Contenuti Open Source :
nuove metodologie per la produzione in Internet di materiale accademico e per l'uso didattico
Milano, Università degli Studi, 9 dicembre 2003

Istanze "open" nella comunicazione scientifica: open archives

Susanna Mornati (*)

(*) CILEA – Consorzio Interuniversitario per l'ICT, Segrate (Milano), mornati@cilea.it

Abstract

Nel 1998 l'espressione "open-source software" comincia a circolare nell'ambiente dei programmatori. Il concetto di "free software" (nell'accezione di "libero", non di "gratuito") vanta una vita ben più lunga, tuttavia il termine "open", ideologicamente meno caricato di "free", ha avuto un successo insperato, tanto da aprire le porte dell'industria e dei servizi basati sullo sviluppo di ICT. Anche nell'ambito della comunicazione scientifica, e più specificamente della disseminazione dei risultati della ricerca tramite pubblicazioni, le nuove tecnologie hanno portato allo scoperto la crisi del modello tradizionale, basato sui costosi servizi delle multinazionali dell'editoria, ed offerto soluzioni alternative per l'"open access" tramite gli "open archives". L'infrastruttura tecnologica, basata sul protocollo OAI-PMH, rende interoperabili gli archivi aperti e prelude alla creazione di un enorme archivio mondiale distribuito della letteratura scientifica. La relazione illustrerà alcuni esempi pratici di implementazione, il "case" E-LIS presso il CILEA ed il progetto AEPIC-OA, una piattaforma nazionale per gli archivi aperti italiani.

1. Introduzione[1]: free software, open source, copyleft.

Nel 1998 l'espressione "open-source software" comincia a circolare nell'ambiente dei programmatori. Il concetto di "free software" (nell'accezione di "libero", non di "gratuito") vanta una vita ben più lunga, tuttavia il termine "open", ideologicamente meno caricato di "free", ha avuto un successo insperato, tanto da aprire le porte dell'industria e dei servizi basati sullo sviluppo di ICT. Entrambi i termini in ogni caso si riferiscono alla libertà accordata dall'autore agli utilizzatori del software di eseguirlo, copiarlo, distribuirlo, modificarlo e ridistribuire le modifiche, purché tutto ciò avvenga nel rispetto dell'unica restrizione imposta dalla licenza, ossia che ciascuna copia o modifica erediti le stesse libertà e sia accompagnata dal codice sorgente "aperto".

La GNU General Public Licence[2] della Free Sofware Foundation ha accompagnato in questo modo codici famosi, dal kernel Linux[3] a tutti i sistemi operativi e applicazioni open source. Questa singolare interpretazione del diritto d'autore è anche nota con il nome di "copyleft", con il doppio senso intraducibile di diritto "lasciato" dall'autore all'utente e di contrapposizione "left vs. right". In sostanza il copyleft non protegge i diritti economici, come il copyright si è ormai ridotto a fare quasi esclusivamente, ma la diffusione gratuita del codice, evitando che qualcuno se ne appropri per farne

[1] I link citati in questo documento sono stati visitati per l'ultima volta il 30 novembre 2003. Data l'alta densità di termini tecnici mutuati dalla lingua inglese utilizzati nel presente lavoro, si è evitato l'uso costante del corsivo per indicarli, sia per non appesantire la leggibilità, sia con il preciso intento di renderli integrati nel fluire del testo.
[2] http://www.gnu.org/copyleft/gpl.html
[3] http://www.linux.org/

prodotti proprietari, come potrebbe invece avvenire se il codice fosse di pubblico dominio e non protetto da alcun tipo di licenza. Non si tratta di impedire la commercializzazione del software open source, di fatto possibile e legittima, ma la sua trasformazione in prodotto chiuso, proprietario, che nessun altro può sviluppare e migliorare.

Questa breve premessa non vuole essere esaustiva della problematica, che viene più estesamente affrontata nel corso dell'intera giornata dedicata ai contenuti open source, ma è tuttavia necessaria per inquadrare l'analogo fenomeno che sta investendo in questi anni le modalità della comunicazione scientifica.

