Gli identificativi persistenti
A Torino un seminario di presentazione
di Oriana Bozzarelli – Maria Cassella

La nuova sede della Biblioteca civica "Italo Calvino" ha ospitato a Torino un seminario di presentazione sugli identificativi persistenti per le risorse digitali, organizzato dalle Biblioteche Civiche di Torino e dal progetto BESS,[1] il 13 aprile 2011.
L'incontro, introdotto da Paolo Messina (Biblioteche Civiche della Città di Torino) e Tommaso Garosci (IRES Piemonte e BESS), si è posto l'obiettivo di illustrare il prototipo di un *network* italiano per la realizzazione di un registro nazionale di risorse digitali e per l'assegnazione, la gestione, la conservazione e la risoluzione di un identificativo persistente (PI) basato sullo standard aperto National Bibliography Number (NBN).
La progettazione dell'infrastruttura tecnologica e organizzativo-gestionale dello standard NBN , oggetto di uno specifico protocollo di intesa siglato nel 2009 tra MIBAC e CNR[2], ha visto coinvolti in un comitato tecnico congiunto anche la Biblioteca nazionale centrale di Firenze, la Fondazione Rinascimento Digitale e l'Università degli studi di Milano[3].
Il seminario torinese è stato condotto da Roberto Puccinelli (CNR di Roma) che ha presentato i risultati della prima fase sperimentale, appena conclusasi, su NBN, proponendo preliminarmente una carrellata, sintetica ma esaustiva, su stato dell'arte e prospettive in materia di identificativi persistenti per le biblioteche digitali e delineando una classificazione operativa, in termini di sistema di risoluzione utilizzato, possibilità e limiti degli stessi.
L'attuale contesto nel quale si registra un'esplosione dei contenuti informativi in rete, una proliferazione quasi incontrollabile di esperienze internazionali e nazionali di digitalizzazione del patrimonio librario e documentale, una crescente massa critica della letteratura di ricerca ad eccesso aperto offre grandi opportunità per la disseminazione della comunicazione scientifica ma, al contempo, porta con sé numerosi problemi relativi alla certificazione e all'identificazione delle risorse digitali nonché alla loro conservazione. In questo scenario assolutamente fluido e in rapidissima evoluzione aprono nuove e promettenti prospettive gli identificativi persistenti ovvero quei codici stabili che, associati alle risorse digitali, ne permettono l'identificazione univoca, la localizzazione e la conservazione a lungo termine oltre a garantirne provenienza, autorevolezza, certificazione dei contenuti e dei diritti (e forse contenerne la diseconomica duplicazione, riuscendo anche a riconoscere e "tracciare" le diverse versioni di un documento digitale).
L'identificazione univoca, peraltro, riveste un ruolo di primo piano nell'accesso alle risorse elettroniche e "costituisce una componente essenziale dell'economia politica dell'informazione, dove il controllo sui mezzi tecnici di accesso alle risorse è vitale quasi quanto quello sulle risorse stesse"[4].
Gli identificatori persistenti sono stringhe numeriche o alfanumeriche di caratteri che, secondo protocolli convenzionali e specifiche modalità di applicazione, vengono associate in modo permanente ad un oggetto digitale al fine di assicurare l'accesso stabile alle risorse e alle informazioni sulle risorse (metadati), riuscendo a garantirne l'unicità e favorendo l'interoperabilità tra sistemi informativi diversi; la persistenza, legata ad un infrastruttura basata su *policy* vincolanti e su regole condivise prima ancora che alla messa in campo di particolari tecnologie, sostanzia gli identificatori persistenti come "fattori abilitanti di natura organizzativa", preziosi in moltissimi ambiti.
"Gli identificatori sono strumenti estremamente potenti all'interno del circuito della comunicazione e tra comunità differenti"[5]. Esistono vari sistemi per l'identificazione persistente delle risorse digitali rispondenti ai bisogni delle diverse comunità ed è difficile prevedere in futuro un'unificazione degli standard. Probabilmente occorrerà sempre convivere con una compresenza di standard diversi - soluzioni *open source* e soluzioni commerciali - dai costi più o meno alti.
Gli identificatori fanno parte di una "grande famiglia" articolabile in una molteplicità di declinazioni e "gradi" di persistenza; operativamente si possono suddividere in due macro-categorie in relazione al sistema di risoluzione usato: gli identificatori che si appoggiano al *Domain Name System (*DNS) e quelli che, diversamente, fanno riferimento a sistemi alternativi al DNS, ad esempio l'*Handle System* (HS).
Avvicinandosi ai numerosi codici di identificazione utilizzati in ambiente digitale uno dei primi acronimi

