DigItalia

Anno V, Numero 1 - **2010**

Rivista del digitale nei beni culturali

ICCU-ROMA

Magazzini digitali: dal prototipo al servizio

Giovanni Bergamin – Maurizio Messina
Biblioteca nazionale centrale di Firenze – Biblioteca nazionale marciana di Venezia

Introduzione

Il progetto Magazzini digitali, avviato nel 2006 dalla Fondazione rinascimento digitale, dalla Biblioteca nazionale centrale di Firenze e dalla Biblioteca nazionale centrale di Roma si propone ora di mettere a regime un sistema per la conservazione permanente dei documenti elettronici pubblicati in Italia e diffusi tramite rete informatica, in attuazione della normativa sul deposito legale (L. 106/2004, d.p.r. 252/2006).

Nella prima parte di questo contributo verrà descritta l'architettura tecnica del progetto, ma è bene tenere presente fin d'ora che la conservazione digitale, come è oramai ben noto, non si esaurisce solo in procedure di natura tecnologica. Le strategie volte ad evitare la perdita dei bit o a prevenire le dipendenze dall'hardware o dal software sono infatti solo una parte del problema. Vanno tenute nel conto dovuto le implicazioni economiche (la sostenibilità), la necessità di selezionare cosa è necessario conservare per le generazioni future, gli aspetti legali legati alla normativa sul diritto d'autore, la necessità della cooperazione fra le istituzioni titolari del deposito legale. Alcuni di questi aspetti verranno esaminati nella seconda parte di questo contributo

Ai fini del progetto, la conservazione digitale può essere definita come un servizio pubblico fornito da depositi *digitali affidabili* o *fidati (trusted* o *trustworthy digital repositories)* in grado di assicurare, per le risorse digitali depositate, la *leggibilità* a livello di bit da parte di una macchina (*viability*), la *interpretabilità* a livello di formato (per esempio: pdf, doc, ecc.) da parte di un elaboratore (*renderability*), l'*autenticità* (*authenticity)* intesa come identità e integrità dell'oggetto digitale, e la effettiva *disponibilità* (*availability*) per le comunità designate (comunità di riferimento, interessate all'uso di quelle risorse).

Il nome del progetto richiama intenzionalmente i magazzini delle biblioteche titolari del deposito legale. Come definito da uno storico progetto europeo sulla conservazione digitale (Networked European Deposit Library – NEDLIB, svoltosi dal 1997 al 2000):

> «For us, as memory organizations, this means we have to move from paper-based stacks to digital stacks».

Per molti aspetti i Magazzini digitali sono comparabili a quelli convenzionali: le risorse digitali devono essere conservate indefinitamente; i Magazzini digitali crescono man mano che si aggiungono nuove risorse; modifiche o cancellazioni di risorse non sono di norma possibili; è impossibile predefinire la frequenza d'uso delle risorse, alcune delle quali non saranno mai utilizzate, o lo saranno raramente. Ed è interessante notare che nove anni più tardi, una ricerca su Google dei termini *Digital Stacks* restituisce la medesima espressione usata nel contesto della conservazione digitale:

> «Digital stacks: rather than boxes, shelves, and climate controlled environments, digital information must be stored in containers, file systems, and secure servers».

L'Architettura tecnologica

Lo scopo del progetto è stato quello di impiantare un'infrastruttura tecnologica con caratteristiche di "permanenza". Dando per assodato che i guasti o le disfunzioni dei vari componenti sono la norma piuttosto che l'eccezione, l'infrastruttura è basata sulla replica dei dati (macchine differenti collocate in luoghi differenti) e su componenti hardware semplici e universalmente diffusi, non dipendenti dai produttori, e che possono essere sostituiti facilmente: in altre parole, semplici personal computer. Inoltre, l'infrastruttura non è dipendente da software proprietario ma si basa su sistemi operativi e servizi a codice sorgente aperto (*open source*).
Attualmente un normale personal computer può facilmente immagazzinare fino ad 8 terabyte di dati, su 4 dischi da 2000 gigabyte, usando tecnologie SATA diffusissime ed economiche. La replica dei dati si basa su comuni programmi di utilità (*utility)* per la sincronizzazione dei dischi come rsync e, per evitare dipendenze hardware come ad esempio le dipendenze da un determinato *disk controller*, non viene utilizzato il RAID (Redundant Array of Independent Disks).
Nel passaggio dal prototipo al servizio, inoltre, è stata modificata l'architettura tecnologica del *dark archive*. Il progetto originario prevedeva infatti l'uso di un sistema di *memorizzazione non in linea (offline storage)* basato su nastri di tipo LTO (Linear Tape-Open); successivamente si è deciso di utilizzare la medesima tecnologia già individuata per i due siti principali, cioè la *memorizzazione in linea (online storage)* su normali personal computer. L'espressione "in linea", comunque, non cambia la funzione del *dark archive*, che è quella di servire da archivio di sicurezza dei dati usabile in caso di *disaster recovery*. I nastri LTO sono sicuramente una soluzione robusta ed affidabile introducono dipendenze e vincoli di natura tecnologica e gestionale (per esempio librerie automatizzate o robot). Per lo stesso motivo si è deciso di non usare un sistema HSM (Hirerachical Storage Management), in quanto le sue diverse implementazioni sono basate su sistemi proprietari.
La comparazione dei costi fra sistemi di memorizzazione in linea e non in linea non è facile: relativamente ai dischi SATA si può dire che il loro costo decresce in ma-

niera proporzionale all'aumento della loro capacità, mentre è difficile stimare i costi complessivi di esercizio (il cosiddetto *total cost of ownership* – TCO) di una soluzione di memorizzazione basata su nastri. Tenuto conto dei pro e dei contro si è deciso che la soluzione più conveniente fosse la memorizzazione in linea su semplici personal computer, facilmente sostituibili (*facilmente sostituibili* significa sostituibili con nessun impatto, o con un impatto trascurabile, sull'architettura complessiva).

