

Relazione tecnica su *DSpace* (MIT):

Antonella De Robbio

[documento 23-10-2002]

Descrizione di *DSpace*

DSpace è il prodotto sviluppato nel 2000 entro un progetto congiunto dalle biblioteche del MIT (Massachusetts Institute Technology) e dalla Hewlett-Packard Company per la costruzione di un repository (deposito) a lungo termine, stabile e sostenibile, che consenta la cattura, la conservazione e la disseminazione della ricerca e dei materiali didattici, generati lungo lo spettro delle discipline al MIT, consentendo accesso remoto via web da qualsiasi punto della rete.

DSpace offre i vantaggi di una distribuzione e conservazione digitale per una vasta gamma di formati inclusi: testi, audio, video, immagini, programmi e datasets. Tutti i materiali possono essere soggetti a restrizioni d'accesso ai fini del copyright elettronico. E' possibile determinare, da parte di singoli o di comunità accademiche, chi avrà accesso alle opere archiviate.

Inoltre *DSpace* offre l'aggancio a versioni successive dei documenti entro il deposito con possibilità di referaggio (peer-review).

Le due configurazioni primarie del sistema *DSpace* si basano sui due presupposti distinti: il primo si riferisce all'aspettativa posta sugli autori che, attraverso l'auto-archiviazione dei propri lavori, forniscono direttamente i metadati alimentando la base di dati. Il secondo è basato sulla considerazione che l'archivio consente di gestire una grande varietà di formati e tipi di oggetti digitali. Infatti da una parte *DSpace* nasce con lo scopo di fornire record sostenibili riferibili alla produzione intellettuale dei ricercatori afferenti alle facoltà del MIT, produzione contenuta nei documenti digitali, dall'altra ha il preciso obiettivo di offrire un valido supporto alla gestione della produzione multimediale didattica, in continua crescita, entro un ambiente configurabile come e-learning (in particolare nella formazione a distanza).

Il software *DSpace* sarà reso disponibile pubblicamente a partire dal 4 novembre 2002, attraverso licenza BSD OpenSource, la quale consente ad altre istituzioni di utilizzare il software *DSpace* per creare propri archivi per il deposito.

La piattaforma *DSpace*

DSpace è una piattaforma software che rende in grado le istituzioni di operare entro un sistema elettronico che consente di:

- ❑ **catturare** materiale di ricerca in vari formati direttamente dai creatori (via facoltà)
- ❑ **descrivere** opere digitali, usando un modulo apposito per la sottomissione dei dati, attraverso l'utilizzo di metadati descrittivi, tecnici e sui diritti (*rights metadata*)
- ❑ **distribuire e comunicare** i contenuti delle opere digitali dell'istituzione via web attraverso un sistema di recupero dell'informazione e, dove necessario, in accesso controllato
- ❑ **conservare** le opere digitali per un periodo a lungo termine

Le caratteristiche di *DSpace* si possono riassumere in quattro punti cardine:.

- ❑ Deposito istituzionale
- ❑ Database per il materiale in formato digitale
- ❑ Luogo per un'archiviazione cumulativa e perpetua
- ❑ Archivio aperto e interoperabile

A livello comparativo tra sistemi esistenti ad oggi entro le biblioteche digitali, **DSpace** assume l'identità di un modello esteso di archivio di tipo e-print su base disciplinare entro un'organizzazione ad archivio di tipo istituzionale, in cui tutti i formati sono accettati e la conservazione diviene paradigma stesso della comunità.

Tale modello differisce da quello di una biblioteca digitale del tipo Greenstone (DL della Nuova Zelanda) o FEDORA, in quanto i contenuti provengono dalle singole facoltà e non attraverso le tradizionali acquisizioni della biblioteca con una responsabilità dei contenuti distribuita.

Per questa ragione **DSpace** necessita di politiche di selezione, sottomissione, politiche sui dati e metadati chiare e conformabili all'istituzione in cui viene collocato.

