JLIS.it

Affidabilità e conservazione a lungo termine per le risorse digitali

Maurizio Lunghi, Chiara Cirinnà
Emanuele Bellini

Introduzione

Internet ha cambiato radicalmente il nostro modo di lavorare, comunicare, vivere, produrre e accedere alle informazioni, interagire con le istituzioni, comprare oggetti e gestire risorse. Oggi infatti sono molte le risorse disponibili, spesso su infrastrutture aperte e flessibili, liberamente accessibili a tutti gli utenti: i contenuti sono utilizzabili da molti servizi basati sui requisiti degli utenti. Il web è stato probabilmente la *killer application* di internet. Negli ultimi anni il web si è spostato dal "web dei documenti" verso il "web dei dati" dove l'informazione non è più impacchettata in un documenti fissi ma è resa disponibile in modo destrutturato e utilizzabile dagli utenti in modo più flessibile. I recenti sviluppi sul web hanno visto l'emergere delle tecnologie del semantic web e Linked Open Data[1] (LOD) associate ad una quantità sempre maggiore di dati disponibili per pubblicare e connettere dati strutturati sul web. Le *best practice*

[1] http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData.

Traduzione italiana a cura dell'Autore.

sui linked data, supportate dal W3C,[2] sono pronte per essere adottate da un numero rilevante di data provider, consentendo la creazione di un spazio di dati globale – il "web dei dati". Sfortunatamente le 5 stelle[3] di LOD sono principalmente orientate verso l'usabilità e la standardizzazione dei dati pubblicati sul web, ma non prendono in considerazione l'affidabilità, la persistenza dei dati e le URI usate per riferirli. Infatti, l'obiettivo dell'approccio di LOD sembra essere orientato a rendere accessibile un'enorme quantità di dati sul web in modo non proprietario (es. CSV invece di Excel) e linkare questi dati ad altri dataset (es. Geonames[4] o DBpedia) per disambiguare il contenuto e fornire un contesto. Tuttavia in alcuni casi, e in particolare nei settori della cultura e dell'istruzione, per ottenere informazioni sull'autenticità, oltre al recupero dei dati necessari o le loro relazioni, sono altrettanto importanti, sia l'integrità che la provenienza. Sistemi di certificazione basati sui Persistent Identifier (PI) per gli oggetti digitali, per gli autori e per le istituzioni possono essere di grande aiuto al fine di migliorare la qualità delle informazioni reperibili da internet e di aumentare in gran parte l'usabilità e lo sviluppo di eventuali nuovi servizi. Questo paradigma, basato sull'identificazione e l'interconnessione dei dati, offre soluzioni a molti dei problemi reali della biblioteca, come la ricerca avanzata sul web, il controllo d'autorità, la classificazione, la portabilità dei dati e la disambiguazione. Nel web dei documenti l'identificazione e la fiducia sono stati garantiti dai siti web e dalle istituzioni che li mantengono, nel web dei dati invece sono integrati in un unico blocco di dati. L'evoluzione di questo paradigma è sempre più importante in prospettiva di una conservazione a lungo termine delle risorse digitali.

[2] W3C - http://www.w3c.it.
[3] http://www.w3.org/DesignIssues/LinkedData.html.
[4] http://www.geonames.org/ontology/documentation.html.

Requisiti per la conservazione a lungo termine delle risorse digitali

Attualmente il numero delle risorse digitali scientifiche e culturali, messe a disposizione su internet attraverso le applicazioni per la biblioteca digitale, è in costante crescita, ed è fondamentale garantire persistenza, autorevolezza, affidabilità e ampia disseminazione delle risorse supportando la loro gestione a lungo termine. Uno dei requisiti principali per affrontare questo problema è quello di adottare sistemi credibili di PI all'interno del ciclo di vita di queste risorse. Il PI deve essere assegnato solo alle risorse che siano stabili, significative per la comunità degli utenti e adatte al campo di applicazione del sistema di identificazione. Attualmente sono disponibili una serie di iniziative, standard e tecnologie, ma può essere difficile per un istituto capire quali di questi sono più appropriati per i propri oggetti digitali. Le tecnologie di PI aiutano a rendere stabile il riferimento a una risorsa digitale, anche se è noto che la persistenza non è solo un problema tecnico. In realtà queste tecnologie non sono, ovviamente, affidabili di per sé, nessuna tecnologia può esistere a tempo indeterminato o garantire servizi senza una organizzazione affidabile e politiche chiaramente definite. Nella nostra visione sistemi di PI sono da intendersi come la tecnologia disponibile, insieme ad un'organizzazione affidabile e a politiche precise per la conservazione digitale implementati dai responsabili della comunità di utenti. Il concetto di persistenza si sposta così dall'essere legato all'impegno di un singolo ente/autorità di registrazione all'impegno della comunità degli utenti serviti da PI. Un sistema di PI può essere considerato quindi come un contratto tra gli utenti finali e i fornitori di servizi responsabili per l'attuazione e il mantenimento del servizio di PI e la funzionalità del sistema. Da questo punto di vista la persistenza di un PI dipende anche dall'impegno della comunità che