L'unico inconveniente in questo approccio è effettivamente un problema ecologico: il consumo di energia e le emissioni di ossido di carbonio. Occorre però osservare che negli ultimi anni i cosiddetti computer ecologici (*green computing)* stanno guadagnando quote crescenti e una diffusa coscienza da parte del mercato. Inoltre si sta sviluppando velocemente la tecnologia delle memorie a stato solido (SSD, Solid State Drive), prive di parti elettromeccaniche in movimento, e questo potrebbe ridurre significativamente nel prossimo futuro il consumo di energia delle apparecchiature di memorizzazione.

Grazie al finanziamento della DGBID (Direzione generale per le biblioteche, gli istituti culturali e il diritto d'autore), l'attuale prototipo di Magazzini digitali sta

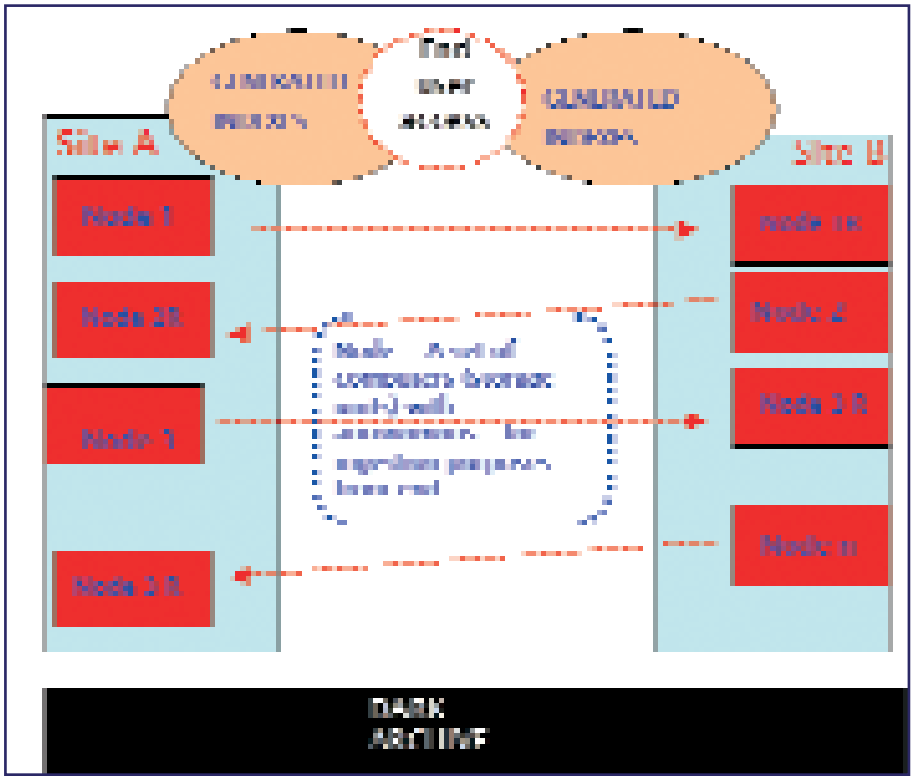

Figura 1. Modello dell'architettura tecnica di Magazzini digitali

ora evolvendo in un servizio operativo basato su due siti principali di deposito, gestiti dalla Biblioteca nazionale centrale di Firenze e dalla Biblioteca nazionale centrale di Roma, e da un *dark archive* gestito dalla Biblioteca nazionale marciana di Venezia. Naturalmente la Fondazione rinascimento digitale continuerà a supportare e promuovere il servizio.
Ciascun sito principale si compone di un insieme di *nodi* indipendenti ed autonomi. A sua volta ogni *nodo è* formato da un insieme di computer che gestiscono in autonomia le attività di acquisizione dei dati (*ingest*). Ciascun *nodo* su un determinato sito ha un corrispondente *nodo replica* sull'altro sito. Magazzini digitali non si basa quindi su un'architettura *sito principale/sito replica (master site mirror site)* dato che ciascun sito contiene, in maniera simmetrica, sia i nodi *master* che i nodi *mirror* (vedi figura 1). Ciascun file fisico è replicato due volte su computer diversi all'interno della stesso nodo. Anche il *dark archive* contiene due copie di ciascun file su due differenti computer. All'interno di Magazzini digitali, dunque, ciascun file fisico è replicato sei volte.
Collocare uno dei siti principali a Firenze sulle rive dell'Arno e l'altro a Venezia in Piazza San Marco, soggetta al noto fenomeno dell'"acqua alta", avrebbe costituito una seria minaccia per la sicurezza complessiva del servizio. Si è dunque deciso di collocare tutto l'hardware presso centri di elaborazione dati esterni, selezionati sulla base del possesso del requisito di base della certificazione secondo lo standard internazionale ISO 27001. Ciascun istituto (BNCF, BNCR, BNM) selezionerà tre differenti centri di elaborazione dati posseduti e gestiti da tre differenti aziende, al fine di ridurre un possibile effetto "domino". Tali centri dovranno inoltre essere distanti l'uno dall'altro non meno di duecento chilometri, al fine di ridurre il rischio derivante da eventi catastrofici naturali. Questa architettura basata sullo standard ISO 27001 formerà la base per una specifica certificazione di Magazzini digitali come archivio digitale affidabile (o fidato). Durante la fase prototipale è stata condotta una sperimentazione con il sistema di autocertificazione DRAMBORA (Digital Repository Audit Method Based on Risk Assessment), e si è tenuto conto anche delle specifiche di TRAC.

