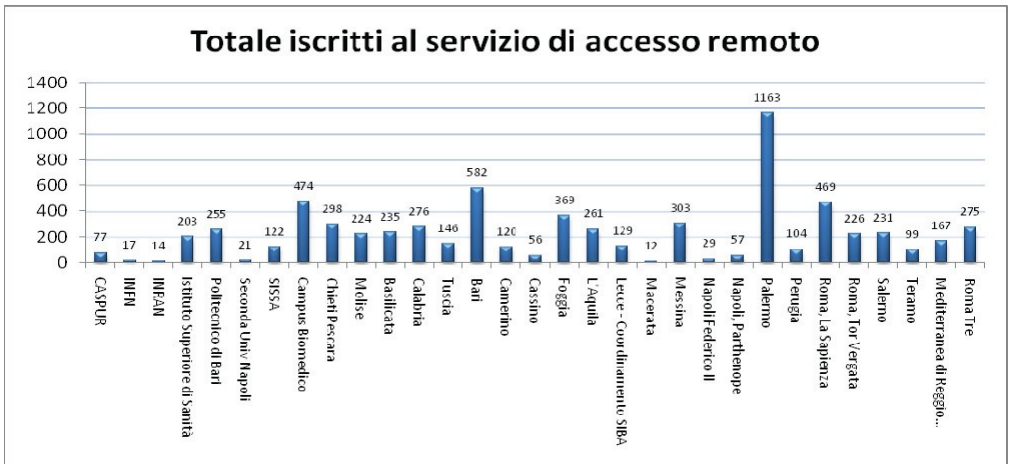


Fig. 14 – Totale degli iscritti al servizio di accesso remoto nel periodo 2004-2009

È evidente da quest'ultimo istogramma come Palermo rappresenti l'università con il maggior numero di iscritti al servizio di accesso remoto (1.163 al mese di Ottobre 2009). Come ulteriore elemento oggettivo dell'apprezzamento di tale servizio, può essere citato l'andamento del rapporto percentuale dei *download* degli articoli da parte di questa classe di utenti rispetto al totale degli articoli scaricati (figura 15) attraverso l'autenticazione classica (basata sull'indirizzo IP).

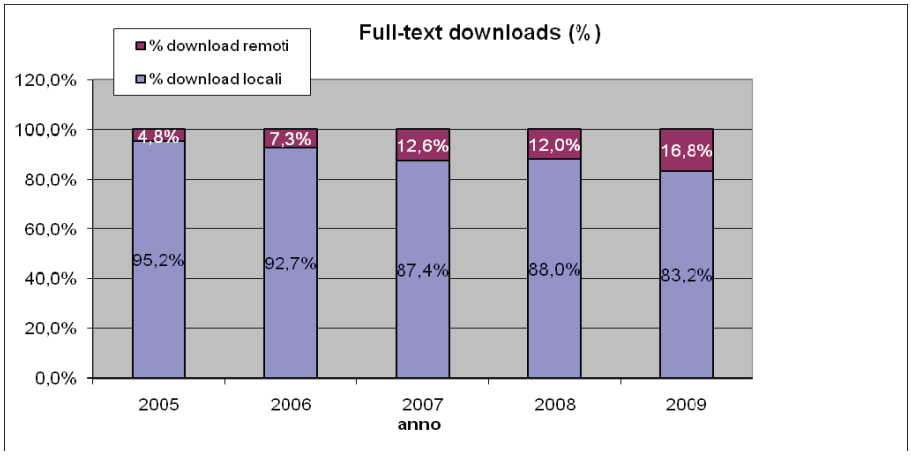


Fig. 15 – Rapporto % tra i downloads *remoti* e *locali*

Per capire invece quale ateneo abbia gli utenti *remoti* più attivi, si faccia riferimento all'istogramma di figura 16, che riporta il numero medio di articoli scaricati nel 2008 da questa categoria di utenti, suddivisi per ateneo di appartenenza. La media degli articoli scaricati per quell'anno è stata di 26,2 articoli ad utente, con picchi superiori ai 70 articoli in media ad utente (atenei di Foggia e di Reggio Calabria).