L'architettura di **DSpace**

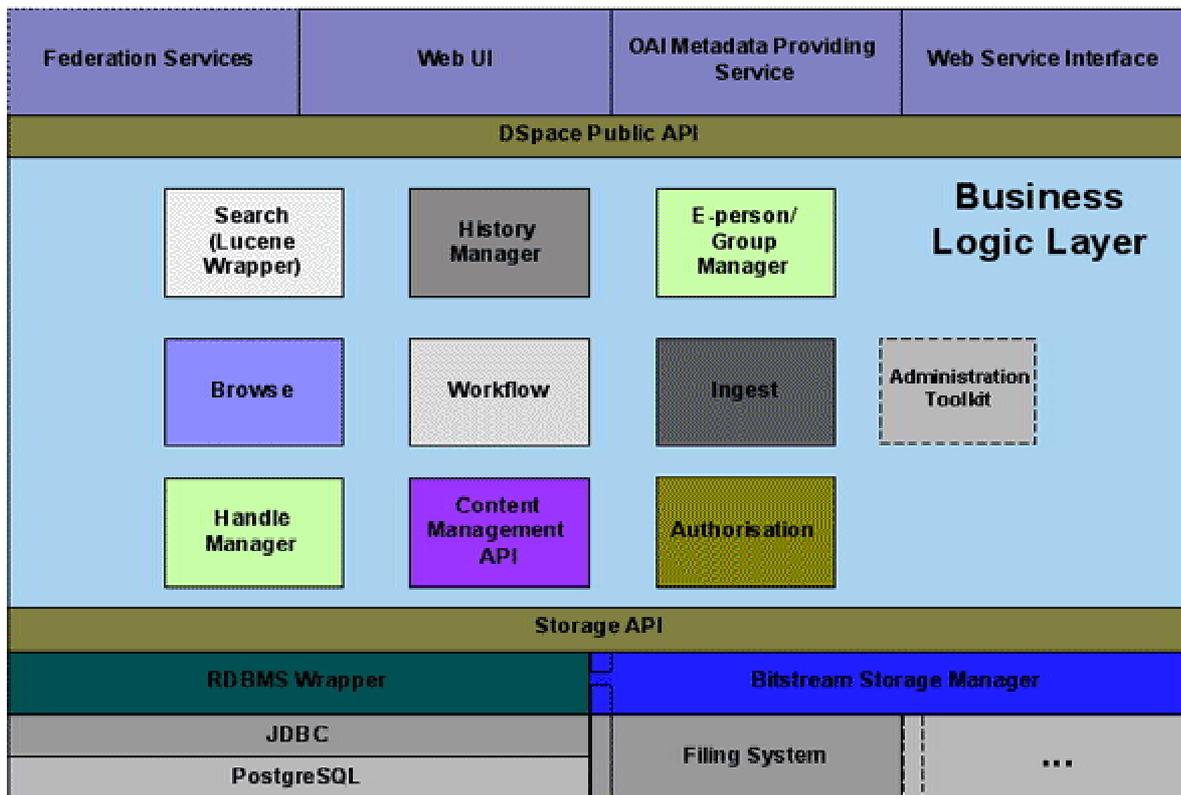
Di seguito la scheda tecnica e dei requisiti

- ❑ Architettura modulare, ben definita APIs
- ❑ 100% open source
 - ❑ Programmato in java
 - ❑ Sistema di gestione base di dati relazionale (RDBMS) e SQL per i metadati
- ❑ CNRI “handles” per gli identificatori permanenti
- ❑ Controllo degli accessi basato su X.509
- ❑ OpenURL linking
- ❑ OAI-PMH per l'esposizione dei metadati
- ❑ Compatibilità OAIS

Requisiti tecnologici:

- ❑ Apache, Tomcat, OpenSSL/mod_ssl
- ❑ Java 1.3, JSP 1.2, Servlet 2.3
- ❑ PostgreSQL 7, JDBC (rdbms)
- ❑ CNRI Handle System 5 (persistent ids)
- ❑ Lucene 1.2 (index/search)
- ❑ Jena (RDF History system)
- ❑ JUnit (testing), Log4j (logging)
- ❑ HP/UX, Linux, Solaris, etc.

Di seguito lo schema dell'architettura di **DSpace**:



Il modello informativo

DSpace è organizzato in modo da accomodare i bisogni organizzativi e multidisciplinari di grandi istituzioni accademiche, offrendo accesso alle opere digitali dell'intera istituzione attraverso un'unica interfaccia.

La sua organizzazione informativa è suddivisa entro due canali, "Comunità" e "Collezioni" ciascuno dei quali mantiene la propria identità entro l'archivio (deposito).

Le "Comunità" sono i dipartimenti, i laboratori, i centri di ricerca, i programmi, le scuole, etc...le quali localizzano le proprie politiche entro la piattaforma ai fini dell'accesso del materiale che esse depositano e ai fini dei processi di validazione e certificazione dei contenuti (approvazione, recensione, lavoro editoriale ...).

