

DSpace e System Integration: soluzioni innovative per le tesi di dottorato

Andrea Bollini, Davide Vitale, Luigi Andrea Pascarelli

CILEA, Roma e Segrate

Abstract

Le tesi di dottorato rappresentano spesso la frontiera delle attività di ricerca di un ateneo e, in quanto tali, meritano la massima visibilità e diffusione, a beneficio degli autori e delle istituzioni che li sostengono, anche per evitare la marginalizzazione della ricerca italiana nel contesto europeo e mondiale. L'evoluzione della situazione normativa spinge inoltre nella direzione dell'automazione delle procedure e della disponibilità dei formati digitali. E' in quest'ottica che sta maturando una crescente richiesta da parte degli atenei italiani di soluzioni software che possano integrarsi nel tessuto informativo istituzionale, contribuiscano all'automazione del processo di deposito delle tesi di dottorato e forniscano nel contempo alle stesse una maggiore visibilità in rete. A tale esigenze il CILEA risponde estendendo il modulo OA della suite SURplus alla gestione delle problematiche specifiche delle tesi.

Doctoral theses often represent the upper edge of academic research activities. As such, they deserve the highest visibility and dissemination opportunities, for the benefit of their authors and the supporting institutions. Their impact would also contribute to prevent the marginalization of Italian research in the European and world-wide context. Changes in norms and rules are also driving universities towards automation of processes and adoption of digital formats. In this situation universities look for software solutions that can be integrated in their information environment, contribute to the automation of the archiving process and provide theses with high visibility on the net. CILEA fulfills these needs with the OA module of the SURplus suite, customized for the specific treatment of dissertations.

Keywords: Tesi di dottorato, Deposito legale, Open Access, Repository Istituzionali, Shibboleth, OAI-PMH, DIDL.

Introduzione

Le tesi di dottorato rappresentano spesso la frontiera delle attività di ricerca di un ateneo e, in quanto tali, meritano la massima visibilità e diffusione, a beneficio degli autori e delle istituzioni che li sostengono, anche per evitare la marginalizzazione della ricerca italiana nel contesto europeo e mondiale.

L'evoluzione della situazione normativa spinge inoltre nella direzione dell'automazione delle procedure e della disponibilità dei formati digitali. Ad esempio il Piano e-gov 2012 [1] prevede sul fronte delle università la semplificazione e digitalizzazione dei rapporti amministrativi con gli studenti e la dematerializzazione dei relativi flussi documentali, fra i quali rientra a pieno titolo anche il processo di deposito delle tesi.

Sempre sul fronte legislativo vi è poi l'importante novità dell'avvio della fase di

attuazione del deposito elettronico presso le biblioteche nazionali avviato come sperimentazione dalla Biblioteca Nazionale Centrale di Firenze.

La CRUI, Conferenza dei Rettori delle Università Italiane, si è espressa a favore dell'accesso online alle tesi di dottorato: l'approvazione delle "Linee guida per il deposito delle tesi di dottorato negli archivi aperti" [2], avvenuta nel novembre del 2007, ha portato a conclusione l'iter iniziato dal Gruppo di Lavoro per l'Open Access nel giugno 2006. Il gruppo, costituito da delegati di oltre 40 atenei italiani, oltre ad un ristretto gruppo di esperti fra i quali anche un rappresentante del CILEA, ha espresso la raccomandazione che gli atenei predispongano, per le tesi di dottorato in formato digitale, dei repository conformi agli standard internazionali per l'accesso aperto alla letteratura scientifica.

E' in quest'ottica che sta maturando una crescente richiesta da parte degli atenei italiani di soluzioni software che possano integrarsi nel tessuto informativo istituzionale, contribuiscano all'automazione del processo di deposito delle tesi di dottorato e forniscano nel contempo alle stesse una maggiore visibilità in rete. A tale esigenze il CILEA risponde estendendo il modulo OA della suite SURplus alla gestione delle problematiche specifiche delle tesi.

