Strumenti aperti per gli archivi documentali

esperienze degli Open Archive nella ricerca

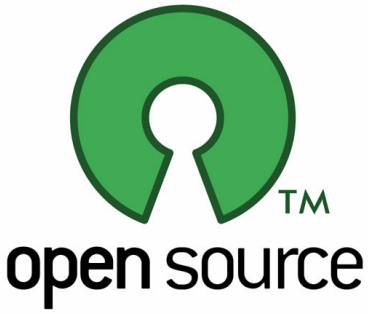
Documento, documentazione, documentarsi
Roma, 1 marzo 2007
dott. Andrea Bollini, CILEA

Agenda

- OAI & OpenSource
- Panoramica degli strumenti “open”
- L’apporto del CILEA
- I repositories: definizione e possibili classificazioni
- La situazione italiana
- I software DSpace & EPrints
- Le iniziative italiane supportate dal CILEA

OAI & OpenSource: un binomio vincente

"The Open Archives Initiative develops and promotes interoperability standards that aim to facilitate the efficient dissemination of content. OAI has its roots in the open access and institutional repository movements."



"The basic idea behind open source is very simple: When programmers can read, redistribute, and modify the source code for a piece of software, the software evolves. People improve it, people adapt it, people fix bugs. And this can happen at a speed that, if one is used to the slow pace of conventional software development, seems astonishing."

OAI & OpenSource: un binomio vincente

OAI:

- Standard di interoperabilità
- Visibilità dei risultati della ricerca
- Accesso ai risultati della ricerca

OSI:

- Condividere conoscenza
- Riutilizzo, adattamento & evoluzione

OAI & OpenSource: un binomio vincente

Diversi studi sulla differenza di impatto tra articoli disponibili in modalità Open Access e articoli disponibili solo tramite canali tradizionali, concordano nell'affermare che gli articoli OA hanno un impatto significativamente più elevato. Per una bibliografia costantemente aggiornata sull'argomento è possibile consultare http://opcit.eprints.org/oacitation-biblio.html

Panoramica degli strumenti “open”

La forte connessione filosofica-metodologia tra OAI e OSI trova ulteriore riscontro anche nel vasto panorama di software che quest’ultima ha prodotto a supporto dell’iniziativa OA. Attualmente sono, infatti, disponibili strumenti open source “maturi” per tutte le componenti dell’architettura OAI, da service provider a data provider, con vari livelli di specializzazione:

- strumenti omni comprensivi: CDS invenio (ex CDSware)
- strumenti per la realizzazione di dataproviders generici: Fedora, DSpace
- strumenti per la realizzazione di repositories: EPrints
- strumenti per la realizzazione di riviste aperte: OJS, HyperJournal
- framework di supporto per l’implementazione di interoperabilità (OAICat, OAIHarvester2)

L’apporto del CILEA

Leader in Italia il CILEA ha attivato da oltre tre anni un team specializzato, AePIC, per assistere e guidare i propri consorziati e tutti gli enti di ricerca che ne facessero richiesta nella delicata fase di avvio, nella pianificazione e nella gestione di progetti documentali “aperti”.

L’utilizzo di software e tecnologie open source permette al CILEA di concentrarsi sulle specificità dei singoli clienti garantendo una maggiore qualità e rapidità nello sviluppo delle proprie realizzazioni.

L’apporto del CILEA

Nel corso di questi anni il CILEA si è specializzato nell’impiego dei maggiori software open source citati selezionando in particolare DSpace ed EPrints per la gestione di repositories e, più in generale, digital library. L’azione del CILEA, coerentemente con lo spirito dei software utilizzati, è stata condotta con una costante attenzione alle evoluzioni in atto nelle rispettive comunità di sviluppatori tanto da entrare a far parte dei contributors di EPrints, OJS e del ristretto gruppo di committers di DSpace.

I repositories: definizione

Per repository si intende un archivio aperto in linea che renda disponibile in forma digitale la cosidetta letteratura grigia (pre-prints, dispense, tesi, ecc…) senza effettuare validazioni qualitative sui contenuti (peer-review); ogni repository prevede però normalmente delle limitazioni al deposito dettate da proprie policy interne.