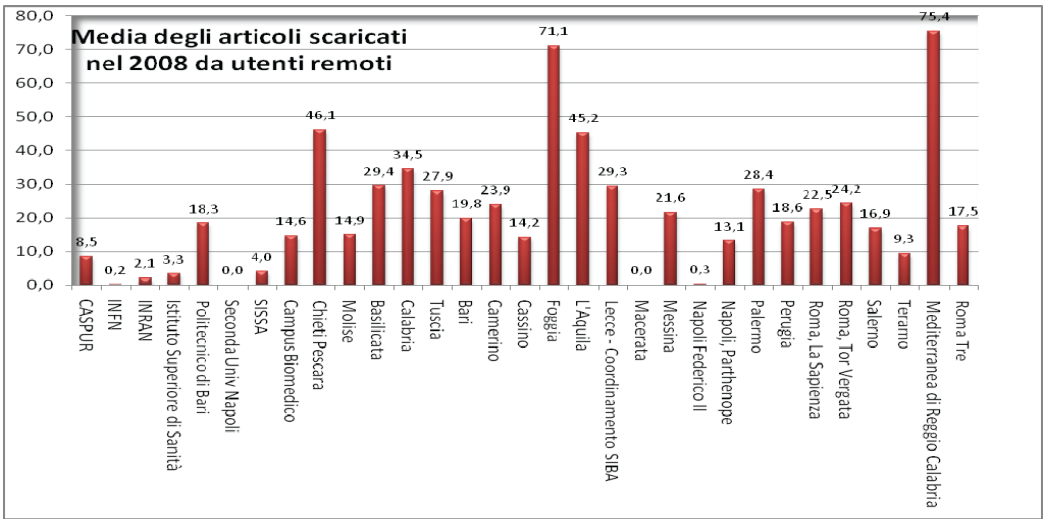


Fig. 16 – Media degli articoli scaricati da utenti remoti

Da quanto mostrato, si può dedurre come questo servizio si caratterizzi per l'emeroteca come un elemento sempre più *critico* dal punto di vista funzionale. È per questo motivo che è attualmente allo studio una nuova soluzione per la gestione dell'*utenza remota*. Questa soluzione, che prevede l'uso del DB MySQL per la gestione degli utenti e dei dati ad essi associati ed offre maggiori garanzie di scalabilità e *robustezza* rispetto a quella attuale, diverrà operativa con il rilascio della nuova versione dell'Emeroteca Virtuale, previsto per la fine del corrente anno.

Evoluzione della piattaforma hardware

Parallelamente ad un processo evolutivo che ha interessato il *software* utilizzato per il servizio di EV, questi dieci anni trascorsi hanno visto l'avvicendarsi di soluzioni tecniche, differenti per caratteristiche e tipologia, ma con un comune denominatore: offrire un servizio *sempre* disponibile e *senza* ritardi. Proprio con queste due parole può essere riassunto lo sforzo progettuale che è stato e che è tutt'ora alla base dell'attività di coloro che progettano e mantengono la piattaforma *hardware* dell'emeroteca all'interno del CASPUR.

Più in dettaglio, gli elementi di scelta della migliore architettura sui quali ci si è basati per offrire il servizio sono sempre stati correlati:

1. alle prestazioni del sistema lato utente (*Users' Quality of Service*, *U-QoS*) e alla disponibilità del servizio (intesa dal punto di vista della *service availability*);
2. al numero di utenti potenzialmente coinvolti;
3. alla scalabilità nello *storage* e nella capacità di I/O (Input/Output) verso le memorie di massa del DB degli Indici degli articoli, dell'area dei metadati e dei *full-text*;
4. alla flessibilità della soluzione per ciò che concerne l'integrazione con sistemi non-proprietari;
5. ai costi sostenibili per questo servizio.

Quest'approccio ha portato, nell'arco degli anni, all'impiego di diverse soluzioni. I primi passi l'emeroteca li ha compiuti su un sistema SUN (SUN-Enterprise 450 [46]), dotato di un'architettura *quadri-processore UltraSPARCII @ 480 MHz,* di una memoria RAM da 4 GB e con un'area disco esterna da 500 GB su *bus SCSI*. Questa architettura ha avuto l'indubbio merito di rappresentare una *robusta* base di partenza sulla quale avviare il servizio per i suoi primi tre anni e mezzo di vita (fino all'agosto del 2003). Col tempo ha però mostrato i suoi limiti, uno dei quali legato al *costo* notevole di manutenzione dell'*hardware* e del sistema operativo (proprietario in entrambi i casi). Un altro elemento che ha pesato nella scelta di una diversa soluzione è stato quello dell'assenza di *ridondanza* di tipo architetturale, cosa che comportava il fermo del servizio per interventi di manutenzione per arresti imprevisti del sistema. Altro limite poteva essere trovato nella sua scarsa flessibilità dal punto di vista della *scalabilità*, o della sua bassa *espandibilità* relativamente alla memoria dedicata ad ospitare metadati e PDF degli articoli caricati.

