



Gli Identificatori Persistenti per i Beni Culturali

È risaputo che le risorse di Internet tendono ad avere una vita breve; la loro identificazione e localizzazione permanente pone problemi complessi che riguardano questioni tecnologiche e organizzative e che implicano la citazione, il reperimento e la conservazione delle risorse culturali/scientifiche. Questo non è solamente un problema tecnico, ma anche organizzativo: l'identificazione degli oggetti digitali, quali testi, musica, video, fotografie, documenti scientifici e così via, è ancora uno dei maggiori problemi che impediscono di considerare Internet una piattaforma affidabile per la ricerca e la disseminazione di contenuti culturali e scientifici.

Perchè c'è bisogno di un "Identificatore Persistente"?

La conservazione a lungo termine, la disseminazione e l'accesso agli oggetti culturali digitali sono adesso tra le missioni prioritarie delle istituzioni culturali, come le università, gli archivi, i musei e le biblioteche. L'uso dell'URL non può essere considerato un approccio affidabile per risolvere queste questioni, a causa dell'instabilità strutturale dei link (ad esempio domini non più disponibili) e delle risorse collegate (rilocazione o aggiornamento). Il corrente utilizzo dell'indirizzo URL accresce il rischio di non recuperare i documenti o di sottoutilizzare le collezioni disponibili. Nel settore dei Beni Culturali è essenziale non soltanto identificare una risorsa ma anche garantirne un accesso continuo nel tempo.

Una soluzione affidabile è quella di associare un Persistent Identifier (PI) a una risorsa digitale, che rimarrà perennemente associato alla risorsa indipendentemente da dove essa sia collocata.

Queste sono le principali attività che devono essere eseguite per implementare un sistema di PI:

- 1) Selezione delle risorse che necessitano di un PI
- 2) Assegnazione di un nome alla risorsa e creazione di un registro
- 3) Risoluzione di un PI con la URL associata
- 4) Mantenimento del registro che associa PI-URL e garanzia di un accesso continuo alle risorse

La prima azione è prerogativa di ogni istituzione culturale, mentre quelle successive possono essere delegate ad altre autorità per garantire una migliore sostenibilità economica e funzionale del servizio.

Ogni istituzione culturale dovrebbe scegliere un sistema di Persistent Identifier tenendo conto dei seguenti requisiti:

- Unicità globale
- Persistenza
- Risolvibilità
- Affidabilità
- Autorevolezza
- Flessibilità
- Interoperabilità
- Costi

Glossario

Oggetto: Qualsiasi entità intellettuale definita dai metadati e dalla terminologia di un dizionario (es. indecs) per assicurare che “ciò che tu intendi sia ciò che io intendo” (interoperabilità). Gli oggetti possono essere fisici, digitali o astratti, es. persone, organizzazioni, accordi, ecc.

Servizio di risoluzione (dereference): il processo in cui un identificatore si configura come l'input (richiesta) di un servizio in rete, per ricevere di ritorno uno specifico output (risorsa, metadati, ecc).

Naming authority: Autorità indipendente che assegna i nomi e garantisce la loro unicità e persistenza. Ogni servizio di risoluzione dei nomi corrisponde ad una naming authority. Un sistema di PI distribuito prevede che la responsabilità della generazione e della risoluzione possano essere distribuite ad altre istituzioni chiamate sub-naming authorities, che gestiscono la porzione del nome dello spazio/dominio.

Namespace: un container astratto che fornendo un contesto per gli oggetti permette la disambiguazione di quelli che hanno lo stesso nome ma risiedono sotto diversi namespace.

Registro: tabella di associazione dei nomi tra URN ed uno o più URL.

Repository: luogo in cui le risorse digitali sono contenute con (DSpace, Fedora, Codex, ecc.) o senza (file system) un sistema di gestione delle risorse.

URI: un Uniform Resource Identifier è il generico insieme di tutti i nomi/indirizzi caratterizzati da brevi stringhe che si riferiscono alle risorse.

URL: un Uniform Resource Locator è un URI che, oltre ad identificare una risorsa ne permette di ottenere una rappresentazione attraverso la descrizione del suo meccanismo di accesso principale o della localizzazione nella rete.

Unicità Globale

Si può considerare l'identifier un'etichetta associata ad un oggetto in un certo contesto. Con “contesto” si fa riferimento sia al tipo di standard usato per la sintassi del nome (ad es. URN:NBN:IT:xxx-xxxx) sia all'identificazione dell'autorità (sub-namespace) che assegna questa etichetta.

Persistenza

La persistenza si riferisce alla vita permanente di un identificatore. Non è possibile riassegnare il PI ad altre risorse o cancellarlo. Il PI sarà globalmente unico per sempre e potrà essere usato come referenza di una risorsa anche oltre la vita della risorsa identificata o della autorità responsabile dell'assegnazione del nome. La persistenza è evidentemente una questione specifica dei servizi e della politica delle istituzioni culturali. L'unica garanzia di utilità e di persistenza dei sistemi di identificazione è l'impegno mostrato dalle organizzazioni che assegnano, gestiscono, e risolvono gli identificatori.