Metadati

Il sistema previsto per Magazzini digitali è in grado di accettare due tipi di file:

1. dati racchiusi in contenitori WARC: un contenitore WARC (ISO 28500) aggrega degli oggetti digitali per un agevole stoccaggio in un *file system* convenzionale;
2. metadati racchiusi in contenitori MPEG21-DIDL: MPEG21-DIDL (ISO 21000) è un contenitore semplice ed agnostico adatto per l'archiviazione dei metadati (insiemi di metadati conformi a "schemi" differenti).

All'interno dell'architettura proposta Magazzini digitali deve affrontare il problema della gestione dei metadati, problema che è stato posto in termini di *modello lago/modello fiume:* un archivio per la conservazione permanente non può basarsi su un *modello lago*, cioè sull'aggregazione di metadati conformi a pochi schemi e alimentati da poche fonti principali, ma deve invece gestire la sedimentazione di metadati conformi a schemi che possono cambiare nel tempo e che sono alimentati da fonti molteplici; deve quindi basarsi sul *modello fiume.* In altre parole, in un archivio per la conservazione permanente dovranno convivere schemi di metadati che derivano, per usare il linguaggio di PREMIS (Preservation Metadata Implementation Strategies), da differenti *agent* (per esempio: *harverster* di metadati OAI-PMH – Open Archives Iniziative-Protocol for Metadata Harvesting, estrattori di metadati come JHOVE, gli stessi bibliotecari, ecc.). Ogni schema è soggetto a cambiamenti nel corso del tempo e la sovrapposizione, a livello semantico, di elementi appartenenti a schemi diversi sarà probabilmente la norma piuttosto che l'eccezione.
Dal momento che i metadati sono l'unico mezzo per controllare i dati è essenziale avere il controllo dei metadati al fine di evitare il rischio di un "modello Babele". Attualmente si sta lavorando su questo, tenendo conto che non sembrano ancora esserci strumenti consolidati e disponibili. Ci sono però delle linee di sviluppo interessanti: tavole di corrispondenza (*crosswalks*) come *Morfrom*, un *web service* dimostrativo di OCLC (On-line Computer Library Center), relativo a dati bibliografici, o i piani di sviluppo di *Dspace* che dovrebbe implementare i risultati del progetto SIMILE (Semantic Interoperability of Metadata and Information in unLike Environments), un progetto di ricerca del Massachusetts Institute of Technology (MIT) e di HP che sta studiando come supportare schemi di metadati differenti usando RDF (Resource Description Framework).

Il Modello giuridico e dei servizi

La seconda parte di questo contributo è relativa agli aspetti giuridici e agli accordi che sottostanno al progetto nonché al modello dei servizi.
La più recente normativa italiana sul deposito legale (L. 106/2004, D.P.R. 252/2006) prevede un periodo di sperimentazione del deposito legale su base volontaria dei documenti elettronici, definiti dalla legge «documenti diffusi tramite rete informatica». Tale normativa può essere considerata come la massima fonte di un impegno formalmente affidato alle biblioteche nazionali depositarie di costituire il nucleo di una rete nazionale per la conservazione digitale che, sulla base dell'esito della sperimentazione o limitatamente a specifiche tipologie documentarie, potrebbe comprendere anche le risorse elettroniche prodotte in altri domini, diversi da quelli in cui operano le biblioteche. Come è noto l'affidamento formale del "compito" di conservare le risorse è uno dei prerequisiti di un archivio digitale affidabile (o fidato).

La sperimentazione è finanziata dal Mibac, Direzione generale per le biblioteche, gli Istituti culturali e il Diritto d'autore, con il supporto umano, organizzativo e finanziario della Fondazione rinascimento digitale. Come precedentemente detto, viene condotta dalle BNCF e dalla BNCR, che operano come siti principali per l'accesso e la conservazione delle risorse elettroniche, e dalla BNM che gestisce il *dark archive* fuori linea, non accessibile al pubblico, per la ridondanza dei dati. I tre principali obiettivi della sperimentazione sono i seguenti:

1. implementare un modello organizzativo adatto a costituire gli archivi, nazionale e regionale, della produzione editoriale elettronica, come previsto dalla legge, e tale da poter essere esteso su scala più ampia;
2. implementare un modello di servizio tale da bilanciare gli interessi dei detentori dei diritti della protezione dei contenuti con quelli degli utenti finali all'accesso alle risorse;
3. implementare un sistema tale da assicurare l'accesso e la conservazione permanente dei contenuti digitali, e la loro autenticità (identità ed integrità).

Al fine di raggiungere tali obiettivi e di bilanciare i diversi interessi dei vari soggetti coinvolti (*stakeholder)* sono necessari degli accordi specifici:

1. un accordo fra le tre biblioteche nazionali e la Fondazione rinascimento digitale per definire le responsabilità ed i ruoli di ciascuna istituzione dai diversi punti di vista, scientifico, tecnico, operativo e finanziario e per istituire un Comitato di coordinamento per tutte le attività di gestione, monitoraggio e valutazione dei risultati. Compito del Comitato sarà anche quello di definire un piano per la sostenibilità finanziaria del progetto dopo i 36 mesi di sperimentazione; l'accordo, sotto forma di lettera d'intenti, è stato firmato il 19 gennaio 2010;
2. un accordo fra le tre biblioteche nazionali e ciascun editore elettronico (o ciascun fornitore di contenuti digitali) che parteciperà alla sperimentazione, relativo all'accesso e all'uso delle risorse digitali oggetto di deposito legale, tale da configurare un modello dei servizi. La normativa corrente (art. 38, D.P.R. 252/2006) prevede un accesso libero per via telematica ai documenti soggetti a deposito legale che siano in origine liberamente accessibili in rete, e un accesso limitato esclusivamente a utenti registrati che accedono da postazioni situate all'interno degli istituti depositari per quei documenti il cui accesso è originariamente soggetto a licenze o condizioni particolari. In ambedue i casi l'accesso deve avvenire nel rispetto delle norme sul diritto d'autore e sui diritti connessi. L'accordo, oramai concluso, prevede i seguenti punti:
   - BNCF e BNCR effettueranno periodicamente la raccolta (*harvesting)* dei documenti elettronici concordati con gli editori (l'*harvesting* è la modalità più