promuove e utilizza il sistema di identificazione usato per le proprie risorse. Questo accade quando lo standard adottato è effettivamente orientato alle esigenze della comunità e l'autorità responsabile per la gestione del sistema è riconosciuta dalla comunità stessa. È ben noto che l'instabilità strutturale di URL semplici (ad esempio domini non più disponibili) e le relative risorse (trasferimento o aggiornamento) è uno dei problemi principali che impedisce l'utilizzo di internet come una piattaforma affidabile per la ricerca e la diffusione di contenuti digitali. L'uso corrente di un approccio con semplici URL usato come identificatore persistente di oggetti digitali comporta molti e documentati rischi in una prospettiva a lungo termine, non solo per il recupero e l'accesso delle risorse, ma anche per quanto riguarda la perdita del riferimento ai documenti digitali o la mancanza di garanzia di autorità e di provenienza. Tali rischi riguardano:

a) i beni culturali e la ricerca, impedendo la realizzazione di servizi affidabili basati sulla citazione, la valutazione della ricerca, la conservazione digitale, l'accesso, ecc,

b) il dominio business, impedendo l'utilizzo dei servizi di acquisto forniti su questi oggetti,

c) la pubblica aministrazione (e-gov), rallentando il processo di dematerializzazione.

È chiaro che il problema non è solo quello di affrontare l'errore HTTP 404, ma di andare verso sistemi di identificazione in grado di sostenere l'autorità, l'affidabilità, la conservazione, la certificazione, lo sfruttamento e l'ampia diffusione di queste risorse. Una soluzione affidabile è quella di associare un PI affidabile alle risorse digitali.

La sfida della fiducia (trust)

La fiducia o *trust*, in generale, riguarda la valutazione e la gestione dei rischi percepiti da ciascun attore nell'avviare una relazione. In altre parole, fiducia comporta rischio. Secondo l'ISO il rischio può essere definito come la combinazione della probabilità di un evento e delle sue conseguenze (ISO/IEC Guide 73). Ci sono un certo numero di eventi con conseguenze negative che possono verificarsi durante la durata del servizio PI, con differenti gradi di probabilità, ma nel caso si verificassero, tutti con costi elevati. Esempi di questi rischi sono:

a) il fallimento nel determinare i costi iniziali e ricorrenti e le relative tariffe del servizio (rischi associati alla sostenibilità finanziaria);

b) l'adozione di tecnologie che sono non più disponibili (rischio connesso con l'adozione di standard);

c) la non disponibilità in rete dell'oggetto identificato (rischio associato all'accordo tra i fornitori di contenuti (*content providers*) e i fornitori di servizi (service provider);

d) la perdita del credibilità da parte della comunità (rischio associato al mandato della comunità), ecc.

Questi fattori possono determinare una diminuzione (lowering) nella fiducia del servizio di PI da parte del content provider e colpiscono la diffusione e la valorizzazione delle risorse digitali. Ad esempio, i vari repository digitali archiviano oggetti ed entità intangibili e li rendono disponibili agli utenti attraverso le reti telematiche: noi accediamo al nostro conto in banca così come all'ospedale o in comune per documenti ufficiali, scarichiamo grandi quantità di file e

chattiamo con avatar. Ma chi certifica l'identità degli attori e garantisce la nostra privacy? Come possiamo contare sull'autenticità dei documenti che scarichiamo? E anche come possiamo fidarci dell'istituzione che rilascia un documento ufficiale? Qual è il rischio se non si può dimostrare che un documento non è valido per i nostri scopi previsti? Quali sono i rischi in generale? È chiaro che una buona dose di fiducia è necessaria per vivere in questo mondo virtuale e artificiale. Un servizio di PI deve affrontare almeno i seguenti requisiti fondamentali:

1. *unicità globale*: il PI è chiaramente parte di un nome di dominio ed è unico e associato ad una risorsa unica;

2. *persistenza*: si riferisce alla durata permanente delle proprietà significative di un identificatore (ad esempio, non è possibile riassegnare il PI ad altre risorse o cancellarlo);

3. *risolvibilità*: si riferisce alla possibilità di recuperare informazioni riguardanti una risorsa o per accedervi direttamente su internet.

Attualmente ci sono diverse tecnologie e standard per l'implementazione di sistemi di PI, ma non c'è un accordo generale sulla loro adozione, spesso perché alcuni di questi sistemi sono nati come soluzioni tecniche, senza il supporto della comunità di utenti che necessitano di specifici livelli di servizi PI. Sistemi come il PURL o il Cool URI(**Bernerslee2009**) hanno considerevoli vantaggi nel supportare l'implementazione del web dei dati grazie alla loro immediata dereferenzialità attraverso il protocollo http, ma ci sono diverse limitazioni causate dal fatto che la loro persistenza non è garantita in principio da una terza parte indipendente e affidabile. È noto che l'approccio alla persistenza dei Cool URI si basa sulla progettazione delle URL. Questo approccio, anche se è considerato

una *best practice* per l'implementazione del semantic web in generale e dei linked data in particolare, è principalmente basato su una approccio tecnico. L'assunzione di base è quella che una URI progettata correttamente dovrebbe ridurre la necessità di dover essere cambiata per garantirne la stabilità nel tempo. Un esempio di questa best practice è quello di evitare di esplicitare nella URI l'estensione della pagine web come .php o .asp così che un cambiamento della tecnologia (ad esempio da PHP ad ASP) di implementazione non si riflette nella sua forma. In questa prospettiva, la persistenza è basata unicamente sull'impegno delle singole istituzioni che stabiliscono un rapporto di fiducia direttamente con gli utenti finali, senza la mediazione di una terza parte. Purtroppo, è ben noto che l'impegno di una singola istituzione non è più sufficiente a garantire né la persistenza a lungo termine delle URL, né l'affidabilità delle risorse in termini di provenienza, autenticità, integrità, conservazione, e così via. In pratica le risorse si muovono in rete, possono essere modificate o cancellate a causa di una moltitudine di fattori che non sempre possono essere predeterminati o regolati da politiche di gestione del contenuto delle istituzioni o disciplinati da best practice. Un caso tipico si verifica quando un ente cessa di esistere perché assorbito da un'altra istituzione, o viene eliminato, o semplicemente perché il suo nome ufficiale è stato cambiato. In questi casi, gli oggetti digitali possono essere rinominati per essere adattati al flusso di lavoro della nuova istituzione, o trasferiti ad altre istituzioni, o nel peggiore dei casi eliminati perché non sono più rilevanti per gli obiettivi istituzionali. È chiaro che tutte queste azioni possono causare la "rottura" della vecchia URL indipendentemente da come è stato costruita. Questo potrebbe non essere un problema se l'ente non gestisce risorse scientifiche, culturali o amministrative, ma diventa un problema critico se questi cambiamenti influenzano istituzioni come gli archivi di dati scientifici, biblioteche, pubblica

amministrazione, e così via. In questi casi, per esempio, bibliografie basate sulle semplici URL o anche Cool URI, che si riferiscono a risorse che erano presenti negli archivi di queste istituzioni, non possono più essere usate in modo affidabile per verificare i lavori scientifici o calcolare indici bibliometrici. Un altro problema critico è collegato alla connessione dei dataset che sono stati aggiornati più volte. In questi casi, può essere difficile o addirittura impossibile verificare la validità del risultato scientifico presentata in un articolo correlato. Ciò che è più importante, tuttavia, è l'impossibilità di implementare sistemi per controllare l'autenticità, la provenienza e l'integrità di tali risorse a causa dell'assenza di un terza parte in grado di garantire l'associazione nome - risorsa. In questo scenario, la maggior parte dei benefici di un ampio accesso a dataset linkati tra loro vengono vanificati dalla mancanza di affidabilità (trust).