Per ovviare a quest'ultimo limite si è preferita una soluzione basata su sistemi di *memoria di massa* esterni al *server* (in gergo tecnico *Storage Area Network*), mantenendo quest'architettura nel tempo, ad eccezione di una sola modifica, quando si è deciso di passare nel 2002 da un collegamento su *bus* SCSI ad uno su *fiber channell* [47] con un *tranfer rate* di 1 Gbit al secondo (1 Gbps).

Come detto in precedenza, la soluzione SUN è stata abbandonata nel secondo semestre del 2003 a favore di una struttura certamente innovativa per questo tipo di servizio, ovvero quella del *server cluster*. Il *cluster* inizialmente utilizzato era rappresentato da due server Linux (chiamati *periodici1* e *periodici2*), con caratteristiche speculari: ciascuno di essi, basato su un'architettura interna a doppio processore Intel Xeon® con *clock* a 2.4 GHz, aveva una RAM di 4 GB ed era dotato di tecnologia *Hyper-Trading* [48], grazie alla quale, sfruttando i cicli di *idle* del processore è stato possibile raddoppiarne virtualmente le capacità computazionali.
L'accesso alla *Storage Area Network* (SAN) avveniva su *bus fibre-channel* con *throughput* complessivo di 2 Gbit/sec in modalità *full-duplex*. I dischi utilizzati negli *array* esterni erano di tipo Ultra-ATA[49], erano installati in sistemi RAID5 della Infortrend[50], e garantivano una capacità di memoria complessiva pari a 4 TB (4000 GB). Si veda la figura 17 per un quadro schematico della soluzione *hardware* adottata.

46 Rif. <sunsolve.sun.com/handbook_pub/validateUser.do?target=Systems/E450/spec> (verifica 20 ottobre 2009)

47 Rif. <www.intel.com/technology/hyperthread/index.htm?iid=ipp_srvr_proc_xeon+feature_f2htt&>
(verifica 20 ottobre 2009)