che si incontrano è l'URI, acronimo appunto di *Uniform Resource Identifier;* si tratta di una stringa che, utilizzando una specifica sintassi[6] organizzata gerarchicamente e relativa all'uso di differenti protocolli di comunicazione, identifica univocamente una risorsa (pagina web, immagine, e-mail, etc.) ma, non necessariamente, fornisce l'accesso alla risorsa stessa. Lo standard URI può essere distinto in due sottoinsiemi: URL e URN.
Senza dubbio l'URL[7] (*Uniform Locator Resource*) è lo strumento di identificazione più conosciuto; i comuni indirizzi web, basati sul *name system* del DNS, sono URL con cui si comunica al browser dove trovare uno specifico oggetto digitale nella rete. Nello specifico l'URL identifica una risorsa elettronica non sulla base del suo nome o dei suoi attributi bensì tramite il suo meccanismo di accesso primario o la sua localizzazione nella rete. La definizione fornita da Tim Berners-Lee "URLs are used to `locate' resources, by providing an abstract identification of the resource location"[8] ne esplicita la problematica commistione tra funzioni di localizzazione e di identificazione.
L'instabilità strutturale dei link, dovuta a cambiamenti per riconfigurare l'hardware con conseguente possibile indisponibilità o variazione dei nomi di dominio e degli oggetti ad essi collegati (riallocazione), manifesta, tuttavia, l'inaffidabilità in termini di persistenza delle URL e profila il rischio concreto di non riuscire a recuperare a lungo termine i documenti, rivelando la natura più autentica delle URL, semplici *locators* non assimilabili a PI veri e propri. In rete è frequente, infatti, il famigerato messaggio di errore "http 404 file not found" con il quale si presentano URL a cui non corrisponde più nessuna risorsa.
L'URN (*Uniform Resource Name*)[9], ideato anche per semplificare la mappatura di altri domini, consente di identificare un oggetto digitale tramite un nome - tratto da un definito dominio di nomi (*namespace*) - in modo univoco e persistente, indipendentemente dalla sua localizzazione fisica; ad esempio l'URN urn:isbn:0-498-47541-1 individua univocamente un libro tramite il suo nome 0-498-47541-1 nel *namespace* dei codici ISBN, ma non indica dove possiamo reperirne una copia. Per gli URN, a differenza degli URL che fanno riferimento al DNS, non esiste una infrastruttura globale per la loro risoluzione. Ragion per cui, se si desidera implementare un sistema di identificazione basato su URN, occorre anche sviluppare una infrastruttura apposita che funga da registro e da risolutore.
Il PURL[10] (*Persistent Uniform Resource Locators*), invece, è un identificativo basato sul DNS e sviluppato da OCLC; identifica un oggetto digitale tramite un normale indirizzo web, alla stregua di un URL, ma non localizza la risorsa direttamente sulla rete bensì si serve in maniera sistematica di un servizio di *redirect*, un *proxy* che svolge il ruolo di risolutore intermedio per reindirizzare la richiesta della risorsa. Il PURL, operativamente poco più che un *proxy*, rimane stabile nel tempo perché viene mantenuta persistente da un amministratore l'associazione tra i suoi codici identificativi e le URL relative, così come l'aggiornamento. Per attivare un PURL occorre registrarsi via web ad un *resolver* dedicato e rispettare alcune pre-condizioni come, ad esempio, possedere già il dominio *top level* che si vuole persistente per le proprie risorse. Il limite del PURL è che mal si adatta ad identificare singole parti di oggetti digitali: ad esempio il caso dei capitoli di un libro per i quali è necessario definire un preciso contesto di riferimento per fa si che assumano un significato.
Anche l'ARK[11] (*Archival Resource Key*), sviluppato dalla US National Library of Medicine è un PI legato ai meccanismi del DNS ma, rispetto al PURL, offre qualcosa di più. Nella sua definizione sono stati considerati aspetti procedurali ed organizzativi piuttosto che elementi tecnologici, seguendo il concetto base che la persistenza sia essenzialmente materia di "servizio". La sua caratteristica principale è quella di connettere tre elementi: l'oggetto digitale, i metadati (in questo caso il set METS) ad esso associati e le modalità di conservazione o per meglio dire l'impegno del gestore dell'identificativo a mantenerlo persistente. E', però, dipendente dal DNS e dal protocollo HTTP per la risoluzione, ragion per cui qualora questa infrastruttura di base a livello di standard dovesse cambiare, l'ARK dovrebbe di pari passo adeguarsi.
La Biblioteca nazionale francese utilizza l'ARK per fornire accesso stabile ai documenti digitali di Gallica e alle notizie dell'opac Bn-Opal plus.
Per superare alcune criticità relative agli identificatori sopra esposti sono stati messi a punto identificativi e architetture di risoluzione che seguono strade alternative al DNS.
Uno degli esponenti più rappresentativi di questo approccio alternativo è l'*Handle System*[12] un'infrastruttura sviluppata dalla *Corporation for National Research Initiatives* (CNRI) concepita allo scopo