I repositories: classificazioni

Le limitazioni riguardano generalmente :

- Tipologia di contenuto: ricerca, didattica
- Produttori del contenuto: Istituzionali, disciplinari, misti

e non limitandosi a repository open access

- Tipologia di accesso: aperto, delayed o ristretto

Dalle policy ai requisiti funzionali

Ogni policy che l'archivio deve implementare ha inevitabilmente un impatto sui requisiti funzionali del progetto.

Autenticazione, Workflow, Metadati & authority files, DRM sono infatti solo alcune delle caratteristiche che vengono influenzate da queste policy

Repository Istituzionali - IR

La realizzazione di un repository istituzionale (IR) comporta normalmente la necessità di integrare il sistema di autenticazione del repository con un sistema centralizzato dell'istituzione (basato ad esempio su LDAP). A livello di workflow di approvazione gli IR non presentano invece particolari necessità ed in taluni casi, anche in virtù del rapporto di fiducia tra autore/dipendente ed ente, potrebbe essere completamente disabilitato (submission=archiviazione). E' infine da segnalare che quando sono previste policy restrittive per l'accesso ai full-text è normalmente richiesta la gestione delle autorizzazioni tramite indirizzo IP/rete di accesso.

Repository disciplinari

Un repository disciplinare si caratterizza dal fatto di contenere materiali di un settore scientifico ben determinato a volte anche molto ristretto. Questo comporta l'adozione di metadati specifici con soggettari normalmente sottoposti ad autority files. Un processo di workflow per l'immissione in archivio è assolutamente indispensabile anche solo al fine di evitare l'inserimento di materiale "improprio". L'invio è infatti normalmente effettuato da autori "sconosciuti" all'organizzazione(i) che supporta l'archivio attraverso la registrazione autonoma nel sistema e procedure di self-archiving.

Repository per la didattica

Un repository di materiale didattico si trova spesso a dover affrontare problemi di amministrazione decentralizzata. Un sistema che preveda l'inserimento di materiale didattico deve risultare sufficientemente flessibile da permettere una gestione autonoma da parte di ogni docente dei contenuti, della loro organizzazione e modalità di accesso. Il riutilizzo del materiale avviene frequentemente deve quindi essere possibile associare il medesimo materiale a più corsi.

Repository per la ricerca

In un repository di materiale di ricerca rivestono particolare importanza i tools di supporto al fruitore finale nell'ottica del riutilizzo e della generazione di nuova conoscenza. Il materiale di ricerca viene costantemente citato dai ricercatori che quindi trarrebbero il massimo vantaggio da tools di export in formato bibliografico e strumenti di ricerca e navigazione contestuali. In una situazione ideale nello stesso repository si potrebbero inoltre trovare più versioni dello stesso articolo (pre-print, post-print, publisher version) tra cui dovrebbe essere garantita una semplice navigazione.