48 Rif. <www.intel.com/technology/hyperthread/index.htm?iid=ipp_srvr_proc_xeon+feature_f2htt&>
verifica 20 ottobre 2009)

49 Rif. <www.seagate.com/support/kb/disc/ultra_ata100.html> (verifica 20 ottobre 2009)

50 Rif. <www.infortrend.com> (verifica 20 ottobre 2009)

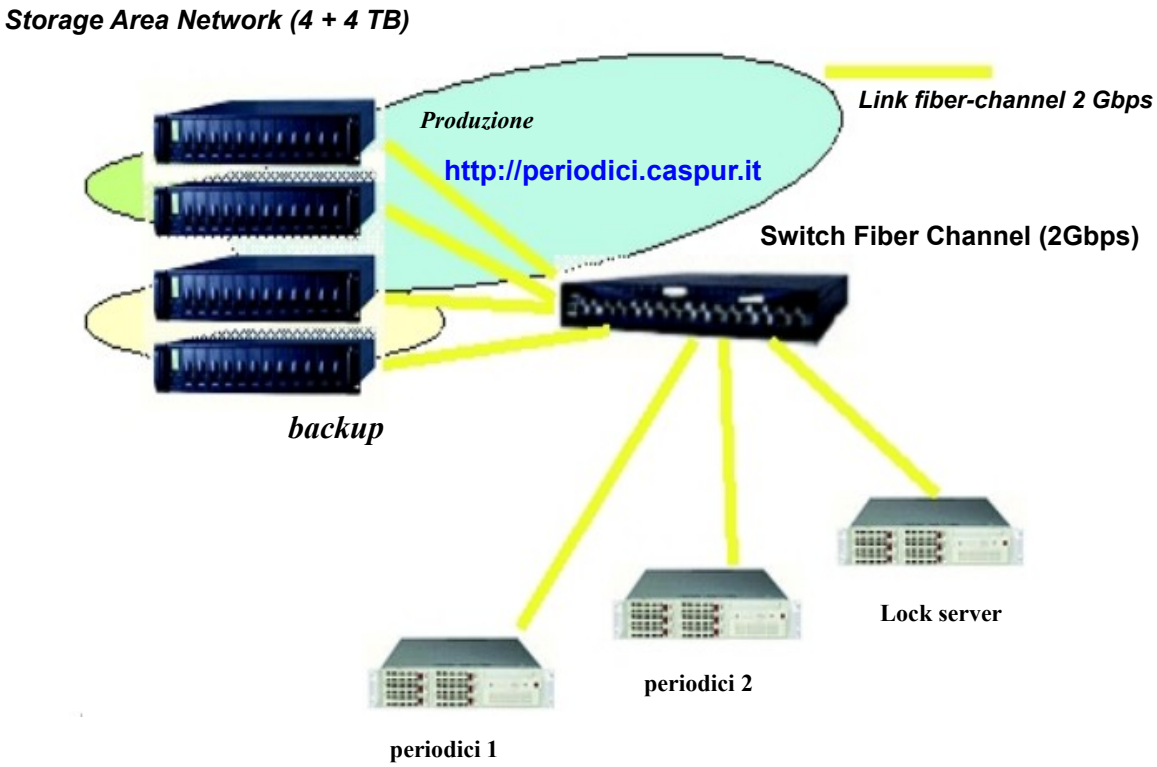


Fig. 17 – Schema dell'infrastruttura adottata a partire dall'agosto 2003

La macchina indicata come *lock-server* aveva lo scopo di gestire l'accesso alla risorsa condivisa rappresentata dall'*array* dei dischi esterni, sui quali risiedevano:

la *Digital Object Repository* (che ospita i file PDF ed SGML degli articoli);
il DB degli Indici e dei *link*;
la gerarchia delle *directory* XML;
le pagine statiche dell'Emeroteca;
il codice eseguibile di Science Server e degli script *custom.*

Elemento certamente innovativo di questa soluzione è stato quello dell'accesso condiviso al disco, per garantire il quale si è fatto uso di un particolare tipo di *filesystem* denominato *GFS*, *Global File System*. Questo sistema proprietario, distribuito dalla *Sistina Software* [51] , permetteva di poter accedere (sia in lettura che scrittura) agli stessi *file* da qualunque nodo del *cluster* in maniera *trasparente*, ovvero senza modifiche sul *software* dell'emeroteca. Oltre ad aver pienamente risposto alle esigenze di scalabilità dell'Emeroteca, questa soluzione ha mostrato notevoli doti di *robustezza* e di *prestazioni*. L'adozione di un meccanismo di *round-robin* sul DNS permetteva di utilizzare una URL univoca per *identificare* il servizio <periodici.caspur.it>, garantendo un implicito bilanciamento di carico tra i due *server* del *cluster.*

L'indubbio vantaggio introdotto da questa architettura, rispetto a quella basata sul *server* SUN, è stato quello di introdurre, per la prima volta nel servizio di EV, un elevato grado di disponibilità dei dati (e quindi dell'informazione ad essi associata), poiché questi continuavano ad essere fruibili anche in seguito alla caduta di uno dei due nodi del *cluster.* Tale approccio architetturale si è dimostrato negli anni vincente e non è stato di fatto più abbandonato.

Ad oggi infatti il servizio di emeroteca viene erogato da tre *server* ciascuno equipaggiato con un doppio processore Intel Xeon® con frequenza di *clock* pari a 3.40 GHz e 8 GB di RAM; per quanto riguarda i dati su SAN si utilizza un sistema RAID della Infortrend con dischi Ultra-ATA, in

grado di garantire una capacità complessiva di 14 TB; un sistema analogo ma con una capacità complessiva di 12,5 TB è dedicato al *backup*. I collegamenti tra i *server* e la SAN sono garantiti da collegamenti in fibra (*fiber-channell*) da 4 Gbps. Nella figura 18 è riportato un quadro schematico della configurazione corrente.