2. La crisi della comunicazione scientifica e l'Open Access.

Anche la crisi della comunicazione scientifica verrà qui affrontata rapidamente, perché già illustrata da altri interventi. Ci limitiamo ad osservare come anche in quest'ambito, e più specificamente in quello della disseminazione dei risultati della ricerca tramite pubblicazioni di articoli in riviste scientifiche, gli autori, che non scrivono per il profitto ma per raggiungere un pubblico più vasto possibile, cedono di fatto i diritti economici di sfruttamento dei loro articoli agli editori, che impongono alle biblioteche costi altissimi per l'accesso. Poche istituzioni possono tuttavia permetterselo, e questa barriera riduce di fatto il numero dei lettori, salvaguardando i profitti economici degli editori ma causando un danno diretto agli interessi degli autori[4]: è già stato dimostrato che gli eprints liberamente accessibili in rete vengono letti e citati molto più degli articoli pubblicati esclusivamente su riviste che richiedono il pagamento di una quota di sottoscrizione[5], influenzando i criteri di valutazione delle carriere dei ricercatori e delle richieste di finanziamento.

In questo paradossale contesto, l'avvento delle nuove tecnologie ha portato allo scoperto la crisi del modello tradizionale, basato sui costosi servizi delle multinazionali dell'editoria, ed ha offerto soluzioni alternative per garantire una più efficace disseminazione della letteratura scientifica, tramite varie strategie, complementari e non concorrenti, di "open access"[6].

Quello di cui tratterò in questa sede è la soluzione offerta dagli "open archives", che coniugano l'impiego di tecnologia open source con il libero accesso ai risultati della ricerca scientifica, con nuove forme di analisi citazionale e indicatori di performance indipendenti dal monopolio dell'Impact Factor[7], con la libera distribuzione a fini non di

[4] Peter Suber, "Removing the Barriers to Research: an Introduction to Open Access for Librarians", in *College & Research Libraries News*, 64 (February 2003) pp. 92-94, 113, testo disponibile all'URL: http://www.ala.org/Content/NavigationMenu/ACRL/Publications/College_and_Research_Libraries_News/Back_Issues__2003/February1/Removing_barriers_to_research.htm.

[5] Steve Lawrence,"Free online availability substantially increases a paper's impact", in *Nature*, 411, 6837, p. 521, 2001, testo noto anche con il titolo "Online or Invisible?" e disponibile al'URL: http://www.neci.nec.com/~lawrence/papers/online-nature01/

[6] Stevan Harnad, "For Whom the Gate Tolls: How and Why to Free the Refereed Research Literature Online Through Author/Institution Self-Archiving, Now", 2001, testo disponibile all'URL: http://www.ecs.soton.ac.uk/~harnad/Tp/resolution.htm.

[7] detenuto da una società privata influenzata dalle multinazionali dell'editoria.

lucro, preservandone i contenuti originali, veicolati da tecnologie digitali, tramite licenze innovative: di fatto, applicando il copyleft[8].

3. Gli Open Archives e il protocollo per l'interoperabilità.

Da oltre dodici anni esistono in rete archivi[9] di eprint, ossia le copie digitali degli articoli prodotti dai ricercatori per la comunicazione dei risultati delle proprie ricerche. Con eprint si intende ogni versione, dal pre-print (l'articolo nella sua versione iniziale) al post-print (l'articolo nella sua versione già sottoposta alla procedura di valutazione da parte di altri ricercatori, detta peer-review o referaggio). Tali archivi prevedono il deposito direttamente da parte dell'autore dell'eprint, accompagnato da una breve descrizione dei metadati essenziali per l'identificazione, quali autore e titolo.

Gli archivi possono essere variamente raggruppati. Possono distinguersi in istituzionali o disciplinari, a seconda che contengano gli eprint prodotti dai ricercatori di una determinata università o ente di ricerca, oppure che raggruppino gli eprint prodotti dai ricercatori che costituiscono la comunità di riferimento per una particolare disciplina, ovunque si trovino ad operare. Possono essere ad architettura centralizzata o distribuita, ossia prevedere il deposito dei lavori su un unico server o su server remoti collegati da un'unica interfaccia di ricerca[10].

Contrariamente a quanto poteva essere previsto intorno alla metà degli anni novanta, quando il World-Wide Web ha iniziato la sua espansione nel mondo al di fuori dell'ambiente della ricerca, il fenomeno della distribuzione parallela di eprint negli archivi e articoli corrispondenti nelle riviste referate non ha avuto la diffusione che il potenziale legato alle nuove tecnologie lasciava sperare[11]. Cosa mancava agli archivi di eprint per catalizzare l'attenzione dei ricercatori?

