1. OPEN ARCHIVE

1.1 INTRODUZIONE

Il presente capitolo, che prende spunto da un articolo della dott.ssa De Robbio, riguardante il ruolo giocato dagli Open Archive per la comunicazione scientifica [B2], rappresenta un'introduzione ad essi, definendo cosa si intende per "Archivio Aperto" e quali sono le sue caratteristiche.
In particolare, vengono prese in considerazione le problematiche riguardanti la diffusione dei lavori scientifici attraverso la pubblicazione di tipo tradizionale e la possibilità di risolverle, rendendo tali documenti liberamente accessibili e a disposizione della comunità, tramite l'adozione di tali archivi informatici.
Importanza fondamentale assume il ruolo dell'OAI [S1], l'organizzazione che regolamenta gli open archive, ed il protocollo da essa definito per garantire l'interoperabilità tra i vari archivi attraverso il web.
Inoltre viene anche accennata l'importanza dei metadati, ovvero le informazioni atte alla descrizione delle risorse.
Infine viene presentata in maniera concisa la storia dell'OAI, partendo dalle origini degli open archive fino alla loro evoluzione attuale.

1.2 OPEN ARCHIVE

Col termine *Open Archive* si intende un archivio informatico, in cui possono

essere depositati dei contenuti, avente le seguenti caratteristiche:

- *meccanismo di sottomissione*;
- *sistema di immagazzinamento a lungo termine*;
- *meccanismo di raccolta dati*.

Per *meccanismo di sottomissione* si intende un sistema che permetta all'utente di depositare i propri contenuti, dove con contenuto si comprende qualsiasi informazione o dato (documenti, immagini, video, audio, …) in formato digitale.

Il *sistema di immagazzinamento a lungo termine* indica la caratteristica di poter mantenere in maniera persistente le risorse memorizzate nell'archivio, garantendo che l'informazione possa essere sempre disponibile 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. Cosa per altro già avvenuta nel caso dei supporti a nastro magnetico ormai quasi del tutto soppiantati dai più efficienti dispositivi di memorizzazione ottici (cd rom, dvd, …).

Il *meccanismo di raccolta dati* dall'archivio è una sorta di interfaccia che permette a terze parti di creare dei servizi a valore aggiunto per gli utenti finali, che supportano la scoperta, la presentazione e l'analisi dei dati nell'archivio.

Quando si parla di dati raccolti da terze parti non si intendono le risorse depositate nell'archivio quanto piuttosto i *metadati*.
I *metadati*, conosciuti come "dati sui dati", sono le informazioni che descrivono i

dati, per esempio i dati bibliografici di un articolo depositato nell'open archive. E' attraverso essi che è possibile raggiungere le risorse cui si riferiscono.

L'organizzazione che regolamenta gli open archive è l'*OAI* .
L'OAI (Open Archives Iniziative) nasce dalla necessità di far interoperare archivi eterogenei per consentire la disseminazione degli *e-print*, principalmente per scopi di studio.

Un e-print è un preprint nella sua forma elettronica, dove per preprint si intende un tipo di documento che riguarda un lavoro tecnico precedente la sua pubblicazione.

Originariamente l'OAI ha preso in considerazione esclusivamente e-print come tipologia di dati trattati, successivamente si è riferita alla disseminazione dei "contenuti" in generale, dove per contenuti si intende qualunque tipo di risorsa digitale.

L'organizzazione dell'OAI si compone di un esecutivo per la gestione e di comitati tecnici per lo sviluppo del protocollo di comunicazione.
Essa ha introdotto una struttura tecnica e organizzativa per risolvere i problemi relativi alla comunicazione o interoperabilità tra archivi elettronici.

Il termine "Archive" di Open Archive è utilizzato dall'OAI per riferire *repository* o depositi di informazioni in senso più ampio. E' infatti noto come nella definizione di archivio aspetti quali mantenimento nel tempo, autorizzazioni di legge e politiche istituzionali debbano essere presi in considerazione a differenza di quanto invece specificato dall'OAI.
Il termine "Open" non riguarda la necessaria gratuità o l'illimitato accesso alle risorse dell'archivio, quanto piuttosto un'architettura per definire e promuovere

interfacce macchina, che facilitino l'"accessibilità" dei contenuti ad una varietà di provider.
Difatti dentro un Open Archive possono essere depositati anche materiali protetti ai quali è necessario accedere previa autorizzazione.