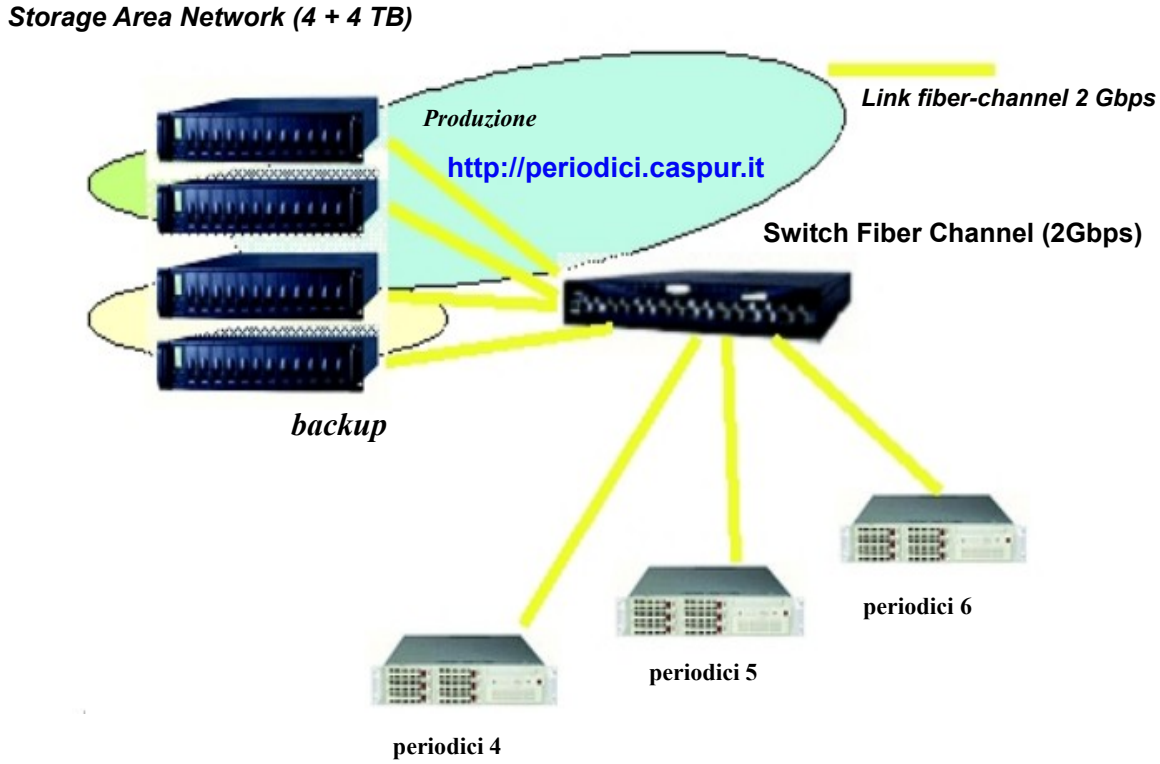


Fig. 18 – Schema dell'attuale infrastruttura del servizio di Emeroteca Virtuale (ottobre 2009)

La figura 19 mostra l'andamento del tasso di crescita della potenza complessiva di calcolo del sistema dedicato al servizio, la RAM complessiva impiegata e le capacità della memoria di massa (su SAN) dedicata ai dati. Dai dati mostrati si può dedurre come nell'arco di un decennio la RAM sia aumentata di un fattore 6, la capacità di calcolo di un fattore 10, mentre la capacità dello *storage* sia aumentata 53 volte, confermandosi come la risorsa più soggetta al cambiamento.

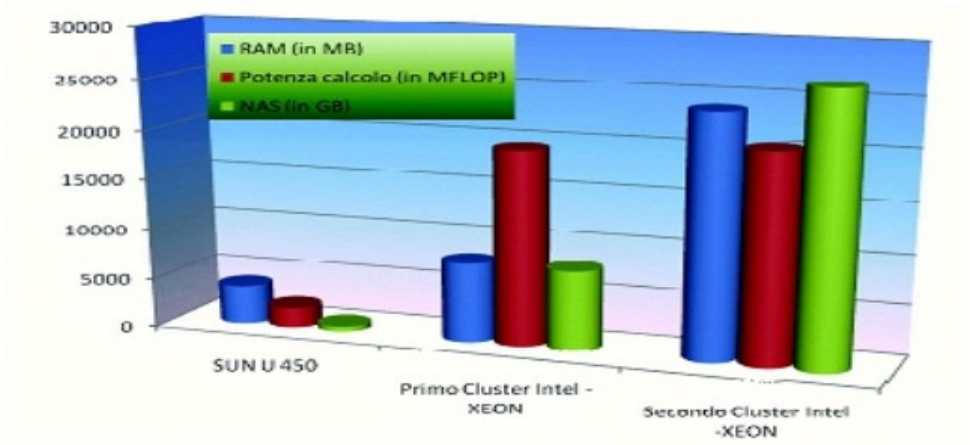


Fig. 19 – Infrastrutture hardware utilizzate per l'EV a confronto (i dati mostrano i parametri della RAM, della potenza di calcolo e della capacità della memoria di massa complessive)

51Rif. <www.sistina.com/products_gfs.htm> (verifica 20 ottobre 2009)

Una nota conclusiva a parte merita il discorso legato al *back-up* dei dati. Come si può

immaginare, tale attività assume per i sistemi di *digital library* qual è quello dell'EV, un aspetto rilevante non solo per ciò che concerne la criticità del servizio, ma soprattutto per la gran mole di informazione immagazzinata all'interno della *repository* dei dati. Una caratteristica comune a tutti i sistemi che offrano accesso *full-text* in locale ad almeno più di un migliaio di testate elettroniche, è che la mole dei dati associati agli articoli disponibili si traduce non solo in una dimensione complessiva della *repository* elevata (> 1.000 GB), ma in un numero complessivo di *file* (di dimensioni comprese tra 1KB ed 1 MB) che può facilmente raggiungere e superare i 10 milioni.