semplice ed economica di alimentare l'archivio, anche dal punto di vista degli editori, a condizione che sia rispettata la normativa sul diritto d'autore);
- nel caso di documenti accessibili su licenza, l'editore fornirà alle biblioteche le necessarie autorizzazioni, e verranno concordati i formati dei file (WARC etc.);
- i documenti verranno immagazzinati in copie multiple (minimo 6 copie) in BNCF e BNCR, ed offline in BNM;
- le biblioteche saranno autorizzate a depositare i documenti presso *data center* esterni, certificati ISO 27001;
- gli archivi digitali saranno conformi ad OAIS (ISO 14721-2003) e saranno certificati come affidabili;
- BNCF, BNCR e BNM assicureranno l'accessibilità e la conservazione permanente dei documenti depositati, e manterranno traccia di qualunque loro modifica, fornendo un rendiconto trimestrale all'editore;
- BNCF, BNCR e BNM saranno autorizzate ad effettuare tutte le operazioni necessarie al mantenimento dell'accessibilità e della conservazione permanente dei documenti depositati (duplicazioni, migrazioni etc.);
- i documenti depositati soggetti a licenza saranno resi consultabili solo da utenti registrati su postazioni multiple prive di stampanti, porte USB etc. poste sulle reti locali di BNCF e BNCR; tutte le operazioni svolte dagli utenti saranno tracciate, secondo le normative vigenti;
- la stampa o il *download* di file sarà soggetto a specifici accordi, e sarà previsto ove necessario un sistema di remunerazione del diritto d'autore (es. per i documenti protetti non disponibili sul sito dell'editore);
- l'accessibilità e la consultazione dei documenti depositati sarà consentita anche alle biblioteche regionali di deposito, con le stesse modalità, ma limitatamente ai documenti prodotti dagli editori la cui sede si trova nella stessa regione della biblioteca regionale di deposito.

Inoltre, al fine di estendere la base della sperimentazione, il progetto prevede il deposito anche dei seguenti tipi di risorse elettroniche, da regolare anch'esso tramite specifici accordi:

1. risorse digitali native prodotte dalle università e soggette anch'esse a deposito, con particolare riferimento alle tesi di dottorato;
2. risorse digitali risultanti dai progetti di digitalizzazione di materiali analogici finanziati dalla Biblioteca Digitale Italiana, soprattutto nell'ambito delle istituzioni della memoria e limitatamente alle copie master.

Per quanto riguarda il primo di questi punti è stato attivato sperimentalmente il sito del deposito legale, in cui compare una pagina informativa sulle procedure di

Errata corrige

Si ripete, nella versione online, il testo corredato delle note mancanti nella versione cartacea.

Magazzini digitali: dal prototipo al servizio

Giovanni Bergamin – Maurizio Messina
Biblioteca nazionale centrale di Firenze – Biblioteca nazionale marciana di Venezia

Introduzione[1]

Il progetto Magazzini digitali, avviato nel 2006 dalla Fondazione rinascimento digitale, dalla Biblioteca nazionale centrale di Firenze e dalla Biblioteca nazionale centrale di Roma si propone ora di mettere a regime un sistema per la conservazione permanente dei documenti elettronici pubblicati in Italia e diffusi tramite rete informatica, in attuazione della normativa sul deposito legale (L. 106/2004, d.p.r. 252/2006).
Nella prima parte di questo contributo verrà descritta l'architettura tecnica del progetto, ma è bene tenere presente fin d'ora che la conservazione digitale, come è oramai ben noto, non si esaurisce solo in procedure di natura tecnologica. Le strategie volte ad evitare la perdita dei bit o a prevenire le dipendenze dall'hardware o dal software sono infatti solo una parte del problema. Vanno tenute nel conto dovuto le implicazioni economiche (la sostenibilità), la necessità di selezionare cosa è necessario conservare per le generazioni future, gli aspetti legali legati alla normativa sul diritto d'autore, la necessità della cooperazione fra le istituzioni titolari del deposito legale[2]. Alcuni di questi aspetti verranno esaminati nella seconda parte di questo contributo.
Ai fini del progetto, la conservazione digitale può essere definita come un servizio pubblico fornito da depositi *digitali affidabili* o *fidati* (*trusted* o *trustworthy digital repositories*) in grado di assicurare, per le risorse digitali depositate, la *leggibilità* a livello di bit da parte di una macchina (*viability*), la *interpretabilità* a livello di formato (per esempio: pdf, doc, ecc.) da parte di un elaboratore (*renderability*), l'*autenticità* (*authenticity*) intesa come identità e integrità dell'oggetto digitale, e la effettiva *disponibilità* (*availability*) per le comunità designate (comunità di riferimento, interessate all'uso di quelle risorse)[3].

[1] Tutti i link sono stati controllati il 27 aprile 2010.

[2] Brian Lavoie – Lorcan Dempsey, *Thirteen ways of looking at... digital preservation*, «D-lib magazine», 10 (2004), 7/8, http://www.dlib.org/dlib/july04/lavoie/07lavoie.html.