L'OAI definisce, tra le altre cose, un *protocollo di comunicazione* per permettere la condivisione di dati tra archivi informatici attraverso il web, rendendo in tal modo possibile la loro pubblicazione ed archiviazione. Difatti, un aspetto che interessa gli open archive è quello della disseminazione di risultati di ricerche scientifiche da parte degli studiosi. Tale disseminazione è stata generalmente resa possibile tradizionalmente attraverso la pubblicazione su riviste o volumi scientifici, ciò ha comportato vari problemi all'effettiva diffusione di tali ricerche che sono stati in parte superati dagli open archive:

- Di solito un preprint evolve nell'articolo di una rivista o in altre forme di pubblicazione a stampa o elettroniche, in tali casi l'accesso ad esso diventa più difficile a causa degli elevati prezzi delle riviste da un lato e dei ritardi nelle pubblicazioni dall'altro.
- A causa delle clausole di copyright sempre più restrittive che vietano la libera riproduzione degli articoli, la disseminazione dei risultati dei lavori degli autori scientifici risulta notevolmente limitata.
- L'effetto del punto precedente pone in essere la difficoltà da parte dei vari autori di potersi citare a vicenda.
- Il peer-review, essenziale strumento di validazione del lavoro dell'autore, tende ad essere sempre più rigido sopprimendo talvolta nuove idee (censura del comitato) e causando rallentamenti nella disseminazione non dovuti alla pubblicazione da parte dell'autore.

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.

Nell'ambito degli open archive *istituzionali* giocano un ruolo importante gli atenei e gli enti di ricerca, che dovrebbero creare i propri, aderendo all'iniziativa OAI e contribuendo così allo sviluppo di questa nuova forma di divulgazione scientifica, poiché la ricerca si svolge, si sviluppa, ma soprattutto si produce entro questi luoghi.

E' necessario che tutti i ricercatori autoarchivino i propri lavori per renderli il più velocemente possibile accessibili all'intera comunità.
Gli open archive introducono un modello più equo ed efficiente per la disseminazione dei risultati di ricerca.

1.3 STORIA OAI

Precedentemente al meeting di Santa Fè del 1999, convegno che diede luogo alla nascita dell'iniziativa degli open archive, il materiale contenuto nei repository veniva depositato tramite autoarchiviazione da parte degli stessi autori. I principali archivi, che successivamente si sono uniformati ai dettami dell'OAI, sono i seguenti:

- 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/
- Il 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
- Il 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.

Gli utenti che interagivano con tali archivi erano costretti ad imparare ad utilizzare interfacce web differenti per archivi differenti dato che i gestori di tali repository adottavano protocolli di comunicazione diversi.
Le soluzioni proposte per fornire servizi di ricerca centralizzati furono fondamentalmente due: la ricerca attraverso vari archivi da un lato, e la raccolta dei metadati dagli archivi dall'altro.
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 studiare la creazione di un'organizzazione universale per l'autoarchiviazione della

letteratura di studio (Universal Preprint Service – UPS), proposta al meeting di Santa Fè nel Nuovo Messico, nell'ottobre dello stesso anno,
Si stabilì la necessità di realizzare una struttura organizzativa e tecnica per la disseminazione degli e-print per trasformare la comunicazione delle informazioni di studio. Tale trasformazione consiste 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 sui 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 attraverso vari archivi (*cross search*) o raccolta di metadati da vari archivi (*harvesting*), la scelta ricadde sulla seconda soluzione.
La ricerca incrociata su più archivi risultava avere il problema principale di introdurre notevoli rallentamenti nel caso fosse presente tra quelli interrogati anche un solo host particolarmente lento; in più i vari host utilizzavano dei linguaggi di interrogazione differenti creando problemi di complessità per gli utenti finali e anche per i software di ricerca.

Il prototipo UPS adottato si basava sulla raccolta dei metadati da archivi multipli e sulla conseguente fornitura di servizi basati su di essi. In tal modo venne risolto il problema dell'interrogazione incrociata poiché si resero possibili richieste da rivolgere all'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" può essere o data provider, o service provider o entrambi, come ad esempio l'applicazione CDSware, che verrà trattata successivamente nell'elaborato.

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".

I fondatori dell'OAI sono la Digital Library Federation (DLF), la Coalition for Networked Information (CNI), e la National Science Foundation (NSF) e il suo successo si basa sulla partecipazione di comunità di persone provenienti da tutto il mondo, in particolare Europa e Nord America.

Da discussioni ed esperimenti successivi al meeting di Santa Fè si arrivò ben presto a definire un protocollo ed un formato di metadati stabili e universalmente riconosciuti.
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 e si basa sugli standard HTTP e XML esistenti.
Il formato di metadati raccomandato dall'OAI è il *Dublin Core Unqualified* benché comunità ristrette di data e service provider possano rappresentare i loro dati in un formato qualsiasi purché accettato dall'intera comunità.
Tali metadati vengono raccolti da diverse sorgenti (data provider) e messi

insieme in un unico archivio gestito da un *harvester* per la fornitura dei servizi in funzione dei metadati contenuti.

Sebbene il protocollo OAI dal punto di vista tecnico sia semplice, realizzare servizi che incontrino le necessità degli utenti risulta abbastanza complesso.
L'OAI-PMH 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

[B2] De Robbio Antonella, "Open Archive per la comunicazione scientifica", *Notiziario del SIMAI*, N° 5, pp. 2-6, 2002.

SITOGRAFIA

[S1] http://www.openarchives.org/