Nello specifico caso dell'EV, il numero attuale di testate elettroniche disponibili *online* supera le 5.300 unità, mentre quasi 8 milioni sono gli articoli disponibili a testo completo (ottobre 2009). La dimensione complessiva dell'archivio dei *digital objects* supera ampiamente i 9 TB (9000 GB), mentre il numero dei *file* associati sfiora i 200 milioni di unità (includendo anche i descrittori XML). Giornalmente, inoltre, vengono caricati 8 *dataset* di 6 diversi editori, per un totale di 2,5 GB distribuiti su circa 30.000 file, ed il processo di caricamento delle riviste di Science Server (*loader*) genera per lo meno altri 30.000 nuovi *file* XML.

Nell'ipotesi di un *back-up* tradizionale su nastro, anche impiegando tecnologie di accesso al disco estremamente efficienti per ciò che riguarda l'I/O (quale quella offerta dal *link fibre-channel* a 4 Gbps ed in un prossimo futuro a 8 Gbps) una copia completa e speculare dell'archivio dell'EV, verrebbe ultimata non prima di qualche settimana nella migliore delle ipotesi, ed escludendo un eventuale *post-processing* di validazione per i dati copiati. È ovvio che durante tale periodo il servizio sarebbe inutilizzabile, costringendo il dirottamento di tutta l'utenza verso gli archivi dei vari editori.

È per questo motivo che è sempre in linea un sistema SAN tecnologicamente identico a quello utilizzato in esercizio (si faccia riferimento agli schemi delle figure 17 e 18). Prima di mettere in esercizio la nuova Emeroteca si è provveduto ad allineare i due sistemi SAN, in modo che risultassero speculari; l'allineamento giornaliero e/o settimanale è garantito da procedure automatiche. In questo modo è possibile garantire all'EV un *fuori servizio* non superiore ad un'ora nel caso in cui il *filesystem* di esercizio dovesse andare incontro ad un grave problema tecnico che compromettesse l'integrità dell'archivio.

Il prossimo futuro

Continua è l'attività di rivisitazione dell'EV, soprattutto in relazione all'esigenza di renderla più rispondente alle aspettative degli utenti, aspettative modulate dai nuovi scenari tecnologici messi periodicamente in campo dalla comunità di Internet. Proprio per rispondere ad una di queste esigenze, ovvero quella di dotare l'emeroteca di strumenti di *social collaboration* propri del cosiddetto Web2.0, nell'agosto del 2008 sono stati resi disponibili una serie di servizi per l'utenza generica e per quella registrata, ovvero: strumenti di *social tagging* di riviste/fascicoli/articoli; *widget* per la ricerca in EV; *feed* RSS per i nuovi fascicoli acquisiti; esportazioni delle sezioni bibliografiche degli articoli in nuovi formati [8].

Inoltre, al fine di rispondere alle esigenze di *ammodernamento* della piattaforma *software* nel 2008 è stato costituito, all'interno del CASPUR, un gruppo di lavoro il cui compito è stato prima quello di definire le aree di intervento sulla struttura del *software* dell'EV sulle quali intervenire e successivamente di prevedere le azioni propedeutiche alla loro sostituzione.

In particolare, si è valutato come necessario un intervento sul *core* del software di *digital library*, ovvero sul motore di ricerca ed indicizzazione degli articoli caricati e sui sistemi di gestione dell'utenza remota. Le aree di intervento sono state:

- le pagine relative al *simple & advanced search*;
- le pagine con i risultati di una ricerca;

le pagine relative alla registrazione di un utente;
le pagine relative all'autenticazione degli "utenti remoti";
le pagine "personali" degli utenti registrati.

Sono quelle, infatti, le sezioni più soggette ad un carico di lavoro maggiore, dovuto al costante aumento e dei contenuti e della base utenti. Come detto nelle precedenti sezioni, l'EV si avvia, infatti, ad avere 10 milioni di articoli caricati ed un numero di utenti remoti che, in un lasso di tempo non troppo lungo, potrà toccare le 10.000 unità. Entrando nel merito poi delle soluzioni scelte per l'implementazione dei nuovi moduli si è deciso di utilizzare in entrambi i casi *software* e linguaggi *open-source*, ovvero: il *software SOLR-Lucene* [52] sviluppato dalla *Apache Software Foundation* [53], per ciò che concerne il motore di ricerca; la base dati *MySQL* [54], e programmi sviluppati in *PHP* [55] per ciò che concerne le sezioni dedicate all'utenza remota [9].