[3] Queste definizioni sono basate su:
- *Trustworthy Repositories Audit & Certification (TRAC): Criteria and Checklist*, http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf (per il concetto di "*trusted digital repositories*");
- Luciana Duranti, *Un quadro teorico per le politiche, le strategie e gli standards di conservazione digitale: la prospettiva concettuale di InterPARES*, «Bibliotime», 9 (2006), 1,

Il nome del progetto richiama intenzionalmente i magazzini delle biblioteche titolari del deposito legale. Come definito da uno storico progetto europeo sulla conservazione digitale (Networked European Deposit Library – NEDLIB[4], svoltosi dal 1997 al 2000):

> «For us, as memory organizations, this means we have to move from paper-based stacks to digital stacks».

Per molti aspetti i Magazzini digitali sono comparabili a quelli convenzionali: le risorse digitali devono essere conservate indefinitamente; i Magazzini digitali crescono man mano che si aggiungono nuove risorse; modifiche o cancellazioni di risorse non sono di norma possibili; è impossibile predefinire la frequenza d'uso delle risorse, alcune delle quali non saranno mai utilizzate, o lo saranno raramente[5]. Ed è interessante notare che nove anni più tardi, una ricerca su Google dei termini *Digital Stacks* restituisce la medesima espressione usata nel contesto della conservazione digitale:

> «Digital stacks: rather than boxes, shelves, and climate controlled environments, digital information must be stored in containers, file systems, and secure servers»[6].

L'Architettura tecnologica

Lo scopo del progetto è stato quello di impiantare un'infrastruttura tecnologica con caratteristiche di "permanenza". Dando per assodato che i guasti o le disfunzioni dei vari componenti sono la norma piuttosto che l'eccezione[7], l'infrastruttura è basata sulla replica dei dati (macchine differenti collocate in luoghi differenti) e su componenti hardware semplici e universalmente diffusi, non dipendenti dai produttori, e che possono essere sostituiti facilmente: in altre parole, semplici personal computer. Inoltre, l'infrastruttura non è dipendente da software proprietario ma si basa su sistemi operativi e servizi a codice sorgente aperto (*open source*).

http://didattica.spbo.unibo.it/bibliotime/num-ix-1/duranti.htm (per rendere valutabile l'autenticità di una risorsa digitale, un pubblico servizio deve essere in grado di stabilire la sua identità e di dimostrare la sua integrità);
- PREMIS 2.0, *PREsevation Metadata: Implementation Strategies*, 2008, http://www.loc.gov/standards/premis/ (per i concetti di "*viability*, *renderability*, *understandability*, *authenticity*, *identity*");
- *OAIS: Reference model for an Open Archival Information System*, ISO 14721:2003 (per il concetto di archivio e di comunità designata: «an organization that intends to preserve information for access and use by a designated community»).

[4] http://nedlib.kb.nl/.

[5] Jim Linden – Sean Martin – Richard Masters – Roderic Parker, *Technology Watch Report: The large-scale archival storage of digital objects*, 2005, http://www.dpconline.org/docs/dpctw04-03.pdf.

[6] http://www.pedalspreservation.org/About/stacks.aspx.

[7] Sanjay Ghemawat – Howard Gobioff – Shun-Tak Leung, *The Google file system*, 2003, http://labs.google.com/papers/gfs-sosp2003.pdf.

Attualmente un normale personal computer può facilmente immagazzinare fino ad 8 terabyte di dati, su 4 dischi da 2000 gigabyte, usando tecnologie SATA diffusissime ed economiche[8]. La replica dei dati si basa su comuni programmi di utilità (*utility*) per la sincronizzazione dei dischi come rsync[9] e, per evitare dipendenze hardware come ad esempio le dipendenze da un determinato *disk controller*, non viene utilizzato il RAID (Redundant Array of Independent Disks)[10].

Nel passaggio dal prototipo al servizio, inoltre, è stata modificata l'architettura tecnologica del *dark archive*. Il progetto originario prevedeva infatti l'uso di un sistema di *memorizzazione non in linea* (*offline storage*) basato su nastri di tipo LTO (Linear Tape-Open)[11]; successivamente si è deciso di utilizzare la medesima tecnologia già individuata per i due siti principali, cioè la *memorizzazione in linea* (*online storage*) su normali personal computer. L'espressione "in linea", comunque, non cambia la funzione del *dark archive*, che è quella di servire da archivio di sicurezza dei dati usabile in caso di *disaster recovery*[12]. I nastri LTO sono sicuramente una soluzione robusta ed affidabile introducono dipendenze e vincoli di natura tecnologica e gestionale (per esempio librerie automatizzate o robot). Per lo stesso motivo si è deciso di non usare un sistema HSM (Hirerachical Storage Management)[13], in quanto le sue diverse implementazioni sono basate su sistemi proprietari.

La comparazione dei costi fra sistemi di memorizzazione in linea e non in linea non è facile: relativamente ai dischi SATA si può dire che il loro costo decresce in maniera proporzionale all'aumento della loro capacità, mentre è difficile stimare i costi complessivi di esercizio (il cosiddetto *total cost of ownership* – TCO) di una soluzione di memorizzazione basata su nastri[14]. Tenuto conto dei pro e dei contro si è deciso che la soluzione più conveniente fosse la memorizzazione in linea su semplici personal computer, facilmente sostituibili (*facilmente sostituibili* significa sostituibili con nessun impatto, o con un impatto trascurabile, sull'architettura complessiva).