Risolvibilità

La risolvibilità si riferisce alla possibilità di recuperare una risorsa solamente se è pubblicata. È importante distinguere il concetto di identificazione da quello di risoluzione. La scelta del namespace di identificazione non implica necessariamente di scegliere una corrispondente architettura di risoluzione.

Affidabilità

Per assicurare l'affidabilità di un sistema di PI, bisogna accertare due aspetti: l'infrastruttura PI deve sempre essere attiva (servizio di ridondanza, servizi di back-up di deposito, etc.) e il registro aggiornato (attraverso sistemi automatici).

Autorevolezza

L'unica garanzia di utilità e di persistenza dei sistemi di identificazione è costituita dall'impegno dimostrato dalle organizzazioni che assegnano, gestiscono e risolvono gli identificatori. Nel settore dei beni culturali la tendenza è quella di far uso di servizi forniti dalle istituzioni pubbliche, come le biblioteche nazionali, gli archivi di stato etc. Requisiti quali l'autorevolezza e la credibilità di un sistema di PI dovrebbero essere attentamente valutati prima di adottare una soluzione.

Flessibilità

Un sistema di identificatori è molto efficace se è capace di conformarsi agli speciali requisiti di differenti tipi di risorse o di collezioni. Per esempio, un sistema di identificatori dovrebbe essere in grado di gestire diversi livelli di granularità poiché, a seconda dei campi di applicazione dell'utente, cambia quello a cui un identificatore punta.

Interoperabilità

Questo aspetto è fondamentale per garantire la possibilità di diffusione e di accesso agli oggetti digitali.

Sono disponibili molte tecnologie e diversi approcci e alcuni di essi sono conformati sulla base dei requisiti di specifici settori. L'interoperabilità tra i diversi sistemi deve essere conseguita almeno al livello dei servizi, in modo da offrire interfacce semplici da utilizzare. Il sistema di interoperabilità si può basare sull'adozione di standards aperti.

Tecnologie

PURLs (persistent URL): PURL è l'acronimo per Persistent Uniform Resource Locator.

Dal punto di vista del funzionamento, un PURL si comporta come un URL. Ciononostante, invece di puntare direttamente alla localizzazione di una risorsa su Internet, un PURL punta ad un servizio di risoluzione intermedio, utilizzando le capacità standard di risoluzione del web server che può reindirizzare alla effettiva localizzazione del documento la richiesta per la risorsa, usando un persistent identifier.

www.purl.org

URN (Uniform Resource Name): l'URN è un URI che usa lo schema URN e non implica la disponibilità della risorsa identificata. Gli URN sono utilizzati come identificatori persistenti della risorsa, indipendenti dalla localizzazione e sono progettati per semplificare la mappatura di altri namespace (che condividono le proprietà degli URN) come URN. Dunque la sintassi degli URN fornisce un mezzo per codificare dati con caratteri in una forma che può essere veicolata tramite i protocolli esistenti, trascritta sulla maggior parte delle tastiere ecc.

www.ietf.org/rfc/rfc1737.txt

HANDLE SYSTEM: l'Handle System è un servizio generico di assegnazione di identificatori che ne consente una risoluzione garantita e l'amministrazione in reti come Internet. L'Handle System gestisce gli "handle" che sono nomi univoci sia per gli oggetti digitali che per altre risorse in Internet. Una naming authority è autorizzata a creare e mantenere gli "handles", l'identificatore deve essere unico per quella authority e non è prevista una sintassi predefinita.

www.handle.net

XRI (OASIS Extensible Resource Identifier): lo scopo del XRI è quello di definire uno schema URI ed un corrispondente namespace URN per servizi di directory distribuite, che permettano l'identificazione delle risorse (incluse persone e organizzazioni) e la condivisione di dati tra vari domini, imprese e applicazioni.

www.oasis-open.org

ARK (Archival Resource Key) (IETF Internet draft): lo schema intende facilitare l'assegnazione del nome e il recupero degli oggetti informativi. Un principio fondatore dell'ARK è che la persistenza è una questione puramente di servizio e non è riferita ad un oggetto o attribuita ad esso da una determinata sintassi di naming. L'identificatore ARK risolve 3 diversi elementi: risorsa, metadato e modalità di conservazione.

www.cdlib.org/inside/diglib/ark

N2T (Name to thing): l'N2T è un consorzio di organizzazioni del settore delle memorie culturali che ha un web server ordinario, ridonato in diverse istanze per l'affidabilità. Questo progetto mira a proteggere 200 URL di organizzazioni dall'instabilità dell'hostname con 200 regole di riscrittura attraverso un semplice reindirizzamento http per ciascuna organizzazione.

Costi

Nel settore dei beni culturali i sistemi di PI adottati dovrebbero essere esenti da spese o almeno a costi sostenibili, perché il ruolo delle istituzioni culturali è quello di garantire libero accesso alle risorse nel tempo e di evitare il "digital divide".

Altre considerazioni

Granularità

La granularità si riferisce al livello di dettaglio di una risorsa a cui i persistent identifiers dovranno essere assegnati. Il requisito di granularità avrà un considerevole impatto sul sistema degli identificatori che un'istituzione adotta.