Sono invece state preservate le sezioni dedicate al *software* di caricamento delle riviste, alla struttura dei metadati descrittivi, al *rendering* degli articoli HTML, al *browsing* delle riviste ed alla navigazione all'interno delle pagine statiche. Il perché non si sia ritenuto di dover intervenire (per lo meno in questa fase) anche su queste sezioni, lo si può ricercare nell'efficienza del codice attualmente utilizzato, nella possibilità di poter comunque interagire con tali moduli, in virtù del fatto che sono stati sviluppati in una logica di *software open-source* e, non ultimo, nell'analisi costi-benefici, che mostra come non profittevole un'attività in questo specifico àmbito.
Nella sottostante figura è mostrato uno schema a blocchi del servizio di EV, nel quale sono evidenziate le sezioni sulle quali si è intervenuto.

Fig. 20 – Sezioni del servizio di EV che saranno aggiornate nel nuovo servizio di EV (previsto per il mese di Novembre 2009)

Parallelamente a ciò, si è avviato un potenziamento della struttura *hardware*: il nuovo servizio che entrerà in funzione per la fine del 2009, vedrà l'impiego di un *server* ulteriore, con le medesime caratteristiche dei tre attualmente in produzione. Relativamente alla distribuzione delle funzionalità all'interno del *cluster*, vale quanto mostrato in figura 21.

52 Rif.<lucene.apache.org/solr> (verificato il 20 ottobre 2009).
53 Rif. <www.apache.org> (verificato il 20 ottobre 2009).
54 Rif. <dev.mysql.com> (verificato il 20 ottobre 2009).
55 Rif. <www.php.org>(verificato il 20 ottobre 2009).

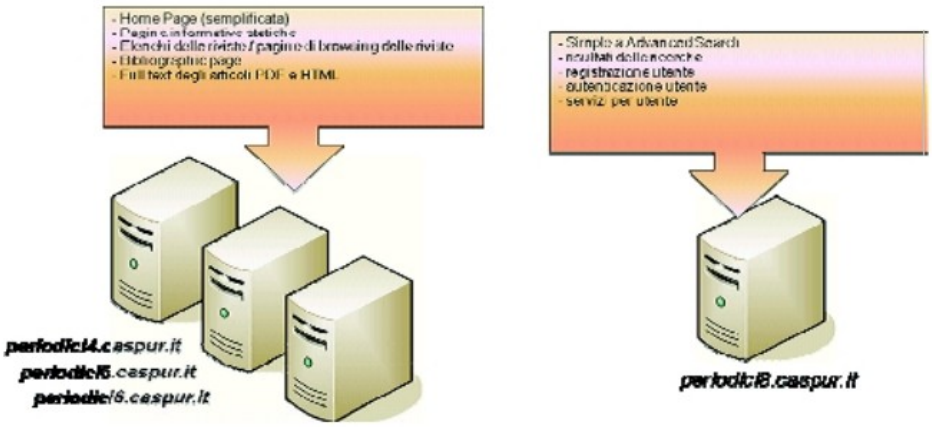


Fig. 21 – Il nuovo cluster dell'EV con le funzionalità associate a ciascun server

Tornando allo schema di figura 20, si può osservare come uno degli elementi, sui quali in futuro maggiormente verterà l'attività di coloro che sono incaricati della gestione tecnica dell'emeroteca, sarà quella relativa alle *policy* di *risk assessment* e *preservazione a medio e lungo termine* dei dati. Relativamente al primo punto il CASPUR ha partecipato nel 2008 ad un corso di *assessment* dei dati di una *digital library* nell'àmbito del progetto Europeo DRAMBORA [10], i cui risultati sono successivamente stati presentati alla comunità CIBER [11]. Si ritiene che quanto appreso in quella occasione e quanto potrà essere in seguito scambiato con la comunità CIBER possa costituire una valida base di partenza per la definizione di una strategia di valutazione dei rischi associati alla qualità dei dati relativi agli articoli memorizzati in emeroteca ed alle metodologie adottabili per il mantenimento di un sufficiente livello di qualità. Analogamente a quanto si è iniziato a fare su questo fronte, verranno studiate soluzioni per affrontare efficacemente il problema della conservazione del posseduto (riviste ed articoli associati) a medio e lungo periodo, problema *straordinario* a causa delle dimensioni del posseduto che si vuole preservare. Si ritiene in questo caso che l'adozione di politiche *inter-consortili*, attraverso il coinvolgimento dell'altro centro che in Italia offre servizi similari, il CILEA, possa essere un modo più efficace per affrontare il problema della preservazione ed in linea con le logiche di ridondanza alla base di una vera politica di conservazione a lungo termine.