L'unico inconveniente in questo approccio è effettivamente un problema ecologico: il consumo di energia e le emissioni di ossido di carbonio. Occorre però osservare che negli ultimi anni i cosiddetti computer ecologici (*green compu-*

[8] http://it.wikipedia.org/wiki/Serial_ATA.

[9] «Rsync è un software per Unix che sincronizza file e cartelle da una posizione all'altra minimizzando il trasferimento di dati», http://it.wikipedia.org/wiki/Rsync.

[10] «RAID è un sistema informatico che usa un insieme di dischi rigidi per condividere o replicare le informazioni, combinandoli in una sola unità logica», http://it.wikipedia.org/wiki/RAID.

[11] http://en.wikipedia.org/wiki/Linear_Tape-Open.

[12] http://www.webopedia.com/TERM/D/dark_archive.html, http://it.wikipedia.org/wiki/Disaster_recovery.

[13] http://en.wikipedia.org/wiki/Hierarchical_storage_management. HSM è una tecnica di gestione dello *storage* che sposta automaticamente i dati da apparati più veloci e di maggiore costo ad altri più economici e meno efficienti, a seconda dei diversi casi d'uso dei dati stessi.

[14] http://digitalcuration.blogspot.com/2009/07/online-and-offline-storage-cost-and.html.

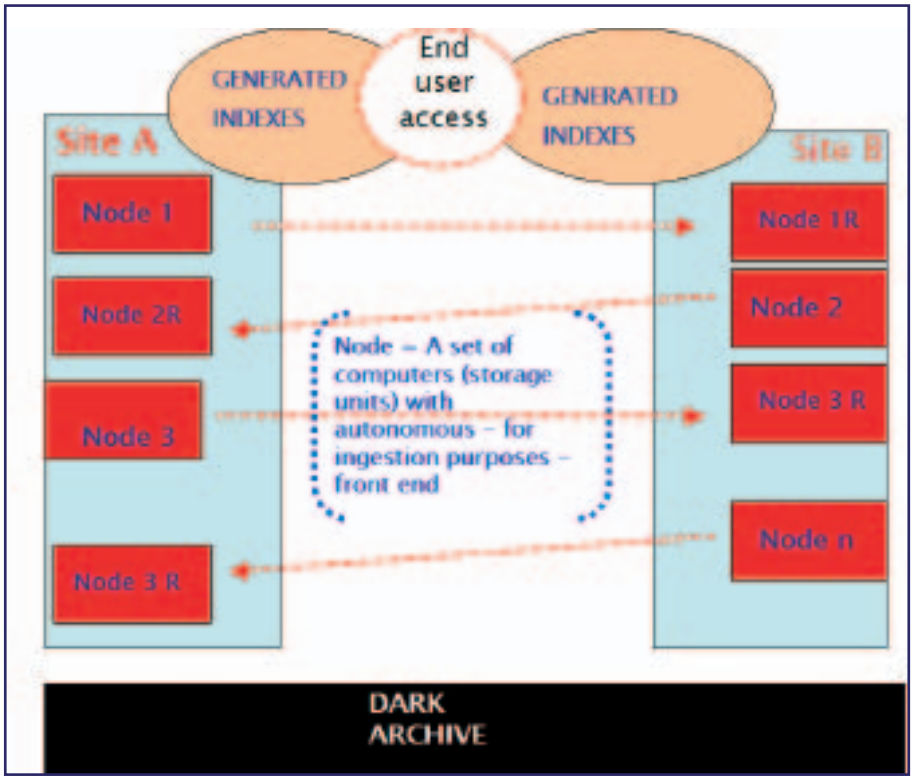

Figura 1. Modello dell'architettura tecnica di Magazzini digitali

ting)[15] stanno guadagnando quote crescenti e una diffusa coscienza da parte del mercato. Inoltre si sta sviluppando velocemente la tecnologia delle memorie a stato solido (SSD, Solid State Drive)[16], prive di parti elettromeccaniche in movimento, e questo potrebbe ridurre significativamente nel prossimo futuro il consumo di energia delle apparecchiature di memorizzazione.

Grazie al finanziamento della DGBID (Direzione generale per le biblioteche, gli istituti culturali e il diritto d'autore), l'attuale prototipo di Magazzini digitali sta ora evolvendo in un servizio operativo basato su due siti principali di deposito, gestiti dalla Biblioteca nazionale centrale di Firenze e dalla Biblioteca nazionale centrale di Roma, e da un *dark archive* gestito dalla Biblioteca nazionale marciana di Venezia[17]. Naturalmente la Fondazione rinascimento digitale continuerà a supportare e promuovere il servizio.

[15] «This includes archival and backup data that would formerly have been saved on tape or other offline storage. The increase in online storage has increased power consumption. Reducing the power consumed by large storage arrays, while still providing the benefits of online storage, is a subject of ongoing research», http://en.wikipedia.org/wiki/Green_computing.

[16] http://en.wikipedia.org/wiki/Solid-state_drive.

[17] Il servizio operativo è stato finanziato per tre anni alla fine del 2009.