Interoperabilità esterna

La piattaforma SURplus [3], ed in particolar modo il modulo OA, si caratterizza per la sua elevata interoperabilità e adattabilità a differenti realtà ed esigenze dei singoli atenei. Il modulo OA, basato sulla piattaforma open source DSpace [4], supporta a vari livelli standard internazionali quali OAI-PMH [5], Dublin Core [6], METS [7], DIDL [8] e altri ancora, che favoriscono l'interoperabilità del modulo non solo verso sistemi interni di ateneo, siano essi altri moduli della piattaforma SURplus o sviluppati per l'ateneo da altri fornitori, ma anche e soprattutto verso piattaforme nazionali e internazionali.

Relativamente alle tesi di dottorato ad esempio la sperimentazione condotta dalla BNCf [9] utilizza il protocollo OAI-PMH e il formato descrittivo delle risorse DIDL. Questo formato, noto anche con il nome di MPEG-21, consente di affiancare alla descrizione bibliografica dei singoli record (tesi di dottorato) anche informazioni puntuali rispetto agli oggetti digitali collegati ed in particolar modo il link diretto per il download della risorsa, utilizzato appunto dalla BNCf per l'harvesting dei "full text". Il formato è comunque altamente estendibile e potrà veicolare in futuro informazioni più articolate, quali ad esempio i diritti di utilizzo della singola risorsa elettronica. E' auspicabile quindi la realizzazione di specifiche raccomandazioni o meglio ancora di un application profile da parte delle Biblioteche Nazionali, in modo da consentire un'interoperabilità non solo applicativa ma anche "semanticamente" più ricca.

L'uscita OAI-PMH delle realizzazioni SURplus OA è configurata di default per esporre i record anche in formato DIDL. Questo implica l'immediata compatibilità di tutte le iniziative basate su SURplus OA con gli attuali criteri di harvesting della BNCf.

A livello tecnico gli interventi sull'uscita OAI-PMH, effettuati per il modulo SURplus OA, non

si limitano alla mera attivazione del formato DIDL: infatti è stata personalizzata la generazione della sezione relativa ai metadati descrittivi Dublin Core, consentendo ai singoli atenei di adattarla alle specifiche scelte descrittive effettuate. A differenza di quanto avviene per un DSpace standard inoltre, la configurazione della sezione DC inclusa nel DIDL viene automaticamente ricalcata dalla configurazione utilizzata per il formato oai_dc, consentendo quindi una più semplice ed economica manutenzione dell'applicativo ed assicurando nel contempo la coerenza trasversale delle scelte di descrizione bibliografica.

Integrazione di processi

Iniziamo l'analisi del processo di gestione delle tesi di dottorato introducendo una semplice categorizzazione degli stakeholder coinvolti:

- Interni all'ateneo: dottorandi, relatori, segreterie amministrative e didattiche, bibliotecari;
- Esterni all'ateneo: biblioteche nazionali adibite al deposito legale, partner di ricerca interessati alla collaborazione scientifica, industrie e partner commerciali interessati al know-how dell'ateneo per il trasferimento tecnologico.

Mentre per i secondi l'interesse principale è l'interoperabilità come descritto nel precedente paragrafo, per i primi si tratta di ottenere una stretta integrazione di tutti i sistemi informativi coinvolti ai fini di un'ottimizzazione dei processi. L'integrazione con i sistemi informativi di ateneo è per ovvie ragioni estremamente variegata. Le soluzioni adottate ed il processo effettivo dipendono fortemente dal contesto specifico in cui viene effettuato l'intervento. E' possibile comunque delineare un processo di massima dell'iter riportato in figura a titolo esemplificativo (*Fig. 1*).

Confrontando questo flusso ideale con il workflow implementato dal software DSpace si ottiene quella che è la nostra proposta standard di modellizzazione in DSpace del flusso di deposito come esemplificato in figura (*Fig. 2*).