Conclusioni

È ormai evidente come questo servizio abbia superato più che positivamente la fase di servizio "sperimentale"; ne è testimonianza l'elemento di attrazione che ha rappresentato per il consorzio CIBER in questi anni e la grande visibilità che ha garantito al CASPUR nel consesso universitario italiano. In conclusione si vuole evidenziare come esso rappresenti, per chi lavora direttamente alla sua gestione o per chi partecipi indirettamente alla sua evoluzione, un impegno costante ed una sfida continua, spesso messa a dura prova dalla sempre più frequente necessità di far quadrare magri bilanci, sempre più ridotti dai costi degli abbonamenti alle riviste elettroniche commerciali. Una diversa distribuzione delle economie che veda lo strumento di diffusione dei contenuti con un peso per lo meno comparabile al costo degli abbonamenti, potrebbe certamente contribuire ad una evoluzione più rapida ed efficace dei servizi, a partire da quello più critico della preservazione.

Bibliografía

(tutti i link sono stati verificati il 21 ottobre 2009)
• Sito di riferimento del consorzio CIBER: <www.uniciber.it>.
• Per una descrizione della struttura interna dell'Emeroteca Virtuale, si rimanda all'articolo: Ugo Contino. "Un portale di accesso a riviste elettroniche multidisciplinari per l'Università e la Ricerca scientifica e

tecnologica: l'esperienza del consorzio CASPUR con il suo servizio di Emeroteca Virtuale". *Convegno "Biblioteche digitali per la ricerca e la didattica: esperienze e prospettive". Parma, La Casa della Musica, 22 novembre 2003* <eprints.rclis.org/785>.
• Per i dati sugli studenti si sono utilizzate le statistiche del Ministero dell'Istruzione, dell'Università e della Ricerca <www.miur.it/UstatNet>, mentre le tabelle sui docenti e ricercatori sono state estratte dal CINECA (fonte <sito.cineca.it/murst-daus/docenti/docenti.shtml>).
• <www.sas.com>.
• <www.projectcounter.org>.
• Ugo Contino. "Il nuovo servizio di produzione di statistiche d'uso compatibili COUNTER dell'Emeroteca Virtuale". *Seminario residenziale CIBER. Villa Pace, Messina. 12-14 giugno 2006* <www.uniciber.it/fileadmin/doc_imm/documenti/SeminarioResidenzialeMessina-Caspur1.ppt%3e.
• Nunzio Femminò, Dario Orselli, Mariella Smedile. "Presentazione del nuovo portale sulle statistiche d'uso per gli atenei del CIBER" *Seminario residenziale CIBER. Villa Pace, Messina. 12-14 giugno 2006* <www.uniciber.it/fileadmin/doc_imm/documenti/villa_pace_nunzio.ppt>.
• Ugo Contino. "Uso di strumenti di "Social Collaboration" nell'Emeroteca Virtuale del CIBER" *Seminario residenziale CIBER. Palazzo Vescovile, Amalfi. 11-13 giugno 2008* <www.uniciber.it/fileadmin/doc_imm/documenti/Amalfi_giugno_2008/EV-e-Web2.0.ppt>.
• Si veda a tal proposito l'articolo: Gino Farinelli, Riccardo Fazio, Ilaria De Marinis, Stefano De Luca. "Il servizio di Emeroteca Virtuale al CASPUR ed il suo nuovo motore di ricerca" *CASPUR Annual Report 2009* <www.caspur.it/Files/annual_report_2009/04-Farinelli_Fazio_DeMarinis_DeLuca.pdf>.
• Informazioni sul progetto DRAMBORA possono essere reperite sul sito istituzionale del progetto: <www.repositoryaudit.eu>.
• Ilaria De Marinis. "Panoramica su DRAMBORA e sugli strumenti di risk assessment dei dati nelle biblioteche digitali. La nostra esperienza per l'Emeroteca Virtuale" *Seminario residenziale CIBER. Palazzo Charamonte-Steri, Palermo. 3-5 giugno 2009* <bib03.caspur.it/ocs/index.php/ciber/pri2009/paper/view/13/6>.