Studio e progettazione di una rete informatica per una biblioteca scientifica basata su tecnologie Open Source

Alessandro Tugnoli, Diego Malipiero, Silvana Mangiaracina
a.tugnoli|diego.malipiero|mangiaracina@area.bo.cnr.it

4 marzo 2010

1 Introduzione

L'accesso all'informazione nell'era digitale è indispensabile per l'attività di ricerca e la competizione scientifica in ambito internazionale e, i servizi offerti dalle biblioteche scientifiche costituiscono quindi uno strumento di lavoro insostituibile per i ricercatori.
Nel corso degli ultimi vent'anni la disponibilità di risorse informative digitali, accessibili ovunque e in qualsiasi momento, ha cambiato le aspettative ed il comportamento degli utenti delle biblioteche: la rapidità e facilità di accesso alle pubblicazioni scientifiche e alle banche dati hanno infatti aumentato in maniera sostanziale il fabbisogno informativo del ricercatore.

Ma qual è la differenza principale per una biblioteca nel rendere disponibile una pubblicazione digitale piuttosto che una cartacea? Il grosso cambiamento risiede nella considerazione che, mentre con l'acquisto di riviste o libri cartacei si diviene materialmente proprietari del materiale, i cui diritti di utilizzo, prestito, riproduzione, etc.. vengono disciplinati dalla legge sul diritto d'autore[1], con la sottoscrizione di una risorsa elettronica ciò che viene pagato è un servizio di accesso al materiale digitale, che il più delle volte risiede sui server dell'editore. In questo caso i diritti e i doveri della biblioteca non sono più regolati dalla legge statale ma diventano diritti e doveri verso l'editore che concede in licenza il servizio. In generale una "licenza" è un contratto che viene accettato da ambedue le parti e che stabilisce degli obblighi contrattuali tra editore e biblioteca (e nel quale ognuna delle parti tende a proteggere i propri interessi o quelli dei propri utenti).
In particolare, nel contratto, oltre alle condizioni economiche, vengono definiti concetti essenziali, quali: chi sono gli utenti autorizzati all'accesso della risorsa, quali sono le modalità di accesso, quali sono gli usi consentiti e quelli vietati.

- Utenti autorizzati: Ci sono due tipologie che vengono usualmente accettate come "utenti autorizzati": la prima comprende gli utenti istituzionali, cioè i dipendenti, gli studenti, gli afferenti dell'ente sottoscrittore (e in genere, tutti coloro che hanno un rapporto di lavoro sistematico con l'ente), la seconda è costituita dai "*walk-in users*" della biblioteca, che sono gli utenti occasionali della biblioteca, che non hanno alcun rapporto con l'ente, ma che si recano fisicamente nella biblioteca per accedere alle sue risorse.

[1] In Italia, per ciò che concerne le pubblicazioni cartacee, le biblioteche godono delle Eccezioni al diritto economico dell'autore che consente loro di effettuare il prestito inter-bibliotecario (ILL, inter-library loan) e la riproduzione, purché in contesti di studio personale o ricerca e, nel caso di riproduzione, purché questa non sia totale: Il prestito eseguito dalle biblioteche e discoteche dello Stato e degli enti pubblici, ai fini esclusivi di promozione culturale e studio personale, non è soggetto ad autorizzazione da parte del titolare del relativo diritto, al quale non è dovuta alcuna remunerazione, cfr. DL n. 68/2003, Capo I, art. 9: nuovo art. 69 comma 1 del Capo V, tit. I, modificati della Legge 22 Aprile 1941, n. 633. La riproduzione di opere per uso personale è consentita nei limiti del 15% di ciascun volume o fascicolo di periodico, escluse le pagine di pubblicità; relativamente alla tecnica di riproduzione e di trasmissione della copia sono consentite riproduzioni effettuate mediante fotocopia, xerocopia o sistema analogo e sono consentiti atti di riproduzione temporanea transitori e accessori e parte integrante ed essenziale di un procedimento tecnologico, eseguiti all'unico scopo di consentire la trasmissione in rete tra terzi, cfr DL 68/2003, Capo I, art. 9: nuovi art. 68 e 68-bis del Capo V, tit. I, modificati della L 633/1941.

- Modalità di accesso: l'accesso alla risorsa deve avvenire attraverso una rete sicura e controllata, alla biblioteca infatti, può essere richiesto in qualunque momento di individuare chi possa commettere un uso non consentito e di inibirlo immediatamente, pena la disconnessione per tutto l'ente.
  Tipicamente vengono registrate e abilitate dall'editore le reti di indirizzi IP dell'ente dalle quali l'uso è consentito (dall'interno di queste reti l'accesso per l'utente è "trasparente", in quanto viene già riconosciuto come un utente appartenente all'istituzione), ma è permesso anche l'utilizzo di *proxy servers* per consentire l'accesso alla risorsa anche da reti non registrate, per esempio da casa, da sedi di lavoro temporaneamente diverse, dall'estero, etc.. Ultimamente, un numero sempre maggiore di editori ha iniziato ad implementare il *framework* Shibboleth[6][7], col grosso vantaggio di poter autorizzare un utente direttamente sul sito dell'editore (evitando quindi l'uso del *proxy* quando l'utente si connette da una rete non registrata), nonché di mettergli a disposizione altri servizi personalizzati, come la possibilità di salvare ricerche o di ricevere avvisi (*alerts*).