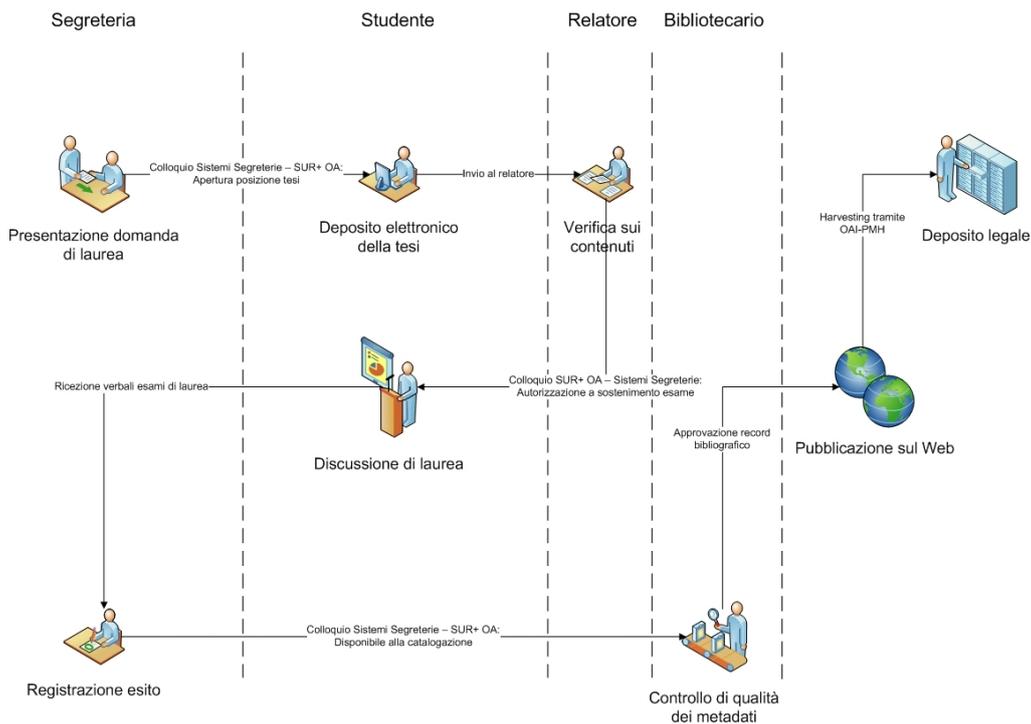


Fig. 1 - Processo di consegna delle tesi di dottorato

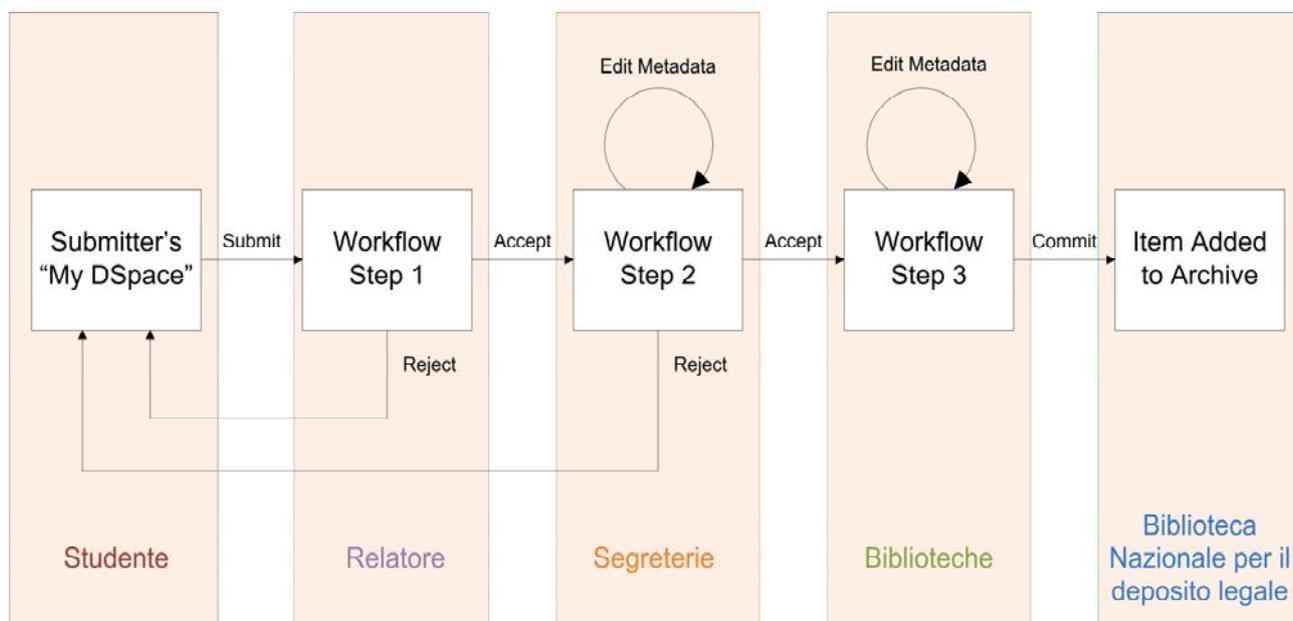


Fig. 2 - Flusso di consegna di DSpace

A livello di architettura generale si possono invece utilizzare due differenti modelli:

- PULL
- PUSH

Nel primo modello, il modulo SURplus OA riceve notifica della presentazione della domanda direttamente dal sistema delle segreterie. Le tecnologie di trasporto di tale notifica sono quelle classiche alla base dell'integrazione di sistemi che vanno quindi dai moderni web service [10], a messaggi JMS [11] o di posta elettronica processati da robot automatici, fino ad arrivare eventualmente all'integrazione tramite middleware di comunicazione più complessi, quali ad esempio CORBA [12].

Il secondo modello prevede invece che sia il modulo SURplus OA ad interrogare periodicamente i servizi di segreteria di ateneo alla ricerca di nuove presentazioni di tesi di dottorato o modifiche ad iter già avviati. Le tecnologie utilizzate in questo secondo caso, oltre a quelle già elencate per il modello PULL, prevedono molto spesso un accesso diretto alla banca dati delle segreterie o ad apposite tabelle/viste di frontiera appositamente predisposte.

Il CILEA ha comunque introdotto un livello di indirizzamento, ottenuto mediante il riversamento dei dati specifici di ateneo in uno specifico database normalizzato mediante

apposite procedure ETL, in modo tale che anche partendo da basi dati differenti tutti gli atenei condividano buona parte dello stack applicativo, riuscendo in questo modo a realizzare economie di scala sia nell'implementazione di nuove istanze, sia nella manutenzione ed evoluzione delle adozioni già in corso. Un diagramma di massima dell'architettura realizzata dal CILEA è riportato in figura (Fig. 3).

Si noti in particolare come l'intervento CILEA si vada ad integrare, per entrambi i modelli (PULL e PUSH), in maniera armonica con quanto già reso disponibile dalla piattaforma DSpace: tale modalità di integrazione garantirà un minor costo di migrazione della soluzione a future versioni di DSpace, con una conseguente garanzia dell'investimento fatto per gli atenei che l'avranno adottata.

Funzionalità specifiche

A completamento delle specifiche esigenze di interoperabilità esposte, la gestione delle tesi di dottorato pongono classicamente un requisito al momento non gestito dalla versione standard di DSpace, ovvero la possibilità di limitare parzialmente o totalmente l'accesso al materiale full text della tesi e, più in generale, la possibilità di impostare specifiche regole di embargo (Fig. 4).

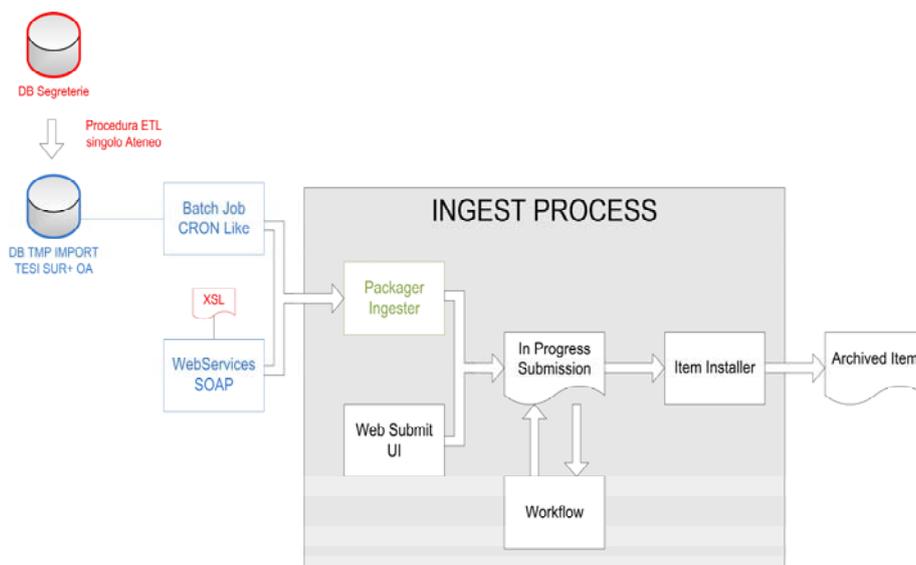


Fig. 3 – Architettura SURplus OA per l'automazione del processo di consegna delle tesi di dottorato

