1. OPEN ARCHIVE

1.1 INTRODUZIONE

Questo capitolo introduce gli Open Archive, ponendo l'attenzione sull'importanza e l'esigenza di una nuova forma di comunicazione scientifica come supporto alla divulgazione della conoscenza umana.

Hanno contribuito in maniera fondamentale per questa introduzione, i concetti e le problematiche relative agli open archive esposti dalla dott.sa De Robbio nella pubblicazione sul notiziario del SIMAI 2002 [B1] e nell'atto di convegno sui metadati svoltosi a Roma nel 2001 [B2].

Importanza fondamentale assume infatti la definizione qui esposta di metadato, entità che permette una migliore percezione delle caratteristiche proprie di una qualunque risorsa informativa e sulla quale si basano gli sforzi che l'iniziativa OAI [S1] ha sostenuto e sostiene tuttora, per garantire l'interoperabilità tra gli archivi nel mondo.

Ancora importante è l'individuazione delle parti generali di un open archive che lo caratterizzano: meccanismo di sottomissione dei contenuti, sistema di immagazzinamento degli stessi e meccanismo di raccolta dei metadati che determina la possibilità di poter creare servizi da offrire all'utente finale che vadano oltre la semplice ricerca di informazioni indicizzate.

Vengono anche esposte le problematiche inerenti alla divulgazione dei risultati di ricerca scientifica e alla possibilità di risolverle con l'adozione degli open archive da parte delle Istituzioni Accademiche.

Infine viene data una breve esposizione della storia dell'OAI a partire dalla

Convention di Santa Fe, evento al quale viene fatta corrispondere la nascita degli open archive, fino alla definizione del protocollo OAI-PMH.

1.2 OPEN ARCHIVE

Le caratteristiche principali di un *Open Archive*, archivio informatico in cui possono essere depositati dei contenuti, sono le seguenti:

- *Un meccanismo di sottomissione*, cioè un sistema dove gli utenti possono depositare i propri contenuti: qualsiasi informazione o dato (documenti, immagini, video, audio,…) in formato digitale.
- *Un sistema di immagazzinamento a lungo termine* che permette di poter conservare in maniera persistente le risorse nell'archivio garantendo la disponibilità dell'informazione indipendentemente dal tipo di supporto utilizzato. Bisogna comunque precisare come l'utilizzo di un determinato supporto dipenda dallo stato di sviluppo tecnologico: ad esempio memorizzare delle informazioni su un supporto ottico, quale un cd rom, potrebbe, tra qualche anno, portare a dei problemi di accessibilità nel caso in cui tali supporti vengano del tutto soppiantati da nuove tecnologie di memorizzazione. D'altronde è noto come il supporto cartaceo che in passato era l'unico strumento di memorizzazione delle informazioni è stato affiancato nel tempo da nuovi supporti digitali riducendone l'utilizzo.
- *Un meccanismo di raccolta dati*, cioè un tipo di interfaccia che permette di creare servizi a valore aggiunto a supporto della scoperta, presentazione e analisi dei dati nell'archivio per gli utenti finali. Dove per dati raccolti si intendono i *metadati* piuttosto che le risorse depositate nell'archivio.

I metadati sono delle informazioni utilizzate sia per descrivere i "dati delle informazioni" presenti nell'archivio che per la loro localizzazione.

Per esempio i dati bibliografici che descrivono un articolo depositato nell'open archive e il modo come reperirlo.

Gli open archive sono disciplinati dall'OAI (Open Archives Iniziative), organizzazione che nasce dall'esigenza di interoperabilità tra archivi eterogenei consentendo la disseminazione degli *e-print* per scopi di studio.

Un e-print è un tipo di documento che riguarda un lavoro tecnico precedente la sua pubblicazione (preprint) nella sua forma elettronica.

Inizialmente l'OAI come tipologia di dati di interesse ha scelto di occuparsi dei soli e-print. In fasi successive ha generalizzato la natura delle informazioni trattate rivolgendosi alla disseminazione di "contenuti", ovvero risorse digitali di qualunque tipo.

Adesso vediamo di analizzare il significato del termine Open Archive:

Il termine "Archive" si riferisce a repository o depositi di informazioni in senso più ampio. In tal caso infatti la definizione rigorosa di archivio comporta la considerazione di concetti quali mantenimento nel tempo, autorizzazioni di legge e politiche istituzionali che sono invece trascurati dall'OAI per semplificarne la comprensione e l'utilizzo.