Nel 1999 nasce a Santa Fe l'Open Archives Initiative[12], con l'intento di sviluppare software open source per l'implementazione semplice e a basso costo di archivi di eprint, o repository (denominati "data provider"), definire un protocollo di comunicazione fra gli archivi che li rendesse interoperabili, dunque "aperti" alla raccolta, ed incoraggiare la creazione di servizi centralizzati (denominati "service

---

[8] Per questo lavoro sono debitrice di molte idee ispiratrici ad Antonella De Robbio, dell'Università di Padova, ed in particolare al suo articolo "Auto-archiviazione per la ricerca: problemi aperti e sviluppi futuri", in *Proceedings* "*Comunicazione scientifica ed editoria elettronica: la parola agli Autori: L'Utente-Autore nel circuito della comunicazione scientifica: editoria elettronica e valutazione della ricerca"*, Milano, 20 Maggio 2003, URL: http://www.cilea.it/convegni/convegnoeditoria/presentazione.html, in corso di pubblicazione, testo disponibile all'URL: http://eprints.rclis.org/archive/00000180/. Naturalmente gli eventuali errori qui contenuti sono invece mia esclusiva responsabilità.

[9] ArXiv, http://arxiv.org/, ora ospitato dalla Cornell University, prima noto con il nome di xxx.lanl.gov, è stato fondato nel 1991 da Paul Ginsparg a Los Alamos per la comunità dei fisici delle alte energie. Oggi include collezioni di altre branche della fisica, oltre alla matematica, l'informatica e la biologia quantitativa.

[10] Mentre arXiv rappresenta il prototipo di server centralizzato, RePEc (http://repec.org/, dedicato agli eprint degli economisti) ha un'architettura distribuita. Entrambi sono server disciplinari: per ovvi motivi (accettando solo lavori provenienti dai propri ricercatori) gli archivi istituzionali non possono raggiungere la stessa notorietà.

[11] Susanna Mornati, "La costruzione delle basi di dati: l'esperienza dei preprint server per la fisica", in *Proceedings "Documentazione: professione trasversale : 5° Convegno Nazionale AIDA. Fermo: Palazzo dei Priori, 23-25 ottobre 1996"*, Roma, CNR-ISRDS, 1998, testo disponibile all'URL: http://www.aidaweb.it/5convegno96/mornati96.html.

[12] http://www.openarchives.org/, origini e obiettivi descritti in: Herbert Van de Sompel e Carl Lagoze, "The Santa Fe Convention of the Open Archives", in D-LIB Magazine, 6 (2), 2000, testo disponibile all'URL: http://www.dlib.org/dlib/february00/vandesompel-oai/02vandesompel-oai.html.

provider") per la raccolta (harvesting), l'indicizzazione e il recupero dei metadati associati ai documenti, nonché il link per l'accesso al full-text.

Gli sviluppi del movimento portano alla creazione dell'infrastruttura tecnologica basata sul protocollo OAI-PMH[13], che rende interoperabili gli archivi aperti, e prelude alla creazione di un enorme archivio mondiale distribuito della letteratura scientifica. Nella definizione di "open archives", il termine "open" indica proprio la compatibilità dei metadati descrittivi con il protocollo di interoperabilità.

4. Il "case" E-LIS.

Nell'ambito dell'Open Archives Initiative è stato sviluppato un software per la gestione di archivi aperti, GNU Eprints[14], realizzato in modalità open source presso l'Università di Southampton. E' costituito da un DBMS relazionale MySQL e una serie di script CGI in linguaggio Perl che si appoggiano ad un server http Apache per le funzioni di deposito, amministrazione, ricerca e recupero dei dati. La leggerezza e la facilità di installazione, gestione ed uso lo rendono molto popolare, ed in pochi mesi decine di installazioni sono state registrate in tutto il mondo[15].