Fig. 4 – Funzionalità di embargo di SURplus OA

La necessità di limitare l'accesso al full text scaturisce da varie situazioni, anche in risposta a specifici regolamenti di ateneo. Ad esempio possono esservi forme di collaborazione con industrie che richiedano una diffusione "controllata" dei risultati conseguiti o comunque un periodo transitorio di indisponibilità dei contenuti della tesi. Il modulo SURplus OA è stato quindi dotato di una specifica funzionalità di configurazione e gestione di questo periodo (embargo). In particolare è possibile definire, a livello di policy di archivio ma anche a livello di macro partizioni dell'archivio stesso (Facoltà, Scuole di dottorato, etc.), quali attori (dottorandi, segreterie, catalogatori, amministratori di sistema) possano stabilire (quali) tempi e (quali) modalità di accesso al singolo file depositato.

Poter definire le caratteristiche dell'embargo per ogni file consente la realizzazione di strategie disseminative anche molto diverse tra loro, ad esempio offrire accesso aperto immediato ad alcuni capitoli della tesi, restringere la disponibilità solo da postazioni

della biblioteca per altri capitoli, porne altri ancora sotto embargo per un certo periodo in rispetto a specifici accordi per la registrazione di brevetti.

Anche le tipologie di utenti sono definibili dal singolo ateneo: l'associazione di un utente ad una o più tipologie avviene mediante automatici determinati dall'anagrafica dell'utente stesso (ruolo istituzionale, etc.) e/o indicazioni esplicite effettuate tramite interfaccia web da parte degli amministratori di sistema.

E' importante sottolineare che la gestione dell'accesso al full text riservato a determinate tipologie di utenti è stata realizzata utilizzando l'infrastruttura di gestione dei permessi preesistente in DSpace. Questo consente un'immediata integrazione della soluzione con qualsiasi modalità di autenticazione venga utilizzata dal repository, sia esso un SingleSignOn di ateneo, una login centralizzata basata su LDAP e similari o una login federata tramite ad esempio il protocollo Shibboleth. In questo ultimo caso è interessante segnalare come, grazie alla presenza della federazione IDEM [13] del GARR, sia possibile ipotizzare

per un prossimo futuro la messa a disposizione delle tesi di dottorato a specifiche tipologie di utenti, trasversali a tutti gli atenei italiani, quali studenti, personale docente etc. senza che l'archivio debba farsi carico della gestione delle credenziali di accesso e dei ruoli dei singoli utenti. L'altro vantaggio immediato derivante dal poter utilizzare l'infrastruttura di autenticazione/autorizzazione nativa di DSpace è la possibilità di consentire l'accesso ai full text da parte dei server delle Biblioteche Nazionali adibite all'harvesting, nonché l'accesso dai terminali delle medesime biblioteche, nel rispetto della legislazione relativa al deposito legale delle tesi di dottorato.

Al momento la comunità open source DSpace ha costituito un gruppo di lavoro [14] con lo scopo di integrare nativamente dalla prossima versione di DSpace, la 1.6, una soluzione per la gestione delle problematiche legate all'embargo. Il CILEA, come sempre a garanzia degli investimenti effettuati dai propri partner, segue con attenzione le attività di tale gruppo allo scopo di proporre la propria soluzione o quanto meno di garantirne la compatibilità futura.