di offrire servizi digitali tra i quali l'assegnazione di identificativi persistenti. L'HS si serve di uno specifico protocollo (*Handle System Protocol*), per creare, conservare e garantire l'accesso sicuro ad un database distribuito che associa in modo permanente ed univoco gli identificativi agli oggetti digitali. Da un punto di vista organizzativo l'identificativo basato sull'HS si struttura in un modello gerarchico distribuito: al *top level* del sistema è collocato *il Global Handle Registry* (GHR), gestore unico di tutte le unità amministrative (*naming authority)* deputate alla creazione, assegnazione e amministrazione degli identificativi; ai livelli inferiori troviamo i singoli *Local Handle Services* (LHS) dipendenti da una specifica *naming authority*.
L'implementazione più importante del meccanismo di risoluzione basato sull'HS è il DOI[13] (*Digital Object Identifier*) un identificativo persistente gestito dall' *International DOI Foundation* un consorzio *no profit* comprendente sia partner commerciali, i maggiori editori scientifici ad esempio, che non commerciali. L'assegnazione dei DOI è tuttavia demandata ad una serie di agenzie di registrazione. Negli Stati Uniti l'agenzia CrossRef[14] assegna gli identificativi per gli articoli scientifici, la Content Directions Inc., invece, assegna il DOI a testi, immagini fotografiche, audiovisivi e documenti sonori, software e databases ecc. In Europa mEDRA[15] (Multilingual European DOI Registration Agency) è l'agenzia di registrazione del DOI. Nata all'interno del programma eContent della Commissione Europea mEDRA è stata nominata ufficialmente dall'International DOI Foundation nel 2003. Nonostante il fatto che l'*International DOI Foundation* sia un'organizzazione *no profit* il DOI viene assegnato a pagamento, ovviamente per sostenere i costi necessari alla gestione del sistema e per garantire la persistenza.
In Italia il gruppo di cooperazione poco sopra citato, nato in seguito all'accordo siglato con il MIBAC, ha concentrato la sua attenzione sullo standard internazionale aperto National Bibliography Number (NBN), promosso dalla Conference of Directors of National Libraries (CDNL) e dalla Conference of European National Librarians (CENL), e sulla creazione di un'infrastruttura - funzionalmente analoga a quella esistente per il DNS e gli URL - e di un software (jNBN) dedicato alla assegnazione/registrazione e risoluzione di PI.
La scelta di NBN, basato sullo standard URN, non é casuale ed è dettata dalla convinzione che sia necessario mettere a punto per le risorse digitali un meccanismo di identificazione persistente ma indipendente dalla localizzazione della risorsa; i metadati associati alla risorsa possono contenere una o più URL come supporto alla localizzazione. Mantenendo aggiornata la lista degli URL assegnati a quella risorsa e in presenza di un'infrastruttura che svolge la funzione di risolutore (cioè mantiene stabile l'associazione tra PI e URL), è possibile avere a disposizione oggetti digitali persistenti e URL sempre "reperibili".
La sintassi di NBN, associabile a qualsiasi tipo di risorsa digitale, è esplicitata in un *request for comment*[16] - scritto dal finlandese J. Hakala all'interno della *Internet Engineering Task Force* (IETF) - che definisce l'utilizzo degli URN per codificare degli NBN. Per questi identificativi è stato registrato il *namespace* URN:NBN. Il registro internazionale di tutti i domini NBN è tenuto dalla Library of Congress; essi sono organizzati in maniera gerarchica (da sx verso dx): il *top level* indica l'identificativo del paese, a cui seguono sotto-domini con un numero virtualmente infinito di livelli. Lo standard ha la peculiarità di attribuire la responsabilità del dominio di più alto livello per ogni paese alla Biblioteca nazionale centrale. Questa "centralizzazione" costituisce, però, anche una delle criticità dello standard e probabilmente è uno dei motivi al quale può essere ricondotta la sua scarsa penetrazione a livello internazionale. Esemplificativo è il caso tedesco. In Germania[17] NBN è notevolmente diffuso e la sua gestione è appunto affidata in via esclusiva alla Biblioteca nazionale centrale (sede di Francoforte) che ha già assegnato oltre un milione di PI. Il volume e il continuo incremento dei dati processati ha determinato un progressivo scadimento in termini di performance del server dedicato e l'insorgere di problemi gestionali tali da rendere necessario un non trascurabile aggiornamento del software.
L'attività della "cordata" italiana, utilizzando unicamente standard aperti per garantire la mappabilità ed l'interoperabilità con altri sistemi, ha portato alla creazione di una innovativa architettura distribuita di sistema, di un software dedicato e all'avvio di un *test-bed,* ora concluso. Facendo propri concetti presenti nel DNS e permettendo l'assegnazione decentrata dei PI, NBN ha introdotto nel sistema una certa scalabilità, garantendogli una crescita potenzialmente illimitata. Sotto il profilo organizzativo, ritenendo