Presso il CILEA è stato realizzato E-LIS[16], un archivio disciplinare specializzato per il deposito di documenti di ambito biblioteconomico e bibliografico, oltre che di scienze dell'informazione e della comunicazione. L'interfaccia è in lingua inglese, ma accetta contributi in tutte le lingue. Installato in ambiente Linux Red Hat su server CILEA[17], è nato e viene mantenuto per iniziativa di alcuni membri del team DOIS[18].

Oltre alle funzionalità di base previste dal software EPrints, che prevede un'area di registrazione, deposito e alerting per gli utenti registrati, e un'area di search, browse e visualizzazione dei depositi più recenti per tutti gli utenti della rete, E-LIS ha implementato un contatore dei record e l'estrazione dei riferimenti bibliografici mediante ParaTools[19]. L'archivio è inoltre corredato da un sistema di classificazione dei lavori, con un unico livello gerarchico per facilitare l'inserimento dei metadati, e di una lista arricchita per definire la tipologia del materiale.

Di particolare interesse lo sviluppo delle policies relative al copyright, tanto da meritare una citazione sulle pagine del progetto RoMEO[20], un'iniziativa britannica per lo studio delle questioni legate al copyright che interessano la comunità accademica. Uno degli ostacoli allo sviluppo degli Open Archive è infatti costituito dalle politiche restrittive di certi editori, che pretendono la cessione completa dei diritti d'autore per lo sfruttamento economico della produzione intellettuale, privando gli autori degli

---

[13] The Open Archives Initiative Protocol for Metadata Harvesting, giunto nel giugno 2002 alla versione 2.0: http://www.openarchives.org/OAI/openarchivesprotocol.html.
[14] http://www.eprints.org/
[15] http://software.eprints.org/archives.php
[16] http://eprints.rclis.org/. Per una presentazione completa dell'archivio, delle sue origini, obiettivi e funzionalità cfr. Antonella De Robbio, "E-LIS: un Open Archive in Library and Information Science", in *Bibliotime*, VI (I), marzo 2003, testo disponibile all'URL: http://eprints.rclis.org/archive/00000201/.
[17] http://www.cilea.it/
[18] http://dois.mimas.ac.uk/team.html
[19] http://paracite.eprints.org/: ParaTools (ParaCite Toolkit) è un set di moduli Perl per la gestione delle citazioni, che comprendono il reference parsing e routines di creazione e gestione di OpenURL.
[20] http://www.lboro.ac.uk/departments/ls/disresearch/romeo/

articoli della possibilità di condurre una disseminazione dei risultati delle proprie ricerche attraverso altri canali, fra cui il self-archiving sul sito istituzionale. Il progetto ha condotto una serie di indagini ed è arrivato a identificare gli editori che consentono il deposito di una copia dell'articolo (pre-print o post-print o entrambe) in un open archive, e quelli che lo negano. I primi costituiscono il 42,5 % del totale, ma rappresentano la maggioranza dei titoli delle riviste indagate (oltre il 54 %), i secondi il 57,5 % (con il 45 % dei titoli), ma di fatto accettano una negoziazione dei diritti.

I motivi che inducono a sottolineare la cura con cui gli editors di E-LIS hanno affrontato questo aspetto sono dovuti non solo all'importanza che le questioni relative agli IPR (intellectual property rights) rivestono nel mondo degli Open Archives, ma anche al tema della giornata, suggerendo l'opportunità di dotare anche la produzione di letteratura scientifica di licenze che garantiscano il copyleft, quali quelle disponibili sul sito di Creative Commons[21].

5. Oltre gli Open Archives: il progetto AEPIC – Open Archives.

Si è già detto che la forza dell'infrastruttura tecnologica che caratterizza gli open archives sta nell'interoperabilità garantita dall'adozione del protocollo OAI-PMH, che consente la raccolta di metadati da parte dei cosiddetti Service Provider, allo scopo di offrire un'unica interfaccia di ricerca ed altri servizi aggiuntivi. Ad oggi tuttavia lo sviluppo su questo versante è ancora scarso e i documenti depositati negli archivi restano isolati o vengono esposti attraverso servizi estremamente generici.

Per ovviare in parte a questo problema ed offrire maggiore visibilità alla produzione di letteratura scientifica in Italia, il CILEA sta realizzando un progetto per la costruzione di una piattaforma nazionale di accesso agli open archives, nella cornice del più vasto progetto AEPIC[22], dedicato all'editoria elettronica per l'ambiente accademico.