- Usi consentiti e vietati: Il download del full text degli articoli o dei libri elettronici cui si accede tramite abbonamento è sottoposto ad alcune regole sempre specificate nella licenza (license agreement) che la biblioteca firma per conto degli utenti e che è tenuta a far rispettare, pena la disattivazione delle singole risorse, prima per gli utenti che non hanno rispettato i vincoli, poi per tutte le reti. É vietato il download, la stampa, la trasmissione, la riproduzione e copia:

  a. sistematici o programmati
  b. indiscriminati
  c. massicci

  intendendo con questi aggettivi il download, stampa etc.. di un intero volume, o fascicolo, o di più fascicoli sequenzialmente. I license agreement prevedono anche l'osservanza di altre due regole. Il full text deve essere destinato:

  - al solo scopo di ricerca e uso personale
  - con divieto assoluto di sfruttamento commerciale

  oltre a divieti su casi specifici che differiscono a seconda del fornitore e che occorre verificare, in caso di bisogno, direttamente sul sito dell'editore[2].

[2]Forniamo qui di seguito alcune clausole del license agreement di Editori di maggior interesse per l'utenza del Consiglio Nazionale delle Ricerche (CNR):
American Chemical Society
http://pubs.acs.org/rates/institutions/terms_conditions.pdf
Individual articles ... are not to be systematically downloaded, re-published in any media, print or electronic form. Individual articles ... may not be downloaded in aggregate quantities or centrally stored for later retrieval. **American Institute of Physics** (e società consorelle)
Systematic or programmatic downloading, printing, transmitting, or copying of the Licensed Material is prohibited. Systematic or Programmatic means downloading, printing, transmitting, or copying activity of which the intent or the effect is to capture, reproduce, or transfer the entire output of a journal volume, a journal issue, or a journal topical section, or sequential or cumulative search results, or collections of abstracts, articles, or tables of contents. Other such systematic or programmatic use of the Licensed Materials that interferes with the access of Authorized Users or that may affect performance of the SCITATION System, for example, the use of 'robots' to index content, or downloading or attempting to download large amounts of material in a short period of time, is prohibited.
American Geophysical Union
http://www.agu.org/pubs/institution_forms/Corporate.pdf
systematically make print or electronic copies of multiple extracts of the Licensed Materials for any purpose, is prohibited;
Knovel
http://www.knovel.com/knovel2/ library_licenseagreement.jsp Knovel Corp. reserves the right to terminate the User's subscription at any time if the User downloads or prints out a substantial portion of Content. The User may not disseminate any portion of Content through electronic means, including mail lists or electronic bulletin boards.
Nature
http://www.nature.com/libraries/site_licenses/
Corp_License_UK2007.pdf (j) make mass, automated or systematic extractions from or hard copy storage of the Licensed Material. **Science**
http://www.sciencemag.org/subscriptions/ inst_terms_unlimited.dtl
Moderate downloading, printing, or saving of material for personal, scholarly, or educational, noncommercial use is permissible, only to the extent consistent with the fair use doctrine. Extensive downloading, printing, or saving of articles by Authorized Users is not permitted.
Springer
"Incidental and non-systematic sharing by Authorized Users with non-authorized individuals of limited amounts of Licensed Material for collaborative research and scholarly purposes, and not for re-transmission, is permitted"

Diviene quindi necessario per una biblioteca scientifica, al fine di garantire alla propria utenza l'accesso alle risorse informative elettroniche nelle modalità più ampie possibili di accesso consentite (connessione da rete fissa, da rete wireless, da reti esterne all'istituzione, etc..) mettere in atto tutto il necessario per garantire la "sicurezza della rete" adempiendo sia alla normativa vigente in materia di accesso alla rete Internet in generale, sia a quanto previsto dalle licenze contrattuali.

In particolare il decreto interministeriale del 16 agosto 2005 [8] stabilisce che i titolari o gestori di un esercizio di qualsiasi specie nel quale sono poste a dispozione del pubblico apparecchi terminali utilizzabili per le comunicazioni telematiche, sono tenuti a:

- adottare misure fisiche o tecnologiche al fine di impedire l'accesso agli apparecchi terminali a persone che non siano preventivamente identificate
- acquisire i dati anagrafici riportati su un documento di identità presentato dall'utente al fine dell'identificazione
- adottare sistemi di monitoraggio delle attività al fine di memorizzare e mantenere i dati relativi alla data ed ora della comunicazione e della tipologia di servizio utilizzato, abbinabili univocamente al terminale utilizzato dall'utente, esclusi comunque i contenuti delle comunicazioni
- i dati acquisiti devono essere raccolti e conservati con modalità informatiche.

2 Tecnologie Open Source per l'implementazione di servizi di rete in una biblioteca scientifica

In informatica, il termine *open source* significa letteralmente sorgente aperto ed indica un *software* i cui autori ne consentono e ne favoriscono il libero apporto di modifiche da parte di altri programmatori della comunità. Chi può modificare il *software*, in che modo, e come diffonderlo, viene regolato dall'applicazione di licenze d'uso (ad esempio la GPL [3]). Grazie ad Internet, la comunità di programmatori *open source* è molto estesa e ormai conta milioni di programmatori geograficamente distanti in grado di coordinarsi e lavorare ad uno stesso progetto raggiungendo livelli di complessità molto più alti rispetto a quanto potrebbe ottenere un singolo gruppo di lavoro.
Ma perchè scegliere un software *open source*? Grazie all'estensione della comunità di programmatori, il software viene aggiornato continuamente, si diffonde rapidamente, è ben testato ed è gratuito !
Avvalersi di software *open source*, soprattutto in ambienti come le biblioteche, consente di abbattere i costi di implementazione di servizi informatici ed offre la possibilità di integrare tra loro diversi software e sistemi in quanto non si è costretti ad utilizzare particolari tecnologie proprietarie.

In un servizio aperto al pubblico, come una biblioteca, realizzare una rete informatica sicura e controllata è una questione di grande importanza in quanto, se ben implementata, permette agli amministratori di sistema di risolvere situazioni critiche, guasti, di prevenire frodi e utilizzi non consentiti nel minor tempo possibile e con maggiore sicurezza.