La situazione italiana

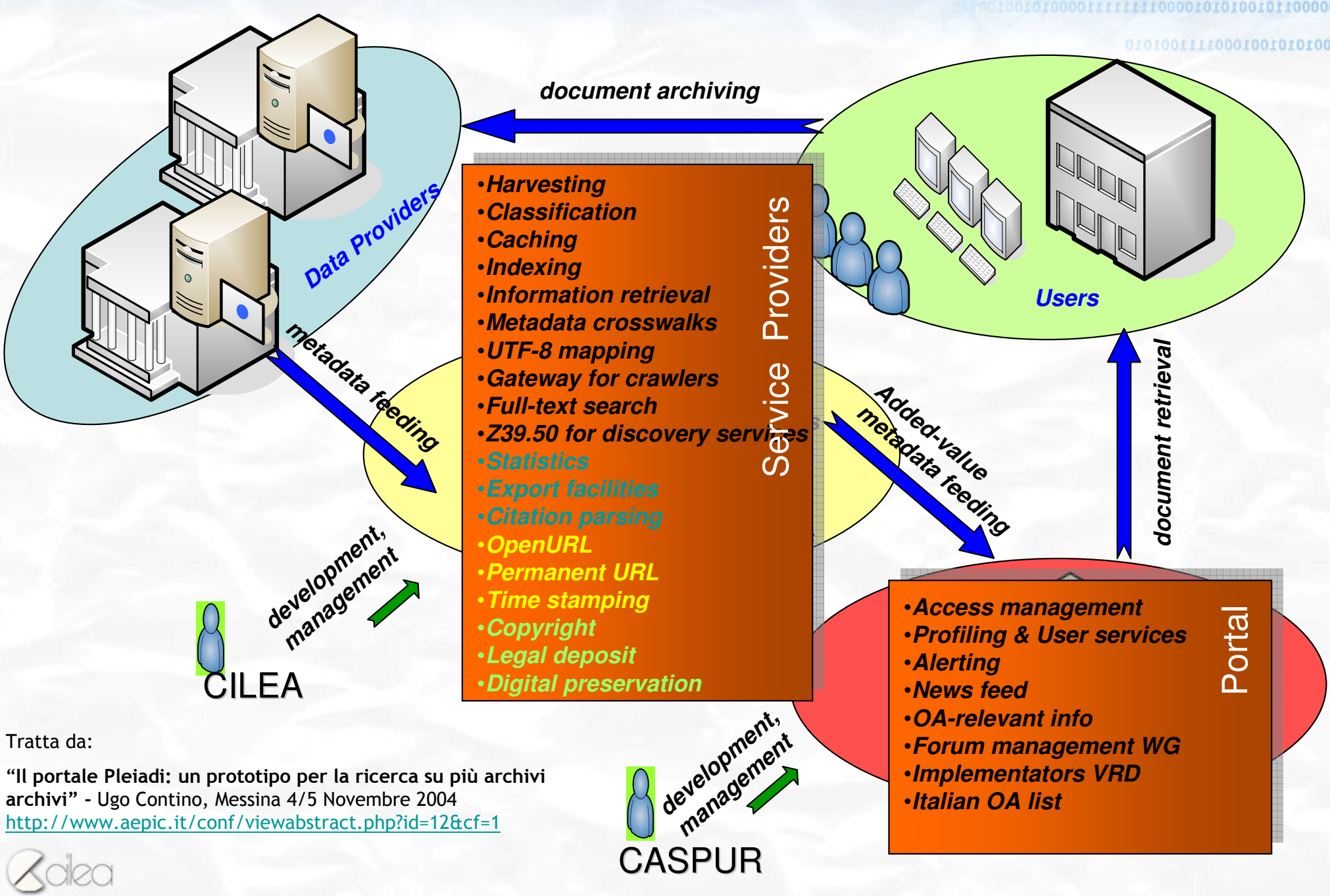
- Dichiarazione di Messina: 4-5 November 2004
- Sottoscritta da 75 università italiane su 77
- Molte università non hanno ad oggi ancora avviato progetti di repository OA e quelli esistenti stentano a decollare
- Atenei & enti di ricerca, soprattuto di grosse dimensione e peso scientifico, hanno difficoltà ad evadere i “debiti informativi” verso il MIUR ed organismi internazionali

La situazione italiana

- Pleiadi Service Provider
  http://www.openarchives.it/pleiadi
- Numero di data providers harvested: 17*
- Numero di documenti harvested: 7424*
- Media documenti/providers: 436*

*Dati al 27 febbraio 2007

Architettura PLEIADI



Tratta da:

"Il portale Pleiadi: un prototipo per la ricerca su più archivi archivi" - Ugo Contino, Messina 4/5 Novembre 2004
http://www.aepic.it/conf/viewabstract.php?id=12&cf=1

La situazione italiana

Attualmente i bibliotecari sono tra i maggiori sostenitori dell'Open Access

Vantaggi:

- metadati di qualità

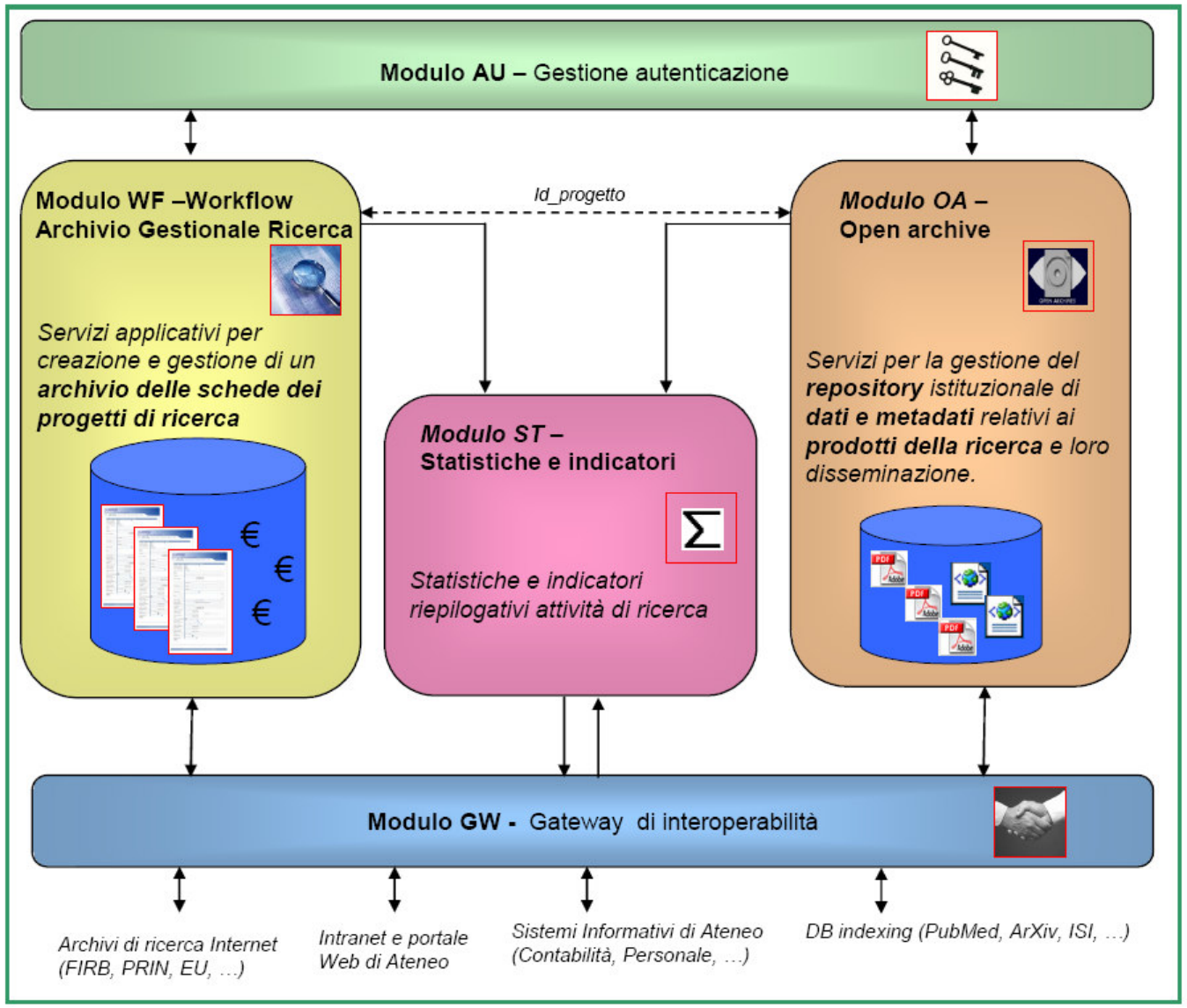
Svantaggi:

- Scarsa influenza accademica
- Non sono i produttori dei contenuti

La situazione italiana

- Autori sono riluttanti a depositore preoccupati dalla loro carriera
- E' possibile incoraggiarli? come?
- **Informazione!** E' necessario far comprendere ai ricercatori i vantaggi conseguenti ad una pubblicazione OA (visibilità, impatto) e ridurre le preoccupazioni relative al copyright (v. Progetto SHERPA/RoMEO)
- **Servizi aggiuntivi!** L'autore deve "vivere" l'archivio come un tool personale di supporto alla propria attività lavorativa e non un ulteriore impegno burocratico (estrazione di liste bibliografiche per la redazione del CV, integrazione con i software di gestione bibliografica, ecc…)
- Attualmente ci sono diversi registri, sia a livello locale sia a livello nazionale, con scopi differenti ma che richiedono essenzialmente I medesimi dati, questo obbliga il ricercatore ad un dispendioso e ripetitivo lavora di data entry