Il servizio NBN:IT come supporto alla trust in LOD

Per affrontare la sfida della trust in LOD, una possibile soluzione potrebbe essere quella di implementare un sistema di PI basato sulla tecnologia URN (Uniform Resource Name).[5] Attualmente, per implementare un sistema di PI l'approccio principale è quello di separare l'identificazione delle risorse della loro localizzazione. Come indicato in precedenza, Tim Berners Lee ricorda che l'adozione di politiche chiare e stabili e linee guida di attuazione sono sufficienti per gestire l'identificazione persistente di risorse su internet. Anche se questo suggerimento è ragionevole e adeguato per alcuni settori,

[5] APARSEN DE22.1 Persistent Identifiers Interoperability Framework - http://www.alliancepermanentaccess.org/wp-content/plugins/download-monitor/download.php?id=D22.1+Persistent+Identifiers+Interoperability+Framework.

è evidente che non possiamo delegare questa responsabilità a ciascun istituto, in particolare nel settore del patrimonio culturale e scientifico per due motivi principali:

1. molte istituzioni non riescono a decidere l'approccio e la strategia da adottare in termini di selezione dei contenuti, formati, denominazione, ecc.

2. domini, infrastrutture, servizi, l'organizzazione istituzionale e impegno possono cambiare nel corso del tempo e, come spiegato sopra, la persistenza è un contratto con una comunità di utenti definita, più che con una sola istituzione.

In ogni caso, le URI sono ampiamente utilizzate nel contesto del semantic web per identificare qualsiasi tipo di risorsa od oggetto che sia reale, digitale, astratto, virtuale, cercando di armonizzare in una visione semantica tutte le applicazioni della comunità utente. Ad esempio, per risolvere questo problema, lo schema info-URI[6] è stato sviluppato dalle comunità delle biblioteche e degli editori per URI di asset informativi che hanno identificatori in *namespace* pubblici, ma che non hanno rappresentanza nell'allocazione delle URI". È chiaro che, per fare riferimento in modo affidabile a un oggetto digitale certificato, l'uso di URN o identificatori che implementano l'RFC 1737 (Requisiti funzionali per URN) è oggi una best practice. Lo scopo di un URN è di fornire un identificatore univoco, globale, persistente, indipendente dalla posizione che può essere utilizzato per l'identificazione e l'accesso alle caratteristiche di una risorsa o per l'accesso alla risorsa stessa. La specifica URN è parte della famiglia di specifiche IETF compresi nel Uniform Resource Identifier (URI) framework. Tale framework comprende anche gli URL, che specificano sia un protocollo che una posizione, al fine di consentire

[6]RFC 4452: http://info-uri.info.

l'accesso alle risorse sul Web. La IANA è l'autorità di registrazione per i namespace URN. Gli URN sono progettate per consentire la mappatura di *namespace* eterogenei su un'unica struttura, e quindi consentire il riutilizzo degli identificatori.
A differenza delle URL, gli URN non sono direttamente azionabili (i browser in genere non sanno cosa fare con un URN) perché non hanno alcuna infrastruttura globale associata che consente la risoluzione (come ad esempio il DNS supporta le URL). Anche se sono state realizzate diverse implementazioni, ognuna propone i propri sistemi per la risoluzione attraverso l'utilizzo di plug-in o proxy, mentre un'infrastruttura che consente la risoluzione degli URN su vasta scala non è stata ancora implementata. Ma le singole implementazioni di namespace come il URN-NBN o il DOI, offrono un servizio di risoluzione disponibile su internet. Il namespace NBN, come un Namespace Identifier (NID), è stato registrato e adottato dal Nordic Metadata Project, ma è stato sviluppato separatamente da sistemi diversi senza implementazioni di riferimento che consentono il coordinamento delle fonti di informazione. In realtà, diverse biblioteche nazionali hanno sviluppato i propri sistemi NBN nel contesto di progetti nazionali; diverse implementazioni sono attualmente in uso, ognuna con metadati descrittivi diversi o livelli di granularità. Appare chiaro quindi che non tutti i sistemi di PI possono supportare la trust del LOD con successo. Il servizio NBN-Italia supporta almeno tre livelli di persistenza(**Bellini2012**):