Realizzare una infrastruttura informatica di questo tipo richiede la presenza di strutture e *server* dedicati all'implementazione di diversi servizi e dal punto di vista *software* la comunità *open source* mette a disposizione tutto il necessario proponendo soluzioni valide e gratuite.
In questo tipo di LAN (*Local Area Network*), ogni utente che vi accede (sia esso remoto o fisicamente presente) ha delle credenziali di accesso (tipicamente *username* e *password*) e può eseguire un numero limitato e controllato di operazioni sui sistemi a lui messi a disposizione.

In particolare, gli utenti che desiderano accedere alle risorse delle biblioteca dovranno possedere un *account* che potrà essere richiesto al personale della biblioteca oppure attraverso una pagina *web* disponibile sul sito della biblioteca. Come vedremo in seguito per memorizzare le credenziali degli utenti si può utilizzare un *server* LDAP[12] di cui ne esiste una implementazione *open source* chiamata OpenLDAP [4]. Gli utenti possono accedere alle risorse della biblioteca anche usando il loro portatile e collegandosi alla rete attraverso un cavo oppure con un collegamento *wireless*. Indipendentemente dal tipo di connessione, dovranno comunque disporre dell'account per accedervi e tali credenziali ver-

ranno richieste (ad esempio tramite l'autenticazione web) al momento della prima richiesta di una pagina internet. Un sistema in grado di gestire questo tipo di problematica è costituito per esempio dalle tecnologie *Captive Portal* e *server Radius*[9], che sono offerte dalla comunità *open source* con prodotti quali *m0n0wall*, *pfSense*[5], *Chillispot*, *WifiDog* e *FreeRadius*[2].

La biblioteca mette a disposizione degli utenti alcuni servizi "federati", ovvero servizi che utilizzano un'autenticazione di tipo SSO (*Single Sign On*), come ad esempio quella fornita dal *framework Shibboleth*, tramite la quale l'utente è in grado di, una volta effettuata l'autenticazione sul sito della propria istituzione, accedere ai differenti servizi senza dover immettere nuovamente le proprie credenziali in quanto viene automaticamente riconosciuto. Shibboleth è un *framework open source* sviluppato dal Middleware Architecture Committee for Education di Internet2 che prevede tre tipologie di sistemi: *Identity Provider* (IdP), *Service Provider* (SP) e *Where Are You From (WAYF) service*. Per un approfondimento sulle architetture di autenticazione e autorizzazione federata per le biblioteche digitali si veda [11].

Tutte le risorse messe a disposizione degli utenti e le loro operazioni devono essere monitorate al fine di evitare frodi ed illeciti. Ciò può essere fatto con l'utilizzo di un *Proxy Server* e di un *Log Server* che possono essere implementati rispettivamente attraverso *Squid* (popolare software libero con funzionalità di *proxy* e *web cache*, rilasciato sotto la GNU *General Public License*) e *Syslog-NG* (una implementazione *Open Source* del protocollo *Syslog* per i sistemi *Unix* ed *Unix-like*).

Al fine di garantire anche la sicurezza dei dati presenti sui diversi sistemi, è possibile avvalersi di un sistema di backup centralizzato tramite il quale vengono copiati i dati provenienti da tutti i server/client di una rete su di un server comune, in modo che, in caso di problemi e/o attacchi (con conseguente perdita di dati) sui vari computer, è possibile recuperare i dati persi.

3 Autenticazione

Come descritto in precedenza una biblioteca scientifica offre servizi di accesso alle informazioni digitali sia per il personale interno che per ospiti esterni.

Dal punto di vista dei servizi informatici e della loro implementazione (in *primis* l'accesso ad Internet), particolare attenzione deve essere rivolta a quest'ultima tipologia di utenza. Il servizio infatti deve essere fruito rispettando le normative vigenti in materia di comunicazioni e deve garantire la sicurezza da attacchi informatici interni ed esterni.

Qualsiasi forma di accesso, ai sensi del decreto interministeriale precedentemente citato, deve essere reso disponibile solo previa identificazione (autenticazione ed autorizzazione) dell'utente.

3.1 Sistemi di gestione delle identità

Quando si ha a che fare con un numero di utenti molto alto e con servizi informatici eterogenei e distribuiti, spesso si tende a creare un numero elevato di *account* su macchine diverse e per scopi diversi, questo rende però la gestione dell'intero sistema nettamente complesso ed insicuro. Un servizio centralizzato per la gestione delle identità (*Identity Management* (IM)) è una *directory* per la gestione delle informazioni degli utenti e costituisce ormai uno standard *de facto* per reti di medie e grandi dimensioni.

Una *directory* è una struttura simile ad un *database*, vi si possono inserire delle informazioni e in seguito recuperarle. In particolare:

- usa un modello gerarchico delle informazioni che sono organizzate in un struttura ad albero
- l'informazione è rappresentata da una *entry*, la quale è composta da attributi che possono avere uno o più valori
- per ogni attributo è possibile definirne un tipo
- il significato delle *entry* è stabilito mediante un attributo particolare: l'*objectclass*
- ogni *entry* è unicamente distinta da un DN (*Distinguished Name*)

Il protocollo di interrogazione più utilizzato è il *Lightweight Directory Access Protocol* (LDAP), RFC 1777 [12].

Come *server* centrale per la gestione di tutti i dati relativi alle utenze ci si può appoggiare ad *OpenLDAP*[4]: un'ottima implementazione *open*

source del protocollo *Lightweight Directory Access Protocol*.

Per l'inserimento, la registrazione e la modifica dei dati personali e delle credenziali di accesso degli utenti è possibile progettare una semplice interfaccia *web*, per esempio in PHP, di facile utilizzo per gli operatori della biblioteca. L'applicativo *web* così creato non solo facilita l'operazione di registrazione degli utenti, ma evita di commettere errori comuni quali creazione di identità con lo stesso *username* o generazione di *password* banali.