La suite SURplus del CILEA



Progettare un archivio documentale

- Dimensione del materiale posseduto e crescita attesa
- Chi possiede il contenuto: legalmente e “fisicamente”
- Chi crea contenuto
- Dimensione e distribuzione dello staff redazionale
- Funzionalità utente e target di riferimento
- Tipologia di contenuto
- Ambiente ICT

I software DSpace & EPrints

- Entrambi sono software open source che consentono la realizzazione di repository OAI
- DSpace presenta un data model più articolato che gli permette di gestire situazioni amministrative più complesse e far coesistere esigenze differenti all'interno dello stesso repository
- EPrints è il software storico, nato appositamente per realizzare repository di materiale di ricerca, ha tra i suoi punti forza la semplicità di utilizzo

DSpace

- Sviluppato inizialmente dal MIT e dalla HP, viene rilasciato come OS nel 2002
- Basato su tecnologie Java J2EE, conforme allo standard OAIS
- giunto alla versione 1.4.1 si caratterizza per una comunità vivace e produttiva
- le linee di sviluppo sono indicate dal gruppo di Committers definito sulla base di criteri meritocratici

DSpace: data models

- Cinque entità principali che permettono la strutturazione e aggregazione delle informazioni
- Comunità: raggruppamento amministrativo/logico di più collezione (e comunità); utilizzato ad es. per creare Aree Dipartimentali all'interno di un unico archivio di ateneo, singole sedi per Enti distribuiti sul territorio, progetti/linee di ricerca, ecc…
- Collezioni: sono un raggruppamento di più item omogenei per tipologia (metadati) e modalità di trattamento (workflow); rappresentano attualmente il punto principale per le personalizzazioni dell'archivio
- Item: rappresenta la scheda bibliografica (metadati) di un singolo documento; contiene al suo interno uno o più bundles di bitstreams
- Bundle: sono una aggregazione di bitstream attualmente utilizzati per separare i documenti originali, quelli risultanti da trasformazioni automatiche (ad es. estrazione fulltext), licenze di archiviazione e Creative Commons
- Bitstream (= fulltext)

DSpace: gestione accounts

La gestione degli accounts in DSpace avviene mediante due entità: eperson e group. La prima rappresenta un singolo utente del sistema, ne permette la gestione dei dati anagrafici e l'accesso al sistema. La seconda è un aggregatore di più accounts (o gruppi) utilizzata per semplificare la gestione dei permessi (il gruppo degli studenti di Fisica, il gruppo dei docenti, ecc…)

DSpace: DRM

La gestione dei permessi è molto flessibile è infatti possibile definire policy di lettura, modifica e creazione per tutte le entità del datamodels. Attualmente la webUI e il server OAI-PMH non supporta però nativamente le eventuali restrizioni alla visibilità dei metadati.

A livello di API è inoltre possibile definire policy di accesso temporali.

DSpace: caratteristiche della WebUI

- Interfaccia web conforme XHTML e WAI
- Multilinguismo
- Semplici possibilità di branding (CMS)
- Facilmente personalizzabile
- Identificativo persistente (Handle) per ogni Comunità, collezione è item

DSpace: retrieval e disseminazione

- Potente motore di ricerca basato su Lucene con indici personalizzabili e ricerca fulltext
- Indici e liste di scorrimento personalizzabili
- Notifica via RSS e email dei nuovi contenuti
- Supporto dello standard OpenURL per l'accesso ai servizi di tipo SFX dell'istituzione
- Supporto allo standard SRW/U
- Identificativo persistente per ogni item, comunità e collezione: handle
- OAI-PMH: METS, MODS, DCQualified, MPEG-21 e semplice creazione di formati personalizzati