1. *Persistenza dell'identificatore.* Se la risorsa non è più disponibile in rete, l'identificatore URN può continuare ad avere senso (ad es. come prova che in un determinato momento quella risorsa esisteva);

2. *Persistenza dell'abbinamento URN e URL.* Si tratta di un impegno che garantisce nel lungo periodo che un URN sia sempre risolto (porti almeno a un indirizzo di tipo URL). Non viene

garantita l'accessibilità della risorsa ma viene garantito l'accesso alla cosiddetta *tombstone* in caso di non disponibilità permanente della risorsa in rete (es. "questo e-book non è più in commercio");

3. *Persistenza della risorsa referenziata da NBN.* Assicurare nel lungo periodo l'esistenza e accessibilità della risorsa referenziata da URN. Questo è il livello di persistenza di NBN reso possibile solo dal deposito (legale o volontario) presso le biblioteche nazionali e dalla descrizione autorevole della bibliografia nazionale.

Grazie a questi livelli di servizio, i nomi del sistema NBN-Italia rappresentano un evidente valore aggiunto se usati nelle architetture LOD per sostenere la trust delle asserzioni nelle triple RDF. Questa proposta va verso l'integrazione dei sistemi di LOD e PI, sfruttando progetti e iniziative in corso, come indicato nel paragrafo successivo.

Prossimi passi: Den Haag Manifesto 2.0 e Firenze Agenda

Il prossimo evento Cultural Heritage On Line 2012, che si terrà a Firenze a dicembre del 2012, si propone di migliorare e rendere efficace il Den Haag Manifesto attraverso il coinvolgimento di iniziative e progetti in corso e di stakeholders come: APARSEN NoE,[7] Datacite,[8] EPIC,[9] e PersID[10] / URN-NBN, W3C,[11] Knowledge Exchange,[12] e

[7] APARSEN - http://www.alliancepermanentaccess.org.

[8] Datacite - http://www.datacite.org.

[9] European Persistent Identifier Consortium (EPIC) - http://www.pidconsortium.eu.

[10] PersID- Building a persistent identifier infrastructure - http://www.persid.org.

[11] W3W - http://www.w3c.it.

[12] Knowledge Exchange - http://www.knowledge-exchange.info.

così via. Due degli obiettivi principali che ci accingiamo a realizzare sono:

1. una revisione del Den Haag Manifesto e il suo miglioramento verso una versione 2.0;

2. la redazione di una Firenze Agenda per definire una strategia comune per l'implementazione di un LOD che sia effettivamente trust.

Den Haag Manifesto 2.0

Nei recenti sviluppi alcune iniziative stanno fondendo l'approccio aperto dei linked open data e le potenzialità del web semantico con il valore aggiunto dato dall'identificazione, autenticità, e provenienza offerte dai sistemi di PI. Il Knowledge Exchange ha organizzato un seminario[13] sugli identificatori persistenti per gli oggetti invitando al confronto sui servizi le varie iniziative correnti per esplorare future possibilità di cooperazione e convergenza. Il seminario si è svolto il 14-15 giugno 2011 presso gli uffici del DANS (Data Archiving and Networked Services) a L'Aia ed è stato ospitato da PersID, la Fondazione SURF[14] e il DANS. Tre i principali attori nel settore degli identificatori persistenti per gli oggetti, Datacite/DOI, EPIC/Handle e PersID/URN-NBN, che si sono scambiati a vicenda informazioni sugli sviluppi recenti, le esperienze degli utenti e le politiche. Nelle sessioni i partecipanti hanno discusso i benefici e le sfide che emergono dalla presenza di più sistemi di PI e la relazione tra i PI ed le comunità dei linked open data: ci fu un chiaro interesse a collegare i sistemi PI agli standard dei linked data. Ciò ha portato al Den

[13] Knowledge Exchange - http://www.knowledge-exchange.info/Default.aspx?ID=440.

[14] http://www.surf.nl/en/Pages/default.aspx.