Avere una singola istanza dei dati per ciascun utente permette di mantenere coerente su più macchine lo stato degli account operando su un unico *server*. Inoltre, cosa non trascurabile, i *server* LDAP offrono un canale sicuro implementato tramite *Secure Sockets Layer* (SSL) per la comunicazione *client - server*.

3.2 Autenticazione in una local area network ad uso pubblico

Per garantire ai servizi offerti caratteristiche quali facilità e semplicità d'uso, efficienza e risposta alle esigenze dell'utente, si è reso necessario utilizzare sistemi e tecnologie di autenticazione differenti, in quanto si deve essere sempre in grado di tener traccia delle azioni compiute dall'utenza (si noti che non si possono memorizzare in alcun modo e in alcuna forma i contenuti delle comunicazioni, ma solo chi e quando ha stabilito una certa connessione verso un determinato servizio).

In particolare nel nostro caso l'accesso può avvenire in due modi diversi:

a. tramite terminali messi a disposizione dalla biblioteca (*Desktop Computer*), oppure

b. *notebook* o altri *mobile device* di proprietà dell'utente.

3.2.1 Desktop Computer

I vantaggi di offrire l'accesso tramite terminali appartenenti all'ente fornitore del servizio sono molteplici sia in termini di sicurezza che facilità di implementazione. I *computer* messi a disposizione devono essere configurati in modo tale che l'accesso al sistema operativo sia limitato ai soli utenti che forniscono credenziali valide ed autorizzate. Nel caso studio, presentato successivamente ne verrà descritta una implementazione.

3.2.2 Mobile device

Per gli utenti che desiderano utilizzare i propri terminali occorre utilizzare teconologie di autenticazione differenti, dal momento che non si puo' in alcun modo modificare o forzare la configurazione del sistema operativo dei computer di proprietà degli utenti. Lo scenario che si viene a creare è quello tipico del servizio di connessione Internet in ambienti pubblici attraverso l'utilizzo di *hot spot* o *captive portal*. Nasce quindi la necessità di creare un sistema capace di captare tutte le richieste di connessione degli utenti e di consentirle solo previa autenticazione e verifica delle credenziali. Ancora una volta orientandoci in ambito *open source* vogliamo segnalare la soluzione in *pfSense*: un ottimo prodotto che racchiude in un'unica distribuzione *custom* del sistema operativo FreeBSD, servizi di *routing*, *firewalling*, *transparent proxy* e *captive portal*.

L'implementazione di un *transparent proxy* sulla LAN destinata all'utilizzo pubblico nasce dall'esigenza di seguire l'articolo 2 del decreto ministeriale [8] che invita gli enti che forniscono accesso alla Rete a memorizzare e mantenere i dati relativi alla data ed ora della comunicazione e alla tipologia del servizio utilizzato, abbinabili univocamente al terminale utilizzato dall'utente, esclusi comunque i contenuti delle comunicazioni.

3.3 SSO e accesso a servizi federati

Un sistema di identificazione federato permette di realizzare il controllo degli accessi degli utenti (autenticazione, autorizzazione e verifica dei diritti) delegando l'autenticazione a specifiche entità autoritative. Di solito questi sistemi si basano sullo scambio di asserzioni codificate in XML sulla base dello standard Security Assertion Markup Language (SAML) [http://saml.xml.org]. I vantaggi di questo approccio sono:

- si evita agli utenti la scomodità di riautenticarsi quando cambiano applicazione (questo meccanismo prende il nome di Single Sign-On (SSO)

- si riduce significativamente la quantità di credenziali da ricordare e conservare,

- gli utenti delle organizzazioni che appartengono alla federazione possono identificarsi ed accedere a servizi e risorse protette usando le stesse credenziali

La federazione IDEM (IDEntity Management per l'autenticazione federata www.idem.garr.it), coordinata dal consorzio GARR, nasce nel 2009 e si prefigge di promuovere l'infrastruttura per l'autenticazione e l'autorizzazione federata per gli Enti di Formazione Universitaria e di Ricerca in Italia. I partecipanti si suddividono in Membri (Enti fondatori GARR, consorzi interuniversitari per il calcolo, università ed istituzioni culturali e scientifiche con sede in Italia) e Partner (fornitori di risorse e di servizi online: informazioni e contenuti multimediali, piattaforme di contenuti scientifici, es. pubblicazioni e banche dati di editori quali Elsevier, Springer, Thomson Reuters, etc..).
L'obiettivo di IDEM è quello di far accedere gli utenti alle molte risorse che la federazione mette a disposizione registrandosi in un unico punto, questo per l'utente vuol dire possedere un'unica coppia di credenziali per molteplici risorse condivise. Nell'architettura di Shibboleth [inserire stessi rif. Bibliografici gia' citati in precedenza], che è il framework software sul quale si basa IDEM, l'autenticazione degli utenti viene effettuata dall'Identity Provider (IdP), un servizio di autenticazione che risiede presso l'istituzione di appartenenza degli utenti. Allo stesso tempo, i fornitori delle risorse (Service Providers, SP) non avranno più bisogno di gestire onerose procedure di accreditamento e di amministrazione degli utenti. La riservatezza degli utenti è maggiormente preservata perché le credenziali sono conservate in un unico repository, posto all'interno della rete sicura dell'istituzione, solitamente vengono trasmessi identificatori opachi (anonimi) degli utenti e i dati personali sono rilasciati solo in caso di effettiva necessità (e solo se l'utente ne ha espressamente autorizzato il rilascio/consenso informato).