necessario progettare un sistema nel quale fosse fondamentale la ripartizione dei flussi di lavoro e dei costi, è stata creata un'architettura articolata in "agenzie" con diversi livelli di responsabilità secondo un modello gerarchico distribuito, ovviamente nel pieno rispetto di *policy* condivise.
Sono previsti domini multilivello, come nell'esempio seguente:

I° LIVELLO	II° LIVELLO	III° LIVELLO	ID PROGRESSIVO
IT:	UR:	CNR:	12345

L'agenzia di *top level,* rappresentata da strutture afferenti al MIBAC[18], si occupa del dominio nazionale attribuito in maniera esclusiva (NBN:IT per l'Italia), l'agenzia di secondo livello (*inner node*) è responsabile di domini relativi a specifici settori culturali (UR per "Università e ricerca"), all'agenzia di terzo livello (*leaf node*) pertiene il dominio della singola istituzione che attribuisce i PI NBN ai propri *e-content*, e così via. Con questo approccio il carico di lavoro per l'assegnazione degli identificativi, la verifica e la gestione delle richieste di risoluzione si decentralizza distribuendosi su più nodi, su più server e i costi di gestione si spalmano su più organizzazioni. Man mano che l'infrastruttura cresce e nuove organizzazioni si aggregano, alcune, tra queste, possono assumere il ruolo di nuove "agenzie" a cui è demandato il compito operativo della gestione del server e del relativo personale informatico-bibliotecario, con una conseguente ri-distribuzione su una molteplicità di soggetti degli oneri. In questo contesto il PI diventa un "percorso a dominio" che riflette l'organizzazione dalla quale proviene. La persistenza e la certificazione delle risorse "dipendono" dalla credibilità ed affidabilità dell'istituzione che gestisce quel determinato *namespace*, oltre che dalla qualità delle *policy* adottate.
Dal punto di vista del funzionamento dell'infrastruttura è solo il "livello foglia" (*leaf node*), quello più granulare, che - tramite il protocollo aperto OAI-PMH - si interfaccia direttamente con le *digital libraries* sotto il suo controllo, effettua periodicamente un *harvesting* di metadati per monitorare quali nuove risorse digitali sono state inserite (a partire dall'ultima operazione di *harvesting*) e produce i PI da associare agli oggetti digitali. Il "livello foglia", a sua volta, è oggetto di *harvesting* dal nodo intermedio e quest'ultimo viene successivamente "harvestato" dal *top level node*; un *back-up* globale di tutti i nomi generati dai *sub-domain* e degli oggetti digitali viene comunque eseguito dal nodo di più alto livello. I livelli superiori operano anche verifiche formali di aderenza alla policy. Dopo un'assegnazione temporanea di identificativi[19] alle risorse, a conclusione dell'intera procedura, gli identificativi vengono confermati, vengono rese permanenti le loro assegnazioni e, conseguentemente, risultano registrati e risolvibili. Ciascuna agenzia stabilisce le policy del suo sotto-dominio, adeguandosi alle policy del livello superiore.