La gestione e definizione delle collezioni viene decisa a livello delle politiche di "Comunità" le quali forniscono i propri formati di metadati.

Da questo modello informativo organizzato attraverso le politiche di "Comunità" si costruiscono e si sviluppano "Collezioni" o aggregazioni di collezioni di differente natura e tipologia e servizi ad alto valore aggiunto:

- ❑ Collezioni tecniche
- ❑ Collezioni virtuali
- ❑ Biblioteche digitali in rete per Tesi o Dissertazioni
- ❑ Indici OAI disciplinari
- ❑ Nuovi modelli di pubblicazioni elettroniche
- ❑ Servizi distribuiti per la disseminazione della ricerca

In pratica si attua il modello disaggregato in cui trovano luogo le quattro funzioni del modello della comunicazione scientifica espresse da Roosendaal e Guerts: registrazione, validazione, consapevolezza, archiviazione.

La personalizzazione sia dei canali "Comunità" e "Collezioni" consente una flessibilità in determinate politiche e per determinati processi che supporta una grande varietà di formati digitali e tipologie di contenuto differenti tra i quali:

- ❑ testi
- ❑ immagini
- ❑ audio
- ❑ video

Alcuni esempi di documenti che *DSpace* può accogliere sono:

- ❑ articoli, preprint,
- ❑ working paper,
- ❑ rapporti tecnici,
- ❑ conference paper,
- ❑ libri,
- ❑ tesi elettroniche,
- ❑ insiemi di dati (datasets): statistici, goespaziali, matlab, ...
- ❑ immagini: visuali, scientifiche, ...
- ❑ file audio,
- ❑ file video,
- ❑ programmi,
- ❑ modelli e simulazioni visuali
- ❑ oggetti per l'apprendimento (Learning Objects)
- ❑ collezioni digitali bibliotecarie riformattate (convertite dal cartaceo)

La conservazione digitale entro *DSpace*

DSpace fornisce immagazzinamento fisico a lungo termine dei documenti digitali in un deposito sicuro, gestito a livello professionale attraverso utilizzo di procedure operative standardizzate come il back-up, il mirroring, il refreshing dei multimediali, e il recupero dei dati per incidenti vari.

Altro aspetto assai interessante è l'assegnazione a ciascun documento di un identificatore persistente al fine di assicurare una sua recuperabilità nel tempo.

DSpace fornisce inoltre un meccanismo per l'avviso ai contributori dei contenuti depositati, sui livelli di conservazione dei supporti per i file che si sottomettono.

La conservazione digitale entro *DSpace* supporta la migrazione dei testi, immagini, audio,... e l'emulazione per software e multimediali.

Inoltre è previsto il registro dei formati digitali

Il modulo per la sottomissione dei dati

Ci si aspetta che la massa critica delle sottomissioni regolari e interne entro *DSpace* avvenga su base individuale, dal creatore stesso dell'oggetto di contenuto o attraverso il supporto di uno staff di bibliotecari della comunità accademica.

Durante il processo di sottomissione viene proposto a chi effettua il deposito il riempimento del form per i metadati al fine di offrire informazioni minimali che descrivono il documento depositato.

Il processo di sottomissione entro **DSpace** permette la descrizione di ciascun documento usando lo schema di metadati Dublin Core. Tutte le descrizioni sono immesse entro un database relazionale il quale è utilizzato dal motore di ricerca per il recupero dei documenti

Lo schema di metadati per **DSpace** dovrebbe essere abbastanza semplice anche per un utente non catalogatore e sufficientemente comprensibile da catturare gli elementi informativi necessari ad un adeguato recupero del documento.

L'input dei metadati avviene attraverso l'utilizzo di template.

DSpace permette ai contributori di limitare l'accesso a determinati documenti o a collezioni intere.

E' possibile depositare nuove versioni per documenti precedentemente sottomessi che possono essere aggiunti e connessi l'uno con l'altro.

Formati multipli per uno stesso documento possono essere depositati, per esempio file TIFF e file GIF per una stessa immagine.

In **DSpace** è possibile effettuare sottomissioni batch per documenti convertiti dal formato cartaceo ad un formato digitale, per esempio collezioni di immagini PDF per volumi fuori stampa o per immagini scansionate di report tecnici o serie di working paper provenienti dai vari laboratori, o per dispense o materiale didattico da mettere a disposizione degli studenti dei vari corsi.