Ciascun sito principale si compone di un insieme di *nodi* indipendenti ed autonomi. A sua volta ogni *nodo è* formato da un insieme di computer che gestiscono in autonomia le attività di acquisizione dei dati (*ingest*). Ciascun *nodo* su un determinato sito ha un corrispondente *nodo replica* sull'altro sito. Magazzini digitali non si basa quindi su un'architettura *sito principale/sito replica* (*master site mirror site*) dato che ciascun sito contiene, in maniera simmetrica, sia i nodi *master* che i nodi *mirror* (vedi figura 1). Ciascun file fisico è replicato due volte su computer diversi all'interno della stesso nodo. Anche il *dark archive* contiene due copie di ciascun file su due differenti computer. All'interno di Magazzini digitali, dunque, ciascun file fisico è replicato sei volte.
Collocare uno dei siti principali a Firenze sulle rive dell'Arno e l'altro a Venezia in Piazza San Marco, soggetta al noto fenomeno dell'"acqua alta", avrebbe costituito una seria minaccia per la sicurezza complessiva del servizio. Si è dunque deciso di collocare tutto l'hardware presso centri di elaborazione dati esterni, selezionati sulla base del possesso del requisito di base della certificazione secondo lo standard internazionale ISO 27001[18]. Ciascun istituto (BNCF, BNCR, BNM) selezionerà tre differenti centri di elaborazione dati posseduti e gestiti da tre differenti aziende, al fine di ridurre un possibile effetto "domino". Tali centri dovranno inoltre essere distanti l'uno dall'altro non meno di duecento chilometri, al fine di ridurre il rischio derivante da eventi catastrofici naturali. Questa architettura basata sullo standard ISO 27001 formerà la base per una specifica certificazione di Magazzini digitali come archivio digitale affidabile (o fidato). Durante la fase prototipale è stata condotta una sperimentazione con il sistema di autocertificazione DRAMBORA (Digital Repository Audit Method Based on Risk Assessment)[19], e si è tenuto conto anche delle specifiche di TRAC[20].

Metadati

Il sistema previsto per Magazzini digitali è in grado di accettare due tipi di file:

1. dati racchiusi in contenitori WARC: un contenitore WARC (ISO 28500) aggrega degli oggetti digitali per un agevole stoccaggio in un *file system* convenzionale[21];

[18] ISO/IEC 27001:2005 «specifies the requirements for establishing, implementing, operating, monitoring, reviewing, maintaining and improving a documented Information Security Management System within the context of the organization's overall business risks».
[19] http://www.repositoryaudit.eu/.
[20] *Trustworthy Repositories Audit & Certification (TRAC)* cit.
[21] ISO 28500:2009 specifica il formato di un file WARC:
– «to store both the payload content and control information from mainstream Internet application layer protocols, such as the Hypertext Transfer Protocol (HTTP), Domain Name System (DNS), and File Transfer Protocol (FTP);

2. metadati racchiusi in contenitori MPEG21-DIDL: MPEG21-DIDL (ISO 21000) è un contenitore semplice ed agnostico adatto per l'archiviazione dei metadati (insiemi di metadati conformi a "schemi" differenti)[22].

All'interno dell'architettura proposta Magazzini digitali deve affrontare il problema della gestione dei metadati, problema che è stato posto in termini di *modello lago/modello fiume*[23]*:* un archivio per la conservazione permanente non può basarsi su un *modello lago*, cioè sull'aggregazione di metadati conformi a pochi schemi[24] e alimentati da poche fonti principali, ma deve invece gestire la sedimentazione di metadati conformi a schemi che possono cambiare nel tempo e che sono alimentati da fonti molteplici; deve quindi basarsi sul *modello fiume.* In altre parole, in un archivio per la conservazione permanente dovranno convivere schemi di metadati che derivano, per usare il linguaggio di PREMIS (Preservation Metadata Implementation Strategies)[25], da differenti *agent* (per esempio: *harverster* di meta-

– to store arbitrary metadata linked to other stored data (e.g. subject classifier, discovered language, encoding);
– to support data compression and maintain data record integrity;
– to store all control information from the harvesting protocol (e.g. request headers), not just response information;
– to store the results of data transformations linked to other stored data;
– to store a duplicate detection event linked to other stored data (to reduce storage in the presence of identical or substantially similar resources);
– to be extended without disruption to existing functionality;
– to support handling of overly long records by truncation or segmentation, where desired»,
http://www.iso.org/iso/catalogue_detail.htm?csnumber=44717.

[22] ISO/IEC 21000-2:2005 specifica:
– «Model: The Digital Item Declaration Model describes a set of abstract terms and concepts to form a useful model for defining Digital Items.
– Representation: The Digital Item Declaration Language (DIDL) is based upon the terms and concepts defined in the above model. It contains the normative description of the syntax and semantics of each of the DIDL elements, as represented in XML.
– Schema: Informative XML schemas illustrating complete grammars for representation of the DID in XML conforming to the normative representation.
– Detailed Examples: Illustrative (non-normative) examples of DIDL documents are provided to aid in understanding the use of the specification and its potential applications.
The ISO/IEC 21000 (MPEG-21) series of International Standards defines an open framework for multimedia delivery and consumption, with both the content creator and content consumer as focal points. The vision for MPEG-21 is to define a multimedia framework to enable transparent and augmented use of multimedia resources across a wide range of networks and devices used by different communities.
This second part of MPEG-21 (ISO/IEC 21000-2:2005) specifies a uniform and flexible abstraction and interoperable representation for declaring the structure and makeup of Digital Items. A Digital Item Declaration (DID) involves specifying the resources, metadata, and their interrelationships for a Digital Item. A DID is done using the Digital Item Declaration Language (DIDL)»,
http://www.iso.org/iso/catalogue_detail.htm?csnumber=41112.

[23] http://orweblog.oclc.org/archives/001754.html.

[24] Il termine *Schema* è usato qui come definito in http://www.w3.org/XML/Schema: «XML Schemas express shared vocabularies and allow machines to carry out rules made by people».