Haag Manifesto (DHM), che delinea una serie di azioni concrete per includere i PI nella comunità linked open data. La Fondazione Rinascimento Digitale ha partecipato al gruppo di lavoro per definire le opportunità di collaborazione tra LOD e sistemi di PI. Nel corso della riunione è emersa una sorta di distanza culturale tra la comunità LOD e quella dei PI. Le principali differenze riguardavano i concetti di identificazione, persistenza e affidabilità. Infatti l'approccio LOD è fortemente orientato alla rappresentazione del flusso di informazioni sul web. Da questo punto di vista la risorsa può cambiare nel corso del tempo in base al flusso di lavoro di pubblicazione. Per esempio, un set di dati può essere aggiornato sul web diverse volte, mentre il suo URI può rimanere lo stesso. Con una visione opposta, i domini PI sono più orientati a identificare risorse stabili gestite da sistemi di Trusted Digital Repositories. Durante i lavori si è cercato di individuare le principali caratteristiche dei sistemi di PI che possono essere importati in LOD. I risultati di questa prima valutazione è stata la definizione di un manifesto in 5 punti che ha impegnato "moralmente" le istituzioni che operano nel campo della PI e LOD ad accertare le loro possibilità di integrazione. I punti identificati sono i seguenti:

1. Un PI può essere un URI http includendo la negoziazione del contenuto;
2. Uso di vocabolari LOD per gli elementi di schema;
3. Individuazione di un insieme minimo di elementi comuni tra gli identificatori spaziali (esempi sono i metadati DOI *kernel*, DataCite kernel etc.);
4. Utilizzare la relazione 'same as' per aiutare l'interoperabilità dei PI;
5. Utilizzare gli IP per i subjects e gli objects delle triple RDF.

Da allora, il DHM viene utilizzato come base per un approccio coordinato tra le comunità PI e LOD ai problemi di identificazione ma, partendo da questi punti, il DHM deve essere rivisto, precisato e ampliato secondo le attuali tendenze e soluzioni. Inoltre deve essere supportato da un programma comune in grado di guidare le prossime implementazioni di LOD e PI, in modo da avere soluzioni armonizzate e interoperabili: la Firenze Agenda.

Una proposta per la Firenze Agenda

Attualmente FRD è responsabile di un Work Package (WP22) specifico sull'interoperabilità dei IP e di LOD nell'ambito del progetto APARSEN. APARSEN è una rete di eccellenza di 34 istituzioni ed è co-finanziata dalla Commissione europea al fine di risolvere la frammentazione della conservazione digitale dei record scientifici in Europa. Nel primo anno il WP22 ha sviluppato un *reference model* per l'interoperabilità dei sistemi di PI. Il lavoro è iniziato con l'individuazione degli *user requirements* per gli identificatori per gli oggetti digitali, le persone e le istituzioni, e successivamente sono stati concordati alcuni criteri per la trust dei sistemi di PI. Infine, è stato proposto un Interoperability Framework in cui qualsiasi sistema di PI affidabile può condividere i suoi dati tramite uno schema comune; il modello propone un'ontologia per l'interoperabilità dei sistemi di PI in linea con l'approccio LOD. L'iniziativa italiana NBN segue lo stesso flusso. Il progetto NBN è guidato dal consorzio italiano per il deposito legale[15] ed ha definito criteri e linee guida per l'assegnazione dei PI. Questo flusso di lavoro, definito in collaborazione con le biblioteche nazionali di Firenze, Roma e Venezia, che gestiscono tale servizio, assicura il corretto livello di trust al PI generato che, attraverso il suo riutilizzo nel dominio LOD, consente l'attuazione

[15] www.depositolegale.it.

del T-LOD (Trusted LOD). La Firenze Agenda si propone quindi di individuare le tappe fondamentali, le linee guida e i criteri che possono essere adottati dalle comunità dei PI e LOD per cooperare nel costruire un web dei dati più affidabile.

Ai fini di una corretta indicizzazione, si invitano i lettori a citare esclusivamente il testo in lingua inglese; l'unico, infatti, che presenta l'indicazione del numero di pagina, l'abstract, le keywords e le date del processo redazionale.

Lunghi, M., C. Cirinnà, E. Bellini. "Trust and persistence for Internet resources". *JLIS.it*. Vol. 4, n. 1 (Gennaio/January 2013): Art: #5494, p. 1–15. DOI: 10.4403/jlis.it-5494. Web.