4 Caso studio: la rete informatica della Biblioteca d'Area del CNR di Bologna

Nel prosieguo si presenta come caso studio quello della Biblioteca dell'Area di Ricerca del CNR di Bologna. La Biblioteca d'Area serve una comunità scientifica composta da circa 1.500 utenti (ricercatori, tecnici, dottorandi, contrattisti, assegnisti, laureandi, ...) che operano presso 6 istituti del Consiglio Nazionale delle Ricerche (CNR) e 2 istituti dell'Istituto Nazionale di Astrofisica (INAF), situati sia all'interno di un "campus di ricerca" che presso una ventina di sezioni staccate (site a Catania, Napoli, etc...). La biblioteca offre servizi di accesso alle risorse online (riviste full-text, e-books, banche dati...), servizi di prestito interbibliotecario e document delivery attraverso il sistema NILDE, servizio di accesso remoto alle risorse da postazioni esterne al campus, via proxy server. E' aperta al pubblico, pertanto offre gli stessi servizi, oltre che ai propri utenti istituzionali, anche a utenti esterni (walk-in users). Nel corso degli ultimi anni, il tema della sicurezza nell'accesso alle risorse elettroniche che la biblioteca acquisisce in licenza si è rivelato sempre più pressante: soltanto nel 2009 la biblioteca ha dovuto fronteggiare ben 3 incidenti occorsi nei confronti di editori (download sistematico e massiccio di articoli) e un incidente con la polizia postale (uso illecito della rete). Ci si è posti, con l'obiettivo di trovare una adeguata soluzione, il tema della sicurezza delle credenziali per l'accesso ai servizi, la necessità di semplificare la gestione e di ridurre il numero delle "username e password" date agli utenti, con l'obiettivo di dar loro "una sola" coppia di credenziali, per l'accesso indifferenziato in-campus e off-campus (da casa, in viaggio, da altre istituzioni) sia a servizi interni che a quelli di fornitori esterni, nonché di preservare la riservatezza sia dei loro dati personali che delle loro ricerche. Alla luce di queste esigenze, la topologia della rete informatica della biblioteca è stata interamente riprogettata. Inizialmente la rete era caratterizzata da una topologia molto semplice: vi era una singola rete, dove i *computer* degli uffici, delle segreterie e quelli destinati all'uso del pubblico, condividevano fisicamente lo stesso apparato di rete e la stessa classe di indirizzi IP. Gli indirizzi IP assegnati ai *computer* erano indirizzi

privati e non raggiungibili dall'esterno in quanto vi era un *firewall* che agiva da NAT (*Network Address Translation*) ovvero un sistema che trasforma gli indirizzi IP dei pacchetti in uscita dal *firewall* in un unico indirizzo (tipicamente quello pubblico del firewall stesso). L'accesso ai *desktop computer* (tre computer con sistema operativo Linux Debian) e alla connessione *wi-fi* (garantita da due *Access Point Wireless*), era inoltre limitato semplicemente da credenziali comuni a tutti gli utenti.

In tale configurazione gli amministratori di sistema perdono totalmente la possibilità di identificare le azioni compiute dagli utenti. La rete descritta infatti ha una struttura troppo aperta nella quale non è possibile tener traccia di chi e quando abbia compiuto un illecito o un abuso dei servizi messi a disposizione dalla biblioteca. Al fine di risolvere questi problemi, aumentare la sicurezza della rete e consentire al personale della biblioteca di avere un controllo maggiore sulle azioni compiute è stata ristrutturata la topologia dell'infrastruttura della rete realizzando quanto rappresentato in fig.1.

La nuova rete è stata quindi suddivisa in due sottoreti private ("Rete Uffici" e "Rete Sala") nascoste dall'esterno attraverso il *firewall* che agisce sempre da NAT (4.5) ed ha inoltre la funzione di separare il traffico tra le due sottoreti e garantire la sicurezza dei *computer* connessi da eventuali attacchi esterni. Come si nota da fig.1, nella sottorete "Rete Uffici", oltre ai *computer* del personale, sono stati posti un *server* di *backup*, un *server* di log, un *server* LDAP ed un *server* RADIUS che descriveremo nei paragrafi successivi. Nella sottorete "Rete Sala" invece, oltre ad alcuni *pc-desktop* e *access point*, è presente un *server* NFS il quale offre un servizio di *mounting* remoto delle *home*: ogni utente può in questo modo ritrovare ad ogni nuova sessione di lavoro, documenti ed impostazioni che aveva salvato precedentemente. Infine, tra la "Rete Sala" ed il *firewall* è stato installato un server *pfSense* che svolge il ruolo fondamentale di *Captive Portal* il quale controlla e gestisce tutti gli accessi provenienti dalla rete aperta al pubblico.

4.1 Desktop Computer

I sistemi operativi *Unix like* offrono varie opzioni per l'autenticazione tra cui le tecnologie *Pluggable Authentication Modules* (PAM) e *Name Service Switch* (NSS).

PAM è una tecnologia che rende trasparente il meccanismo di autenticazione alle applicazioni che necessitano di autenticare gli utenti e permette di usare LDAP come meccanismo di autenticazione su qualsiasi sistema supporti PAM senza dover modificare le applicazioni stesse. La gestione del servizio di autenticazione avviene tramite *file* di configurazione (es: /etc/pam.d/*) nei quali vengono dichiarati i moduli che PAM deve prendere in considerazione. Il modulo di interesse è *pam_ldap.so*, che, come si può intuire dal nome, permette l'autenticazione attraverso un servizio LDAP.