L'architettura della piattaforma è strutturata in due oggetti principali: un insieme di Service Provider e un portale. I metadati e, dove necessario, i documenti a testo pieno dei lavori scientifici, saranno raccolti dai Data Provider disponibili ed appropriati - possibilmente dedicati al deposito delle pubblicazioni accademiche e di ricerca prodotte a livello nazionale, sia in campo scientifico-tecnico sia umanistico/giuridico - mediante l'impiego del protocollo OAI-PMH, quindi verranno sottoposti ad una serie di operazioni:
- attribuzione di una classificazione per soggetto a livello generale, basata sulla suddivisione in aree di ricerca operata dal Ministero[23];
- funzionalità centralizzate di caching e indicizzazione;
- funzionalità avanzate di ricerca, scorrimento e recupero delle informazioni;
- analisi citazionale operata nel testo del documento per la registrazione presso servizi di indicizzazione delle citazioni[24], nonché per l'estrazione automatica di OpenURL[25], che possano essere utilizzati da servizi di *resolving* esistenti;
- allestimento di *crosswalk* per la conversione, ai fini dell'importazione/esportazione da/verso basi di dati che adottano differenti standard per i metadati;

[21] http://creativecommons.org/
[22] http://www.aepic.it/
[23] http://www.murst.it/atti/2000/alladm001004_01.htm.
[24] Ad esempio CiteSeer, noto anche come ResearchIndex, vedi: http://citeseer.nj.nec.com/
[25] http://library.caltech.edu/openurl/Public_Comments.htm

- servizio di *gateway* per *web crawlers*;
- funzionalità di ricerca nel testo pieno;
- statistiche di accesso;
- funzionalità di esportazione per la costruzione di report o pagine *Web* individuali o istituzionali, a fini di presentazione o valutazione;
- implementazione di un server Z39.50 per offrire la configurazione di servizi di *discovery* da inserire fra le risorse di sito, ad esempio cataloghi di biblioteche, delle istituzioni partecipanti;
- certificazione temporale, protezione dei diritti d'autore, deposito legale, indirizzo permanente e conservazione dei dati.

Attraverso un'interfaccia di ricerca, il portale offrirà un punto di accesso singolo ai dati centralizzati e all'informazione rilevante nel contesto OAI, oltre ad una varietà di altri servizi personalizzati, basati su un sistema di gestione dei profili di accesso, quali *alerting*, *newsfeed*, e così via. Il sito del portale ospiterà anche una lista delle iniziative italiane in ambito OAI, un *testbed* per gli strumenti ad uso degli sviluppatori, un *virtual reference desk* per gli implementatori e un *forum* per la condivisione a livello nazionale di linee guida relative all'implementazione di archivi aperti.

6. Problemi aperti e conclusioni.

Fra i contenuti originali veicolati da tecnologie digitali e messi liberamente a disposizione in rete, possiamo dunque comprendere la letteratura scientifica depositata negli open archives. Molta strada resta tuttavia ancora da percorrere.

Le università e gli enti di ricerca devono modificare i criteri di valutazione della ricerca, adottare politiche sistematiche di archiviazione sul sito istituzionale ed avviare con gli editori una negoziazione collettiva dei diritti. Occorre dotare ogni documento di una licenza di copyleft, definire un set di metadati descrittivo dei termini di accessibilità e costruire strumenti di DRM (digital rights management) che utilizzino i metadati espressi in formati machine-readable.

Sul fronte dei servizi, la collaborazione anche a livello internazionale si è rivelata un fattore essenziale per la diffusione degli archivi aperti. I punti critici comprendono sia i costi a carico dello sviluppo di progetti complessi quale AEPIC-Open Archives, sia la scarsa integrazione delle risorse digitali - dai periodici elettronici alle immagini, dagli strumenti per la didattica multimediale al cartaceo digitalizzato, dalle tesi agli OPAC - gestite da una molteplicità di protocolli e piattaforme che non garantiscono il colloquio reciproco.

La strada è lunga e irta di ostacoli, parallela al cammino dei contenuti e degli strumenti open source, ma l'auspicio è che possa essere percorsa sino in fondo e che possa contribuire ad una più ampia diffusione della conoscenza e ad un più rapido progresso dell'umanità.