In alcune situazioni può essere necessario citare una pagina web che serve come accesso a una collezione di web files, o citare un articolo di giornale, un item o un capitolo. Ad ogni modo per la gestione dei diritti potrebbe essere richiesto un maggior dettaglio. Ogni istituzione dovrebbe valutare se un servizio di PI fornisce il giusto livello di granularità per i propri tipi di risorse.

Identificatori opachi o "parlanti"

Un persistent identifier può non contenere alcuna informazione riguardo l'oggetto che identifica (opaco) per il fatto che si basa su caratteri casuali che non hanno unità semantiche associate.

Un identificatore opaco richiederà sempre un servizio di risoluzione per essere identificato, ma potrebbe avere qualche dato semantico incorporato (parlante) sufficiente ad identificarlo.

Di solito è più facile memorizzare e usare identificatori mnemonic-based, invece di quelli che contengono una sequenza di caratteri senza significato, sebbene ciò non abbia alcuna rilevanza nella elaborazione.

Gestione delle versioni

Ogni nuova versione di una risorsa richiederà un persistent identifier separato. Una nuova versione può essere considerata come un oggetto digitale differente perché il suo contenuto o il formato fisico può essere stato modificato. La gestione delle differenti versioni può essere realizzata attraverso regole di denominazione o tramite i metadati.

In che modo le tecnologie possono aiutarci?

L'applicazione dei PI richiede un database che possa tenere traccia dell'attuale localizzazione di un oggetto digitale, chiamato "resolver database". Un resolver database mappa la localizzazione della risorsa e reindirizza l'utente alla localizzazione corrente. Il resolver database ed il relativo servizio di risoluzione possono essere implementati in modo centralizzato o distribuito, ed utilizzando o meno il DNS.

Centralizzato: questa architettura è basata su un nodo centrale che genera i nomi delle risorse e assicura la loro risolvibilità e affidabilità nel tempo. Questo tipo di soluzione implica una centralizzazione delle responsabilità e della gestione dei costi; perciò un servizio di risoluzione centralizzato ha un solo punto debole.

Distribuito: questo tipo di architettura necessita di registri distribuiti e di servizi di risoluzione per ogni autorità di secondo livello impegnata nella gestione dei nomi dei propri PI; una autorità di primo livello gestisce il processo di reindirizzamento della richiesta di risoluzione all'appropriato servizio.

Riferimenti

ERPANET workshop Persistent Identifiers

Thursday 17th - Friday 18th June 2004-University
College

Cork, Cork, Ireland

www.erpanet.org/events/2004/cork/index.php

DCC Workshop on Persistent Identifiers

30 June - 1 July 2005

Wolfson Medical Building, University of Glasgow

<http://www.dcc.ac.uk/events/pi-2005/>

URN:NBN

<http://www.ietf.org/rfc/rfc3188.txt>

URN:NBN:DE

<http://www.persistent-identifier.de>

URN:NBN:IT

<http://www.rinascimento-digitale.it>

DOI

<http://www.doi.org>

ARK

<http://www.cdlib.org/inside/diglib/ark/>

PADI

<http://www.nla.gov.au/padi/topics/36.html>

PILIN

<https://www.pilin.net.au/>

OpenURL

http://www.niso.org/committees/committee_ax.html

DNS-based: il protocollo HTTP è usato per attivare il link di citazione sul web attraverso una richiesta HTTP ad un servizio di risoluzione. Questa modalità basata sul DNS non ha bisogno di specifici clients o di plug-in per browsers standard di Internet.

Non DNS-based: ulteriori implementazioni hanno aiutato a sviluppare un protocollo specifico per la gestione dell'assegnazione del nome e per la risoluzione dei PI (per esempio il DOI). In questo caso uno specifico client (o un programma di browser) deve risolvere uno specifico identifier e deve far accedere agli oggetti digitali o ai loro metadati associati. Questa soluzione può prevedere un proxy per estendere il servizio al protocollo HTTP.

Opportunità di ricerca

Con la crescita delle società ICT, molta più attenzione è stata data alla questione della stabilità dell'URL quando si accede alle risorse su Internet. I Persistent Identifier sono una risposta relativamente recente a questo problema. Il contesto estremamente dinamico in cui operano questi sistemi sta facendo emergere ampi margini di ricerca. Di seguito sono riportati alcuni aspetti interessanti e ancora non risolti da studiare più a fondo:

- la tendenza attuale è quella di adottare sistemi che si riferiscono al dominio d'uso (per esempio l'NBN nel settore bibliotecario). Comunque una risorsa può far parte di più domini e può essere identificata con diversi sistemi. Perciò è necessario garantire l'interoperabilità tra diversi sistemi di identificazione ed implementazioni basate su uno stesso namespace;

- I Persistent Identifier consentono di accedere alle risorse ma anche ai loro metadati, che sono fondamentali per permettere all'utente di identificare il contenuto. Perciò è sempre più importante sviluppare una gestione avanzata dei metadati e dei servizi utenti;

- Le relazioni semantiche tra gli oggetti multimediali possono essere prese in considerazione per definire ontologie e per una migliore conoscenza delle risorse Internet.