Particolarmente interessante è il problema del riconoscimento delle risorse digitali identiche e di quelle fortemente simili. La duplicazione è monitorata tramite controlli *bitwise*[20] che permettono di effettuare confronti a partire dal contenuto di bit della risorsa ed appurare se esistono oggetti digitali identici già dotati di PI. Decisamente di non facile soluzione, invece, è la questione del riconoscimento di oggetti digitali molto simili ma non identici. Esistono algoritmi *near book duplicate* - sinora non implementati nel prototipo presentato dal CNR - in grado distinguere risorse fortemente simili, con ottimi risultati (90-95%). Puccinelli sottolinea che un idoneo meccanismo di identificazione persistente garantisce un collegamento tra le varie versioni digitali di una stessa risorsa, permettendo al tempo stesso di coglierne la diversità.
jNBN[21] è il software sviluppato dall'Ufficio Sistemi Informativi del CNR con la collaborazione della Fondazione Rinascimento Digitale[22] a supporto dell'infrastruttura distribuita italiana per risolvere identificativi NBN; può essere installato e configurato ad hoc su qualsiasi nodo dell'architettura. Il software, scritto in linguaggio *java*, viene rilasciato sotto la European Union Public Licence (EUPL), la licenza europea per software libero[23]. L'interfaccia prototipale web presenta un box di ricerca *Google style*, dove inserendo l'identificativo NBN si ottengono la risorsa e/o i metadati descrittivi associati alla risorsa (questi ultimi in XML nativo o formattati in Dublin Core); esiste un profilo di amministratore (dedicato alle attività sistemistiche di configurazione, creazione degli account per gli utenti, calendarizzazione delle operazioni di *harvesting,* etc.) ed un profilo riservato all'operatore di biblioteca.
A livello italiano è stata condotta una sperimentazione con l'installazione di nodi presso le Biblioteca nazionali di Firenze e di Roma, la Fondazione Rinascimento Digitale e il CNR; prove di *harvesting* sono

state effettuate su *digital libraries* nell'area di Pisa e Potenza, così come prove di assegnazione di PI a risorse digitali e analisi di pre-fattibilità per verificare quali istituzioni intendessero adottare questo standard. La fase sperimentale, afferma Roberto Puccinelli, adesso è conclusa e il futuro è tutto da definire. Il prototipo, infrastruttura e software, è disponibile. Anzi, l'ingresso di nuovi interlocutori presso i quali attivare nuovi nodi è fortemente auspicabile perché permetterebbe di verificare ulteriormente sia la scalabilità sia la validità delle policy adottate nel sistema. I requisiti necessari alle istituzioni per contribuire all'implementazione dell'infrastruttura, oltre all'adozione di soluzioni *open source,* sono la disponibilità di un server sul quale configurare il "nodo" e di personale, informatico e bibliotecario, dedicato.
La discussione ed il confronto sono aperti.