DSpace: metadati

- Gestione di schemi multipli di metadati
- Possibilità di definire il formato di immissione per ogni metadato (nome, data, lista controllata, ecc…)
- UTF-8
- Supporto per multilinguismo dei metadati

DSpace: autenticazione

- Supporto per LDAP
- Supporto per CAS
- Supporto per certificati X.509
- Facile scrittura di metodi personalizzati
- Inserimento in gruppi in base all'IP di provenienza
- Possibilità di utilizzare simultaneamente più modalità di autenticazione

DSpace: Import/Export

- Import da formato XML nativo, XML+XSL, PDF, personalizzato
- Export in XML secondo vari formati con la possibilità di utilizzare tutti i formati definiti per OAI-PMH (METS, MODS, DCQualified, MPEG-21)

DSpace: uno sguardo al futuro…

- E’ da poco terminata la revisione architetturale che porterà alla nascita della versione 2.0; la transizione avverrà in maniera progressiva mantenendo la migrazione dalle versioni precedenti più semplice possibile
- Le maggiori novità riguarderanno la revisione dell’architettura software che semplificherà ulteriormente l’estendibilità, la modularità e la manutenzione del prodotto rafforzando quindi la possibilità di riutilizzo come componente
- Il data model sarà esteso consentendo di associare metadati personalizzati ad ogni oggetto dell’architettura (possibile supporto per il MAG)
- Miglioramenti alle performance per garantire la corretta gestione di masse critiche di documenti
- L’introduzione di un motore di workflow che consentirà la massima flessibilità nei processi di immissione, verifica, validazione e accettazione dei documenti

Alcune iniziative *by* CILEA

- E-lis - http://eprints.rclis.org/
- Earth-prints (Ist. Nazionale Geofisica e Vulcanologia) http://www.earth-prints.org
- AIR (Univ. Milano) – http://air.unimi.it
- AperTO (Univ. Torino) – http://aperto.unito.it
- Archivio digitale del ceum (Univ. Macerata) - http://archiviodigitale.unimc.it/
- E-ms (Ist. Medicina Sociale) -
- ArmiDA (Univ. Milano) – http://armida.unimit.it
- Biblioteca Digitale Medievistica – di prossima attivazione
- SSPAL.doc – di prossima attivazione
- DSpace@UniPr – http://dspace-unipr.cilea.it
- Envimarch (CNR di Potenza) – di prossima attivazione
- …

E-LIS

- Il più vasto repository di Library & Information Science, coordinato da uno staff readazionale internazionale che garantisce metadati qualitativamente elevati.
- Attualmente possiede oltre 5100+ documenti ad accesso aperto
- Basato sul software open source EPrints

Earth-prints.org

- Giovane repository disciplinare: raccoglie materiale di ricerca sulla geofisica
- Promosso dall'Istituto Nazionale di Geofisica e Vulcanologia che l'utilizza come proprio IR
- Attualmente possiede un patrimonio di oltre 1600+ documenti quasi tutti disponibili ad accesso aperto

AIR: Archivio Istituzionale della Ricerca – Univ. degli studi di Milano

- IR dell'Università degli Studi di Milano utilizzato dai dipendenti per l'evasione di debiti informativi nei confronti dell'istituzione
- Avviato nel maggio del 2006 possiede attualmente un patrimonio di oltre 15000+ schede bibliografiche purtroppo solo in minima parte corredata da fulltext
- Attivazione di nuove soluzioni tecnologiche e organizzative con servizi di supporto ai docenti per incentivare l'inserimento dei fulltext

SSPAL.DOC: l'archivio istituzionale

- IR della Scuola Superiore per la Pubblica Amministrazione Locale diventerà accessibile pubblicamente a breve
- Prima esperienza significativa di utilizzo delle tecnologie OAI in una P.A.
- Recupero del pregresso della Scuola Nazionale, con revisione e normalizzazione dei metadati da parte dei bibliotecari
- Funzionalità di ricerca, scorrimento e navigazione dell'archivio orientata agli utenti tipici della scuola
- 300+ documenti ad accesso aperto