Allo stesso modo di PAM, NSS rende indipendenti le applicazioni dai *name service* in modo trasparente, ovvero permette di reperire tutte quelle informazioni che tipicamente sono contenute nei file di sistema *passwd, shadow, groups, hosts, etc...* da un *server* LDAP. Questo è possibile grazie all'utilizzo di alcuni moduli che mappano le chiamate delle librerie C GNU, quali *getpw*, getsh*, etc...* in azioni che dipendono dalla tecnologia del *name service* sottostante che può essere un generico NIS (*Network Information Service*).
In particolare la libreria necessaria ad NSS di nostro interesse è *nss_ldap.so* che permette di recuperare le informazioni relative agli *account* degli utenti dal *server* LDAP. Il file di configurazione di NSS che bisogna quindi modificare è /etc/nsswitch.conf che dovrà essere simile alle seguenti righe:

```
passwd: files ldap
group:  files ldap
shadow: files ldap
```

Questa configurazione indica al sistema operativo che le credenziali dell'utente che sta cercando di effettuare l'autenticazione devono essere cercate prima nel *file system* locale, e in caso di insuccesso, in un *server*, tipicamente remoto, LDAP. Invertire la sequenza di ricerca non è mai consigliato in quanto, se per qualche ragione il *server* LDAP non fosse raggiungibile, non sarebbe più possibile effettuare l'autenticazione sul *personal computer*, neppure con le credenziali dell'utente *root*.

É importante infine sottolineare come la comunicazione tra PAM, NSS ed LDAP avvenga attraverso un canale sicuro grazie al protocollo SSL (*Secure Socket Layer*), entrambe le librerie *pam_ldap.so* e *nss_ldap.so* forniscono infatti autonomamente il supporto a tale protocollo.

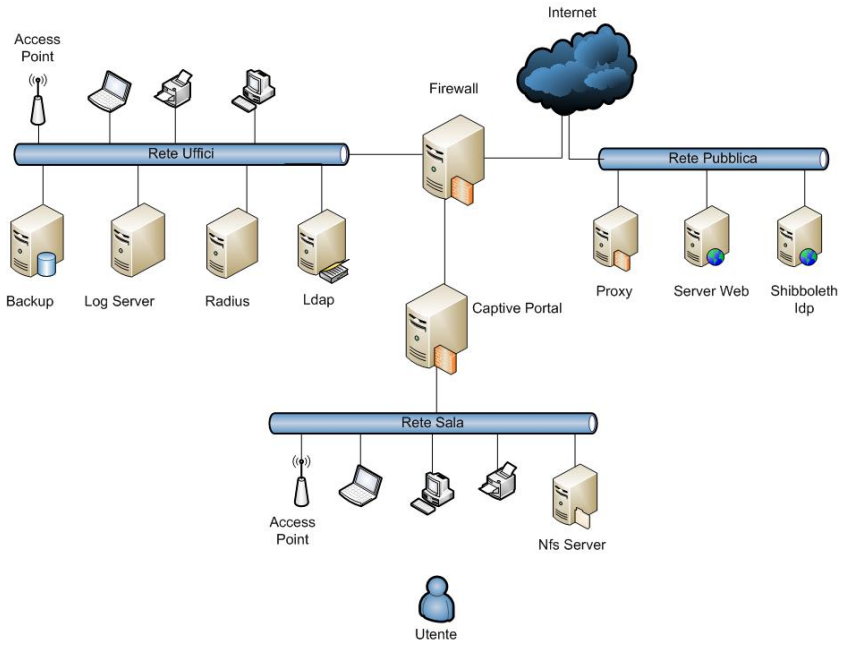

Figura 1: Struttura della nuova rete della biblioteca dell'Area della ricerca CNR di Bologna

4.2 Il captive portal e pfSense

Per consentire l'accesso ai *notebook* ed ai dispositivi mobili di proprietà degli utenti della biblioteca è stato necessario implementare un *Captive Portal* Lo scopo di questa tecnologia è quella di forzare un *client http* connesso ad una rete, a visitare una speciale pagina web nella quale l'utente può leggere i termini e le condizioni di utilizzo dei servizi messi a disposizione dall'ente e dove deve effettuare l'autenticazione tramite l'inserimento della propria *username* e *password* prima di poter accedere alla rete.
Esistono molti software in grado di fornire un servizio di *Captive Portal* ma, nel caso della rete informatica della Biblioteca d'AdR di Bologna si è scelto di utilizzare pfSense.
PfSense è una distribuzione *custom* del sistema operativo FreeBSD sviluppato per essere utilizzato principalmente come *firewall*, *router* e/o *captive portal*, ma non solo, grazie alla sua flessibilità e ad a un sistema di *packaging* automatizzato, permette una facile installazione di una ricca serie di pacchetti software che ne estendono e potenziano le funzionalità di base, oltre ad offrire una comodissima e chiara interfaccia *web* per la gestione della configurazione di tutti i servizi.

Nello specifico la versione di *pfSense* utilizzata è stata la 1.2.3-RC1 che si appoggia su FreeBSD 7.1 ed i servizi aggiuntivi installati sono stati: il *server proxy* Squid e l'analizzatore dei *log* lightSquid, in grado di generare statistiche dettagliate sull'utilizzo del proxy. Tecnicamente il meccanismo di connessione alla "Rete Sala" è totalmente gestito dal *server* pfSense. Esso infatti rappresenta il *gateway* della rete e svolge funzioni di *routing* statico verso il *proxy* Squid. Dapprima rilascia dinamicamente le configurazioni di rete ai vari *client* attraverso un servizio di DHCP, successivamente capta ogni pacchetto e lo dirotta al servizio di *Captive Portal* il quale secondo delle *access list* prestabilite nella configurazione del *firewall*, stabilisce se filtrare o meno la richiesta, ovvero: se presentare una pagina *web* per l'autenticazione o eseguire un *pass-through* della richiesta. Per quanto concerne l'autenticazione degli utenti il sistema offre due possibilità:

- creare una lista di utenti locali
- interrogare un *server* Radius esterno

La prima è sicuramente la soluzione di più facile implementazione ma porta con sè un grosso limite:

rende necessario permettere l'accesso (con privilegi di *Administrator*) alle pagine di configurazione di tutto il sistema anche al personale degli uffici che svolgono il ruolo di registrazione di nuovi utenti. Un'operazione di questo tipo può rilevarsi "pericolosa" sia in termini di sicurezza che di stabilità del *server* stesso.