La natura multidisciplinare di DSpace

Nel decidere sugli schemi per l'utilizzo di **DSpace** è importante considerare la tipologia di contenuto che ci si aspetta di ricevere ed elaborare entro il sistema, sia a breve sia a lungo termine.

Poiché **DSpace** è un deposito di tipo istituzionale, esso dovrà poter accogliere le numerose discipline che rappresentano tutte le scuole, i dipartimenti, i laboratori e i centri dell'istituzione.

Inoltre in un ambiente reso ancor più eterogeneo dalla presenza di numerosi formati differenti, è assai improbabile che un solo schema di metadati descrittivi possa andar bene per tutti i contenuti depositati dalle varie "Comunità". Risulta anche impossibile stabilire a priori quali potranno essere le tipologie di contenuto predominanti nel futuro. Le comunità che producono e collezionano differenti tipologie di oggetti informativi devono usare differenti schemi di metadati per soddisfare le proprie necessità di didattica e/o ricerca.

I metadati in DSpace

In un ambiente fatto di collezioni eterogenee è inoltre assodato che esistono già metadati specifici per specifiche collezioni, per esempio i LOM Learning Object Metadata per la didattica, metadati catalografici e record MARC per documenti digitalizzati dopo scansione, metadati per le singole comunità (matematici, chimici ...), Dublin Core per alcuni file testuali, schema FGDC per dati geospaziali, VRA per le immagini digitali e così via ...

Alla luce delle configurazioni di **DSpace**, fu creata, all'epoca della creazione del prodotto, una task force di bibliotecari al MIT per la scelta di un set Dublin Core Metadata Element con qualificatori per i metadati **DSpace** nell'applicazione al MIT. La scelta fu basata su LAP Libraries Application Profile. La task force creò anche un crosswalk dal formato MARC al Dublin Core sulla base di quello della Library of Congress.

Attualmente **DSpace** utilizza, come punto di partenza, una versione qualificata dello schema Dublin Core basato sul Dublin Core Libraries Working Group Application Profile (LAP).

Vedi tabella:

<http://DSpace.org/technology/metadata.html>

Osservazioni conclusive

Possibili sfide risiedono entro i seguenti issues:

- ❑ Accettazione del prodotto da parte del sistema accademico (facoltà e comunità di utenti)
- ❑ Valutazione e certificazione di un archivio istituzionale (data provider) con possibili ricadute a livello di archivi disciplinari a livello internazionale (service provider)
- ❑ Miriadi di discipline differenti per storia e cultura
- ❑ Politiche del copyright e sostenibilità di tali politiche nel tempo
- ❑ Risorse finanziarie
- ❑ Conservazione digitale

Senza dubbio il prodotto sembra assai allettante, in particolare in virtù del fatto di essere un software libero. E' necessario valutare il suo impatto sull'attuale ambiente informativo di Ateneo e la sua integrazione con gli attuali archivi e collezioni presenti entro lo SBA, soprattutto nell'ottica di un'eventuale acquisizione da parte dello SBA di un portale-gateway che funga da MetaSearch per una ricerca incrociata entro archivi differenti.

Va attentamente vagliata la possibilità di utilizzarlo come strumento per la costruzione di un Object Learning Repository. Va inoltre considerato che *DSpace* implementa funzionalità di Open Archive per la ricerca e quindi questa sua configurazione lo renderebbe un prodotto completo, orientato alle due funzioni chiave dell'Ateneo: ricerca e didattica.

Ma è necessario indagare meglio su questi aspetti, soprattutto in relazione alla sperimentazione già condotta sul software Eprints entro il nostro Ateneo, che risulta un eccellente strumento per la creazione di archivi digitali e-prints per la ricerca.

Eprints offre notevoli potenzialità, flessibilità di configurazione e caratteristiche di amichevolezza, oltre ad una certa leggerezza nei software utilizzati. Il supporto sia via liste di discussione, sia come documentazione disponibile è altamente professionale e di sicura efficacia.

Va verificato se anche *DSpace* offre gli stessi servizi di supporto e se l'ambiente risulta agevole e sufficientemente leggero sia per l'utenza sia per lo staff.

Riferimenti bibliografici

Smith MacKenzie, MIT Dspace

<http://documents.cern.ch/cgi-bin/setlink?base=agenda&categ=a02333&id=a02333s13t2/transparencies>

DSpace Home

<http://www.dspace.org/>

DSpace Technology and Architecture

<http://dspace.org/technology/architecture.pdf>

DSpace Internal Reference Specification

<http://dspace.org/technology/functionality.pdf>