

Harvesting delle tesi di dottorato delle Biblioteche Nazionali tramite DSpace

Versione: 1.0 (14 settembre 2010)

Autore: Andrea Bollini, Nilde De Paoli



Premessa

Scopo di questo documento è illustrare la modalità di configurazione del software Open Source DSpace per il corretto harvesting di tesi di dottorato in formato elettronico da parte delle Biblioteche Nazionali. Cilea è il maggior service provider Italiano per il software DSpace e, in collaborazione con alcune Università di cui gestisce le installazioni, ha analizzato e realizzato l'esposizione dei metadati delle tesi di dottorato come richiesto dalle Biblioteche Nazionali. Le configurazioni apportate da Cilea rendono accessibili alle Biblioteche Nazionali anche i file non consultabili pubblicamente, ossia coperti da embargo sia temporaneo che illimitato.

Caratteristiche di DSpace per la gestione dell'harvesting

Come indicato nel documento *“Consegna alle Biblioteche Nazionali delle Tesi di Dottorato in formato digitale indicazioni tecniche per la raccolta automatica (harvesting)”* prodotto dalla Biblioteca Nazionale Centrale di Firenze, Biblioteca Nazionale Centrale di Roma e Fondazione Rinascimento Digitale, il software DSpace può essere predisposto per generare strutturalmente i metadati in formato MPEG21DIDL ed esporli tramite protocollo OAI-PMH. Questo formato è stato prescelto dalle Biblioteche Nazionali in quanto permette di rappresentare risorse digitali anche composte da molteplici file.

Il Cilea ha quindi analizzato l'esposizione dei metadati in formato MPEG21-DIDL dai repository di alcune Università che utilizzano DSpace di cui gestisce l'installazione e, in collaborazione con queste ultime, ha stabilito le specifiche da seguire per rendere disponibili i file come richiesto.

SUR+ OA estende la gestione dell'embargo di DSpace da parte dell'autore ad ogni singolo file di un singolo record. Per cui permette di harvestare le tesi mantenendo le politiche di accessibilità decise dall'autore per ogni file.

Inoltre, grazie all'integrazione con i software delle segreterie dell'Università che gestiscono le tesi di dottorato realizzata da Cilea, si ha la sicurezza di esporre i metadati che sono stati validati a livello istituzionale.

L'uscita DIDL è automaticamente sincronizzata con la configurazione oai_dc.

Esposizione tramite protocollo OAI-PMH

I record esposti via OAI-PMH necessitano un identificativo univoco. Generalmente questo è legato alla “base” URL del repository da cui provengono. DSpace genera automaticamente l'identificativo OAI ricavandolo dalla base url di installazione. Quindi non sono necessarie ulteriori implementazioni alla configurazione standard.

Configurazione di DSpace per l'uscita MPEG21-DIDL

La configurazione dell'uscita dei metadati in formato, MPEG21-DIDL sull'installazione DSpace dell'Università che ne fa richiesta, verso la BNCF viene effettuata da Cilea. Il formato MPEG21-DIDL permette di affiancare alla descrizione bibliografica delle tesi informazioni puntuali rispetto ad oggetti digitali collegati. In particolare consente di esporre il link diretto per il download dei full-text per

DSpace e l'harvesting delle tesi di dottorato per BNCF

l'harvesting da parte di BNCF e di rappresentare i documenti complessi come le risorse digitali composte da più di un file.

Gli step necessari alla "messa in opera" della funzionalità richiedono anche alcune operazioni da parte dei responsabili dei repository delle Università. Di seguito vengono elencati gli interventi da effettuare dopo la configurazione dell'uscita MPEG21-DIDL realizzata da Cilea (tra parentesi viene specificato chi deve eseguire l'operazione):

1. (UNIVERSITA') creazione di una policy di BITSTREAM_DEFAULT_READ per ogni collezione in cui verranno inserite le tesi per il gruppo BNCF. Le collezioni coinvolte dovranno essere harvestabili senza tener conto delle singole opzioni di accesso decise per i file in sede di submission (le istruzioni per effettuare questa operazione sono riportate nel paragrafo seguente)
2. (UNIVERSITA') comunicazione a Cilea dei singoli indirizzi ip oppure del range di ip che saranno associati al gruppo DSpace BNCF
3. (CILEA) associazione in configurazione degli indirizzi ip comunicati e breve restart del servizio