Le *slides* relative al seminario sono accessibili *online* sul sito del progetto BESS.

---

[1] BESS è un'iniziativa di acquisizione di risorse elettroniche che coinvolge diciotto biblioteche di area socio-economica in Piemonte <http://www.bess-piemonte.it/checosebess/convenzione.htm>

[2] Si tratta di un accordo di collaborazione tra la Direzione Generale per le Biblioteche, gli Istituti Culturali ed il Diritto d'Autore del Ministero dei Beni e delle Attività Culturali (MiBAC) e il Consiglio Nazionale delle Ricerche (CNR) firmato dal Direttore Generale del MiBAC, dott. Maurizio Fallace e dal Vice Presidente del CNR, Prof. Roberto de Mattei. L'intesa ha avviato una prima fase di sperimentazione, oggetto del seminario di Roberto Puccinelli.

3 Inizialmente anche l'Agenzia spaziale italiana.

4 Cfr. GIUSEPPE VITIELLO, L'identificazione degli identificatori, "Biblioteche oggi", 22 (2), 2004, p. 68.

5 ANTONELLA DE ROBBIO, URI, URN e URL, una questione di definizioni: universal versus uniform, "Biblioteche oggi", 20 (1), 2002, p. 34.

6 Si veda lo schema dei vari URI pubblicato da IANA, organismo che tra l'altro sovrintende l'assegnazione di indirizzi IP a livello mondiale <http://www.iana.org/assignments/uri-schemes.html>, ultimo accesso 28 aprile 2011.

7 Per una definizione formale si veda TIM BERNERS-LEE et al. (Dicembre 1994), Uniform Resource Locators (URL), IETF http://www.faqs.org/rfcs/rfc1738.html.

8 Ibidem.

9 Si veda l'IETF RFC 2141< http://www.ietf.org/rfc/rfc2141.txt>

10 Si veda <http://purl.oclc.org/docs/index.html >

11 Si veda <http://tools.ietf.org/html/draft-kunze-ark-15>

12 Si veda <http://www.handle.net/>

13 Si veda <http://www.doi.org/>

14 Si veda <http://www.crossref.org/>

15 Si veda <http://www.medra.org/it/index.htm>

16 Si vedano le specifiche dell'IETF RFC 3188 <https://datatracker.ietf.org/doc/rfc3188/>

17 Lo standard NBN è operativo anche in Finlandia e soluzioni prototipali sono presenti in Olanda.

18 Le Biblioteche nazionali centrali e l'ICCU.

19 Se una risorsa possiede già un PI (es. ISBN, DOI, etc.) è possibile utilizzare come sua parte specifica questo identificativo pregrresso.

20 Il controllo dell'impronta digitale (digest) viene effettuato attraverso l'algoritmo MD5 (Message Digest algorithm 5) e le funzioni di hash.

21 Si veda <http://www.jnbn-supported.org/index.php/it/il-software>

22 CNR e FRD partecipano al progetto internazionale PersID <http://www.persid.org> finalizzato allo sviluppo di un meta-resolver a livello europeo in grado di risolvere da un punto di accesso unico i vari PI. E' stato sviluppato un prototipo, basato unicamente su "tecnologie" open source, la cui infrastruttura recepisce le specifiche dell'RFC 3188

(NBN). Interessante è notare che - a differenza dell'infrastruttura gerarchica distribuita italiana in cui le singole associazioni vengono conservate anche a livello centrale - in questo contesto, per decisione comune, ciascun paese partecipante si occupa di conservare i propri dati localmente.

23 EUPL è la prima licenza free/open source europea per il software. La versione 1.0 è stata approvata dalla Commissione Europea il primo gennaio 2007. È una licenza copyleft che ha ricevuto l'approvazione dell'OSI (Open Source Initiative).

Pubblicato su BIBLIOTECHE OGGI, XXIX, 8, (2011), pp. 66-70.