Il termine "Open" riguarda un'architettura per definire interfacce che facilitino l'accessibilità ai contenuti da parte dei provider e non l'eventuale gratuità o l'accesso illimitato alle risorse dell'archivio.

La struttura dell'OAI si compone di:

- un esecutivo per la gestione dell'iniziativa
- comitati tecnici per lo sviluppo del protocollo di comunicazione.

I problemi che tale struttura vuole risolvere riguardano la comunicazione o interoperabilità tra archivi elettronici.
Per affrontare le precedenti problematiche, l'OAI definisce un *protocollo di comunicazione* per la condivisione, la pubblicazione e l'archiviazione di dati tra archivi informatici attraverso il web.

Lo scopo primario degli open archive è la *disseminazione* dei risultati di studio da parte dei ricercatori scientifici, il cui interesse di poter pubblicare i loro risultati su riviste o altri mezzi di diffusione scientifica comporta i seguenti problemi superati appunto dagli Open Archive:

- Accedere agli articoli di una rivista a stampa o elettronica diventa più difficile a causa dell'aumento crescente dei prezzi da un lato e dei ritardi nelle pubblicazioni dall'altro.
- Le clausole di copyright sempre più restrittive, che vietano la libera riproduzione degli articoli, limitano ulteriormente l'ampia disseminazione dei lavori degli autori.
- L'ostacolo alla possibilità da parte degli autori di poter citare i loro colleghi, come conseguenza del punto precedente.
- Il peer-review, pur essendo ritenuto un utile mezzo di validazione per i lavori dei diversi autori, a volte risulta essere piuttosto rigido sopprimendo nuove idee (censura del comitato) e rallentandone la disseminazione.

Gli Open Archive in funzione della politica organizzativa adottata si suddividono in:

- *Istituzionali*, che raccolgono tutti i lavori di una determinata istituzione o ente (università, dipartimenti, …). Di conseguenza i materiali raccolti possono coinvolgere varie discipline.
- *Disciplinari*, che raccolgono i lavori di una determinata disciplina.

Per quanto riguarda gli open archive istituzionali, le aspettative riguardano principalmente gli atenei e i centri di ricerca i quali potrebbero realizzare i propri, compatibili OAI, unendosi all'iniziativa di sviluppo per una nuova forma di divulgazione scientifica, dato che la ricerca si svolge, si sviluppa, ma soprattutto si produce entro questi luoghi.
Un'altra speranza risiede nella volontà da parte dei ricercatori di autoarchiviare i propri lavori per incrementare il processo di disseminazione all'intera comunità.

1.3 STORIA OAI

La nascita degli open archive viene fatta corrispondere al primo meeting mondiale tenutosi a Santa Fè nel 1999. Prima di tale data gli autori si limitavano ad archiviare la loro produzione scientifica in repository. Nella seguente lista sono riporatati i principali archivi che si sono uniformati successivamente all'iniziativa dell'OAI:

- *xxx* fu il primo archivio di e-print, successivamente chiamato arXiv. Esso nacque come repository nel campo della fisica di energia e si allargò in seguito agli altri campi della fisica, della matematica, delle scienze non lineari e dell'informatica. L'indirizzo internet di arXiv è http://arXiv.org/
- *CogPrints* è il repository per le scienze cognitive, la psicologia, la linguistica e le neuroscienze. L'indirizzo internet di CogPrints è

http://codprints.soton.ac.uk/

- *Networked Computer Science Technical Reference Library*, meglio conosciuto come NCSTRL detto "ancestral" più che un data è un service provider in quanto fornisce accesso ai rapporti tecnici informatici depositati in arXiv e/o in altri repository. L'indirizzo internet di NCSTRL è http://www.ncstrl.org/
- *RePEc* permette agli autori che si occupano di economia di sottomettere i loro lavori ai propri archivi dipartimentali. L'indirizzo internet di RePEc è http://repec.org
- *Networked Digital Library of Theses and Dissertations* (NDLTD) costruì una biblioteca digitale di tesi e dissertazioni autorizzate dagli studenti delle istituzioni partecipanti. Creò un iter di sottomissione dei lavori in questione, sviluppando anche un DTD (Document Type Definition) XML per la validazione dei documenti quali tesi e relazioni.