Solo gli item che al momento di essere pubblicati nel repository sono associati a una collezione con policy adeguate all'harvesting da parte di BNCF vengono automaticamente esposti. Eventuali item che dovessero essere pubblicati prima che la collezione di cui fanno parte non fosse configurata adeguatamente dovranno essere "sbloccati" per l'harvesting manualmente, utilizzando la funzionalità web di gestione delle autorizzazioni del singolo item. In questi casi occorrerà creare policy di READ per il gruppo BNCF per ogni file che si vuole rendere visibile (la procedura di dettaglio ricalca quella riportata nel paragrafo successivo per le collezioni). L'uscita DIDL è automaticamente sincronizzata con la configurazione oai_dc.

Step per la creazione delle policy di accesso

Per creare le policy corrette per le collezioni che saranno harvestate da BNCF, gli amministratori dei repository dell'Università devono effettuare i seguenti passaggi.

1. entrare in edit sulla collezione da abilitare all'harvesting "completo" e premere su 'modifica'
2. accedere al pannello di gestione delle autorizzazioni della collezione (pulsante "Autorizzazioni relative alla collezione" in coda alla pagina di edit)
3. creare una nuova policy (pulsante "Aggiungi un nuovo profilo di autorizzazioni")
4. selezionare, nella pagina di definizione della policy, BNCF come gruppo e DEFAULT_BITSTREAM_READ come azione e quindi salvare la nuova policy

Procedura dell'harvesting da parte di BNCF

L'Università deve chiedere a BNCF, scrivendo direttamente al Direttore, la disponibilità a effettuare l'harvesting delle proprie tesi di dottorato. Per attivare l'harvesting è sufficiente utilizzare l'applicativo presente sul sito <http://register-oai.depositolegale.it/> o inviare una mail oai@depositolegale.it per registrare l'url oai-pmh del proprio repository.

La BNCF esegue la procedura di raccolta una volta al mese in maniera incrementale. Come conferma dell'avvenuta ricezione delle tesi, BNCF si incarica di inviare una mail con allegato due file, in formato xml e xls, contenenti la lista delle URI delle tesi depositate.

Configurazione tecnica di DSpace per l'utilizzo del plugin di autenticazione tramite IP

In DSpace si può abilitare l'autenticazione di alcuni IP aggiungendo il metodo nella coda della configurazione di DSpace, ad esempio:

```
plugin.sequence.org.dspace.authenticate.AuthenticationMethod =  
org.dspace.authenticate.IPAutentication
```

Dopodichè sarà possibile mappare i gruppi di DSpace agli IP in dspace.cfg settando authentication.ip.GROUPNAME = iprange[, iprange ...], ad esempio:

```
authentication.ip.MY_UNIVERSITY = 10.1.2.3, \ # IP completo  
13.5, \ # IP parziale  
11.3.4.5/24, \ # con CIDR  
12.7.8.9/255.255.128.0 # con netmask
```

Se invece si volessero escludere alcuni IP dall'accesso, è possibile farlo inserendo il prefisso '-'. Per esempio se si volesse includere tutti gli IP di una certa rete di classe B escludendo però gli utenti della rete contenuta di classe c, si può scrivere: 111.222,-111.222.333.

Se il nome del gruppo contenesse degli spazi vuoti è necessario evitare di scrivere, ad esempio, Department\ of\ Statistics.

Configurazione tecnica di DSpace per l'esposizione strutturale dei metadati

Per attivare il DIDL Crosswalk è necessario eseguire le seguenti operazioni:

- Decomentare oai.didl.maxresponse in config/dspace.cfg
- Decomentare la riga DIDL Crosswalk (Crosswalks.didl=org.dspace.app.oai.DIDLCrosswalk) nel file config/templates/oaicat.properties

DSpace e l'harvesting delle tesi di dottorato per BNCf

- Eseguire bin/install-configs
- Riavviare Tomcat
- Verificare l'attivazione del crosswalk all' URL <http://myspace/dspace-oai/request?verb=ListRecords&metadataPrefix=didl>