Casi d'uso

Il CILEA ha al proprio attivo la realizzazione di numerosi repository con software DSpace, sia nell'ambito della suite SURplus sia come installazione autonoma. Tra i nostri partner che utilizzano DSpace per la gestione delle tesi di dottorato citiamo:

- l'Università di Parma con DSpace@UniPr [15], primo repository DSpace realizzato dal team AePIC nel 2003, che utilizza da quasi un anno il protocollo Shibboleth fra le modalità di autenticazione dei propri utenti;
- l'Università Cattolica del Sacro Cuore che, in produzione da oltre un anno, ha realizzato DocTA [16], repository dedicato esclusivamente alla gestione delle tesi di dottorato integrato nei flussi delle segreterie; sono in corso di sviluppo nuove funzionalità che semplificheranno ulteriormente l'iter amministrativo affrontato dai dottorandi, ad esempio la produzione automatica della dichiarazione di conformità in sostituzione della compilazione manuale;
- l'Università di Bergamo, il cui archivio istituzionale AISBerg [17] è appena stato reso pubblicamente accessibile e ospiterà, tra l'altro, le tesi di dottorato di tutte le scuole presenti in ateneo;

- l'Università Statale di Milano, prima ad adottare nel 2005 il precursore del modulo SURplus OA con il proprio archivio AIR [18] che, nel corso dei prossimi mesi, terminerà la migrazione alla versione più recente del modulo in modo da potersi avvantaggiare tra l'altro della funzionalità di gestione dell'embargo;
- l'Università di Milano Bicocca, il cui archivio aperto BOA [19] sta diventando rapidamente un punto di riferimento tra gli archivi istituzionali italiani, che nei prossimi mesi aprirà nuove collezioni per la gestione delle tesi di dottorato integrata nei flussi delle segreterie e che è stata coinvolta con successo dalla BNCf, assieme all'Università di Parma e poche altre in queste prime fasi di testing del protocollo di harvesting per il deposito legale basato sul protocollo OAI-PMH ed il formato DIDL.

A queste esperienze vanno inoltre aggiunte altre iniziative in corso con altri atenei italiani. Ad esempio abbiamo avviato con un importante ateneo del Nord Italia un progetto per la realizzazione di un archivio DSpace, integrato nei flussi delle segreterie e accessibile tramite Shibboleth, che accoglierà con una politica di deposito obbligatorio le tesi di tutti i livelli (triennali, specialistiche e di dottorato) dell'ateneo.

Questo progetto risulta di particolare avanguardia nel panorama italiano non solo per la scelta di affrontare le problematiche relative alle tesi di qualsiasi livello, ma anche perché è al momento orientato ad utilizzare la procedura online in completa sostituzione della modulare cartacea di deposito delle tesi, in particolare la liberatoria. Questo comporterà una importante semplificazione nell'iter amministrativo di deposito delle tesi a vantaggio dei laureandi e degli uffici preposti e rappresenterà uno dei più significativi passi in avanti nei processi di dematerializzazione avviati dagli atenei italiani.

Di questo e degli altri progetti in corso vi daremo notizia in uno dei prossimi numeri del Bollettino del CILEA.

Bibliografia

- [1] URL: <http://www.e2012.gov.it/>
- [2] URL: <http://www.cruil.it/HomePage.aspx?ref=1149>
- [3] URL: <http://www.cilea.it/index.php?id=surplus>
- [4] URL: <http://www.dspace.org>

- [5] URL: <http://www.openarchives.org/>
- [6] URL: <http://dublincore.org/>
- [7] URL: <http://www.loc.gov/standards/mets/>
- [8] URL: <http://xml.coverpages.org/mpeg21-didl.html>
- [9] URL: <http://www.bncf.firenze.sbn.it/>
- [10] URL: http://it.wikipedia.org/wiki/Web_service
- [11] URL: <http://java.sun.com/products/jms/>
- [12] URL: <http://it.wikipedia.org/wiki/CORBA>
- [13] R. Gibellini "Il CILEA entra nella Federazione IDEM del GARR", *Bollettino del CILEA*, n. 111, giugno 2008.
URL: <http://www.idem.garr.it/>
- [14] URL: http://wiki.dspace.org/index.php/Embargo_1.6
- [15] URL: <http://dspace-unipr.cilea.it/>
- [16] URL: <http://tesionline.unicatt.it/>
- [17] URL: <http://dspace-unibg.cilea.it/oa/>
- [18] URL: <http://air.unimi.it/>
- [19] URL: <http://boa.unimib.it/>