Per questa ragione, nel nostro caso si è deciso di sfruttare la seconda possibilità, ed abbiamo implementato un *server Radius* il quale non memorizza le credenziali degli utenti ma le ottiene tramite l'interrogazione di un *server* centralizzato LDAP (vedi fig.1) che viene aggiornato dal personale degli uffici tramite una semplice *web application* appositamente costruita.

4.3 Il server log

L'intera rete è monitorata tramite l'utilizzo del protocollo *syslog*. In particolare è stato implementato un *server* con ampio spazio disco dove si è installato Syslog-NG, uno strumento di tracciamento degli eventi di sistema. Esso fornisce una gestione centralizzata dei *log* di tutti i *server* della rete sulla quale agisce indipendentemente dalle piattaforme presenti. Inoltre fornisce diverse caratteristiche aggiuntive, inclusi filtri basati sul contenuto dei messaggi di *log*, gestione personalizzabili della memorizzazione delle informazioni raccolte e diverse capacità di analisi delle stesse. Il *server* ascolta su una porta prestabilita che di default corrisponde alla 514 su protocollo UDP. I *client* sono stati quindi configurati in modo tale da ridirigere tutte le loro informazioni di *logging* verso l'indirizzo iIP del *server log* con destinazione la porta appena citata.
Il vantaggio è quello di avere in un unico *server* la memorizzazione in tempo reale delle attività di tutti i *server* connessi. La configurazione del *server syslog-ng* (nelle distribuzioni Debian) si trova in */etc/syslog-ng/syslog-ng.conf*.

La struttura del *file* è molto semplice e intuitiva, ed è costituita da 5 sezioni: *options*, *source*, *destination* e *filter*.

options

```
options {
    # enable or disable directory creation
    # for destination files
    create_dirs(yes);

    # default owner, group, and permissions
    # for log files
    owner(root);
    group(root);
    perm(0600);

    # default owner, group, and permissions
    # for created directories
    dir_owner(root);
    dir_group(root);
    dir_perm(0700);

    #Disable statistic log messages.
    stats_freq(1);
};
```

come si può notare è possibile definire non solo la possibilità da parte del sistema di creare o meno nuove *directory*, ma anche i permessi con i quali i *file* e le *directory* sono scritte nel *file system*, inoltre è possibile generare in modo automatico delle statistiche sui *log* utili per carpire delle informazioni aggiuntive sul sistema che si sta monitorando.

source

```
source s_all {
        ...
    udp();
            ...
};
```

in questa sezione vengono definite tutte le sorgenti che generano *logs*. Nel nostro caso abbiamo abilitato come sorgente la porta udp in quanto il *server* ha la sola funzione di tener traccia delle operazioni compiute da *client* e *server* remoti.

destination

```
        ...
# some standard log files
destination d_apache_err {
        ...
```

con la parola chiave *destination* si esplicitano le destinazioni in cui andranno memorizzati i *log* captati. Syslog-ng offre la possibilità di definire destinazioni quali *file* oppure *database* esterni tramite l'utilizzo di *pipe*.

filter

```
filter f_apache_err { facility(local7)
and not match (slapd)
and not match("SQUID")
and not match("dhcpd");
};
```

con la parola chiave *filter* si definiscono un insieme di filtri necessari per catturare specificatamente i *log* di interesse. Dall'esempio riportato sopra si

può notare come sia possibile costruire filtri anche complessi tramite l'utilizzo degli operatori logici AND, OR e NOT. In particolare, nell'esempio vengono filtrati tutti i *log* con *facility* local7 [10] che contengono le parole "slapd", "SQUID", "dhcpd".

log

```
#Apache
log {
source(s_all);
filter (f_apache);
destination(d_apache);
};

log {
source(s_all);
filter (f_apache_err);
destination(d_apache_err);
};
```

con la parola chiave *log* si definisce formalmente una regola di log, mettendo in relazione gli attributi *source*, *filter* e *destination*. Ovvero si definisce una regola dove si dichiara da quale sorgente il log deve provenire, attraverso quale filtro deve passare e dove deve essere memorizzato.

4.4 Il server di backup

Il *server* di *backup* è stato implementato installando e configurando il *software open source BackupPC*[1] disponibile anche in pacchetto per la distribuzione Debian utilizzata per il *server*. La gestione dei *backup* può essere effettuata attraverso una comoda interfaccia *web* che permette la configurazione degli *host* di cui effettuare il *backup* e di definirne le modalità di accesso. *BackupPC* è in grado di copiare i dati da *computer Windows*, *Linux* e anche *Macintosh* in quanto supporta i protocolli *samba* ed *rsync*. Consente inoltre di eseguire *backup* incrementali e, tramite *rsync*, è possibile ottimizzare il tempo necessario alla copia dei dati in quanto diminuisce il traffico di rete effettuando una vera sincronizzazione copiando solo i *file* che sono stati modificati. Inoltre, per ragioni di sicurezza, si è scelto di configurare *rsync* in modo che venisse eseguito (tramite *ssh*) da un utente con bassi privilegi (e quindi abilitando *sudo* per questo utente).

4.5 Il firewall

Il *firewall* è costituito da un *server* con installato il sistema operativo Debian con alcuni pacchetti tra cui *iptables* per l'implementazione delle regole di *routing* e di filtro e il pacchetto dhcp3 per il servizio di DHCP (*Dynamic Host Configuration Protocol*). Il *firewall*, oltre a garantire la sicurezza della rete interna è in grado di nascondere entrambe le sottoreti dalla rete esterna facendo NAT, permettere l'accesso al *server* di backup e al *server* di log da parte degli altri *server*, permettere l'accesso alle stampanti dai computer degli uffici (per quelli che si trovano su reti diverse) e fornire indirizzi IP nella rete degli uffici tramite il *server* DHCP.