La problematica che interessava la comunità scientifica era la diversità delle interfacce di ricerca per i vari archivi, da qui la necessità da parte degli utenti di dover utilizzare interfacce web differenti per archivi differenti.
Le soluzioni proposte per fornire servizi di ricerca centralizzati furono fondamentalmente due:

- ricerca attraverso vari archivi (*cross-search*)
- raccolta dei metadati dagli archivi

Nel luglio del 1999 Paul Ginsparg, Rick Luce, e Herbert Van de Sompel del Los Alamos National Laboratory riunirono un ristretto gruppo di tecnici per partecipare al meeting di Santa Fè nel Nuovo Messico ad ottobre dello stesso anno per proporre la creazione di un'organizzazione universale per

l'autoarchiviazione della letteratura di studio (Universal Preprint Service – UPS).
Da tale convegno emerse la necessità di realizzare una struttura organizzativa e tecnica per la disseminazione degli e-print in modo da trasformare la comunicazione delle informazioni di studio. Tale trasformazione consistette nella definizione di una struttura di pubblicazione aperta, su cui stabilire livelli liberi (free) o commerciali.
Lo scopo di questo meeting fu anche quello di discutere i problemi di interoperabilità e iniziare a lavorare su di un prototipo di biblioteca digitale basato sugli archivi di e-print esistenti.

Relativamente al problema di scelta tra servizi di ricerca venne privilegiata la raccolta di metadati da vari archivi (*harvesting*) piuttosto che la ricerca "cross search" che introduceva il problema di notevoli rallentamenti qualora tra gli archivi interrogati fosse presente anche un solo host particolarmente lento. A questo si aggiungano anche problemi di complessità per gli utenti finali e per i software di ricerca dato che i vari host utilizzavano linguaggi di interrogazione differenti.

Il sistema di raccolta che si concretizzava nel prototipo UPS raccoglieva metadati da archivi multipli e forniva servizi basati su di essi. In tal modo si resero possibili interrogazioni da rivolgere ad un unico host centrale.
Da qui nasce la distinzione tra due soggetti all'interno dell'organizzazione dell'UPS: i Data Provider e i Service Provider.
I primi gestiscono l'esposizione dei metadati nel repository ed eventualmente anche le risorse ivi contenute.
I secondi raccolgono i metadati dai data provider per fornire servizi agli utenti finali.
In generale emerse che un'organizzazione di tipo "provider" poteva essere o data provider, o service provider o entrambi come l'applicazione CDSware che

costituisce l'argomento centrale della parte applicativa della tesi.
I metadati che venivano scambiati potevano essere rappresentati in un formato accettato da una comunità ristretta di data e service provider, anche se il formato raccomandato dall'OAI è tuttora il Dublin Core Unqualified.

I metadati raccolti da diverse sorgenti (data provider), potevano essere messi insieme in un unico archivio gestito da un service provider per la fornitura di servizi in funzione dei tipi contenuti.

Il nome UPS fu successivamente cambiato sia per evitare confusione, dato che tale sigla è un marchio registrato appartenente ad una società di spedizioni, sia perché non tutti gli e-print sono preprint come il nome in questione lasciava intendere.
Il nome scelto in sostituzione fu inizialmente OAi acronimo di Open Archives iniziative che divenne ben presto OAI per sottolineare l'importanza dell'"Iniziativa".
Dopo il meeting di Santa Fè diverse furono le occasioni per ulteriori incontri che portarono alla definizione di un protocollo ed un formato di metadati stabili e universalmente riconosciuti.
I fondatori dell'OAI sono la Digital Library Federation (DLF), la Coalition for Networked Information (CNI), e la National Science Foundation (NSF).

L'OAI-PMH è il protocollo elaborato dall'OAI per la raccolta di metadati dai vari repository che aderiscono all'iniziativa. Esso serve a renderli disponibili ai soggetti che si occupano di fornire servizi a valore aggiunto (service provider) in funzione delle informazioni ricavate dai metadati.
Tale protocollo potrebbe divenire parte integrante dell'infrastruttura del web, così come è già avvenuto per il protocollo HTTP, se la sua semplicità verrà a combinarsi con l'interesse da parte di organizzazioni di ricerca ed editori.

BIBLIOGRAFIA

[B1] – De Robbio Antonella, "Open Archive per la comunicazione scientifica", *Notiziario del SIMAI*, N°5, pp.2-6, 2002.

[B2] – De Robbio Antonella, "Metadati per la comunicazione scientifica", *Biblioteche Oggi*, dicembre 2001.

SITOGRAFIA

[S1] – http://www.openarchives.org/

(Sito ufficiale dell'iniziativa OAI, che ha dato luogo alla nascita degli open archive e al protocollo tecnico ed organizzativo per la loro formazione)