Il *firewall* dispone di tre interfacce di rete: una viene utilizzata per l'accesso alla rete WAN (Internet), una per la rete degli uffici ed infine un'altra per la rete della sala. Si è scelto di utilizzare tre schede di rete appositamente per separare fisicamente il traffico di queste reti per motivi di sicurezza. Inoltre, dato che per l'interfaccia esterna (WAN) passano i pacchetti provenienti dalla rete della sala e degli uffici, si è scelto di differenziarle effettuando un *postrouting* dei pacchetti cambiando l'IP del mittente (nell'intestazione del pacchetto TCP) assegnandogli un indirizzo IP che indicasse la rete di provenienza. Per fare ciò e assegnare quindi due indirizzi IP alla stessa scheda di rete, è stato necessario utilizzare lo stratagemma dell'*ip-alias*.

Le regole per il *postrouting* sono del tipo:

```
iptables -t nat -A POSTROUTING -s $RETE_UFFICI
        -o $EXT -j SNAT
        --to $IP_NAT1

iptables -t nat -A POSTROUTING -s $RETE_SALA
        -o $EXT -j SNAT
        --to $IP_NAT2
```

dove $EXT è la scheda di rete da cui escono i pacchetti verso Internet e $IP_NAT1 e $IP_NAT2 sono 2 IP pubblici associati a tale scheda.

Per quanto riguarda l'accesso ai *server* dietro firewall da parte dei *server* pubblici per alcuni servizi (come il Syslog e come l'accesso all'LDAP), è stato necessario eseguire il *port forwarding*, un meccanismo che consente di assegnare una porta sul *firewall* per ridirigere il traffico verso un *server* che ascolta su una determinata porta. Un esempio di questo meccanismo è proprio l'accesso al server di log (che si trova dietro firewall, nella rete degli uffici), in cui *syslog* (è un demone che ascolta sulla porta 514 UDP) deve essere contattato dai *server* fuori firewall. Per fare questo è sufficiente inserire in *iptables* queste regole:

```
iptables -t nat -A PREROUTING -p udp -i $EXT
```

```
        -d $IP_NAT1 --dport $PORTAVIRT
        -j DNAT --to $IPSERVER:514
```

(in tale regola viene mappata la porta $PORTAVIRT con la porta 514 sul logserver $IPSERVER)

```
iptables -A FORWARD -i $EXT -d $IPSERVER
        -p udp --dport 514
        -o $ETH_UFFICI -j ACCEPT
```

(la seconda semplicemente fa il forward dei pacchetti verso il logserver che ascolta sulla 514)

Nella topologia rappresentata in fig.1 si presenta il problema di come effettuare il *routing* dei pacchetti destinati alla "Rete Sala", dato che il *firewall* conosce solo la rete in cui si trova il *Captive Portal*. Per risolvere tale problema è stata definita una regola di *routing* statica che indica al firewall su quale interfaccia di rete mandare i pacchetti destinati alla rete che si trova "dietro" al *Captive Portal* e di cui non è immediatamente a conoscenza. La regola è stata creata tramite il comando *"route"* in questo modo:

```
route add -net $RETE netmask $MASK gw $GATEWAY
```

dove $GATEWAY è l'IP del *gateway* cui inviare i pacchetti destinati alla rete $RETE

Analogalmente a quanto eseguito per il *Captive Portal* sono state aggiunte altre regole di *routing* che permettono di:

- effettuare il *routing* dei pacchetti provenienti dall'esterno verso la "Rete Uffici" qualora l'indirizzo di destinazione contenga un IP in questa rete;
- effettuare il *routing* dei pacchetti provenienti dall'esterno verso il *Captive Portal* qualora l'indirizzo di destinazione contenga un IP in questa rete

Inoltre, sono state aggiunte alcune regole di forward dei pacchetti del tipo

```
iptables -A FORWARD -i $ETH_SALA -s $IP_CAPTIVE
        -d $IP_RADIUS -p udp
        -m multiport --dports 1812,1813
        -o $ETH_UFFICI -j ACCEPT
```

per effettuare il *routing* dei pacchetti provenienti dal *Captive Portal* verso il RADIUS e la regola

```
iptables -A FORWARD -i $ETH_SALA -s $PC_PUBBLICO1
        -d $IP_LDAP -p tcp
        -m multiport --dport 636,389
        -o $ETH_UFFICI -j ACCEPT
```

per effettuare il routing dei pacchetti provenienti dai pc della biblioteca verso l'LDAP

Riferimenti bibliografici

[1] Backuppc - (http://backuppc.sourceforge.net).

[2] freeradius - (http://freeradius.org).

[3] Gnu general public license - (http://www.gnu.org/copyleft/gpl.html).

[4] Openldap software - (http://www.openldap.org).

[5] pfsense - (http://www.pfsense.org).

[6] Shibboleth enable service providers. https://wiki.internet2.edu/confluence/display/seas.

[7] Shibboleth. http://shibboleth.internet2.edu.

[8] Stanca Pisanu, Landolfi. Decreto ministeriale interno del 16 agosto 2005 - (http://www.comunicazioni.it/binary/min_comunicazioni/normativa/di_2005_08_16.pdf), 2005.

[9] C. Rigney, S. Willens, A. Rubens, and W. Simpson. Remote authentication dial in user service (radius), 2000.

[10] Gene Spafford Simson Garfinkel. *Practical Unix & Internet Security Second Edition.* O'Reilly & Associates, Inc. 1996.

[11] Giacomo Tenaglia. Studio e realizzazione di una infrastruttura di autenticazione e autorizzazione (aai) italiana per servizi di biblioteca digitale. Master's thesis, Dipartimento di Informatica, Alma Mater Studiorum - Università di Bologna, 2007.

[12] W. Yeong, T. Howes, and S. Kille. Lightweight directory access protocol, 1995.