[25] http://www.loc.gov/standards/premis/.

dati OAI-PMH – Open Archives Iniziative-Protocol for Metadata Harvesting[26], estrattori di metadati come JHOVE[27], gli stessi bibliotecari, ecc.). Ogni schema è soggetto a cambiamenti nel corso del tempo e la sovrapposizione, a livello semantico, di elementi appartenenti a schemi diversi sarà probabilmente la norma piuttosto che l'eccezione.
Dal momento che i metadati sono l'unico mezzo per controllare i dati è essenziale avere il controllo dei metadati al fine di evitare il rischio di un "modello Babele". Attualmente si sta lavorando su questo, tenendo conto che non sembrano ancora esserci strumenti consolidati e disponibili. Ci sono però delle linee di sviluppo interessanti: tavole di corrispondenza (*crosswalks*) come *Morfrom*[28], un *web service* dimostrativo di OCLC (On-line Computer Library Center), relativo a dati bibliografici, o i piani di sviluppo di *Dspace*[29] che dovrebbe implementare i risultati del progetto SIMILE (Semantic Interoperability of Metadata and Information in unLike Environments)[30], un progetto di ricerca del Massachusetts Institute of Technology (MIT) e di HP che sta studiando come supportare schemi di metadati differenti usando RDF (Resource Description Framework)[31].

Il Modello giuridico e dei servizi

La seconda parte di questo contributo è relativa agli aspetti giuridici e agli accordi che sottostanno al progetto nonché al modello dei servizi.
La più recente normativa italiana sul deposito legale (L. 106/2004, D.P.R. 252/2006) prevede un periodo di sperimentazione del deposito legale su base volontaria dei documenti elettronici, definiti dalla legge «documenti diffusi tramite rete informatica»[32]. Tale normativa può essere considerata come la massima fonte di un impegno formalmente affidato alle biblioteche nazionali depositarie di costituire il nucleo di una rete nazionale per la conservazione digitale che, sulla base dell'esito della sperimentazione o limitatamente a specifiche tipologie documentarie, potrebbe comprendere anche le risorse elettroniche prodotte in altri domini, diversi da quelli in cui operano le biblioteche. Come è noto l'affidamento formale del "compito" di conservare le risorse è uno dei prerequisiti di un archivio digitale affidabile (o fidato)[33].
La sperimentazione è finanziata dal Mibac, Direzione generale per le biblioteche, gli Istituti culturali e il Diritto d'autore, con il supporto umano, organizzativo e finanziario della Fondazione rinascimento digitale. Come precedentemente detto, viene condotta dalle BNCF e dalla BNCR, che operano come siti principali per l'ac-

[26] http://www.openarchives.org/OAI/openarchivesprotocol.html.
[27] http://hul.harvard.edu/jhove/index.html.
[28] http://journal.code4lib.org/articles/54.
[29] http://www.dspace.org/, in particolare: http://www.dspace.org/faq/FAQ.html.
[30] http://simile.mit.edu/.
[31] http://it.wikipedia.org/wiki/Resource_Description_Framework.
[32] L. 106/2004, art. 4.
[33] *Trustworthy Repositories Audit & Certification (TRAC)* cit.

cesso e la conservazione delle risorse elettroniche, e dalla BNM che gestisce il *dark archive* fuori linea, non accessibile al pubblico, per la ridondanza dei dati. I tre principali obiettivi della sperimentazione sono i seguenti:

1. implementare un modello organizzativo adatto a costituire gli archivi, nazionale e regionale, della produzione editoriale elettronica, come previsto dalla legge, e tale da poter essere esteso su scala più ampia;
2. implementare un modello di servizio tale da bilanciare gli interessi dei detentori dei diritti della protezione dei contenuti con quelli degli utenti finali all'accesso alle risorse;
3. implementare un sistema tale da assicurare l'accesso e la conservazione permanente dei contenuti digitali, e la loro autenticità (identità ed integrità).

Al fine di raggiungere tali obiettivi e di bilanciare i diversi interessi dei vari soggetti coinvolti (*stakeholder*) sono necessari degli accordi specifici:

1. un accordo fra le tre biblioteche nazionali e la Fondazione rinascimento digitale per definire le responsabilità ed i ruoli di ciascuna istituzione dai diversi punti di vista, scientifico, tecnico, operativo e finanziario e per istituire un Comitato di coordinamento per tutte le attività di gestione, monitoraggio e valutazione dei risultati. Compito del Comitato sarà anche quello di definire un piano per la sostenibilità finanziaria del progetto dopo i 36 mesi di sperimentazione; l'accordo, sotto forma di lettera d'intenti, è stato firmato il 19 gennaio 2010;
2. un accordo fra le tre biblioteche nazionali e ciascun editore elettronico (o ciascun fornitore di contenuti digitali) che parteciperà alla sperimentazione, relativo all'accesso e all'uso delle risorse digitali oggetto di deposito legale, tale da configurare un modello dei servizi. La normativa corrente (art. 38, D.P.R. 252/2006) prevede un accesso libero per via telematica ai documenti soggetti a deposito legale che siano in origine liberamente accessibili in rete, e un accesso limitato esclusivamente a utenti registrati che accedono da postazioni situate all'interno degli istituti depositari per quei documenti il cui accesso è originariamente soggetto a licenze o condizioni particolari. In ambedue i casi l'accesso deve avvenire nel rispetto delle norme sul diritto d'autore e sui diritti connessi. L'accordo, oramai concluso, prevede i seguenti punti:
   - BNCF e BNCR effettueranno periodicamente la raccolta (*harvesting*) dei documenti elettronici concordati con gli editori (l'*harvesting* è la modalità più semplice ed economica di alimentare l'archivio, anche dal punto di vista degli editori, a condizione che sia rispettata la normativa sul diritto d'autore);
   - nel caso di documenti accessibili su licenza, l'editore fornirà alle biblioteche le necessarie autorizzazioni, e verranno concordati i formati dei file (WARC etc.);
   - i documenti verranno immagazzinati in copie multiple (minimo 6 copie) in BNCF e BNCR, ed offline in BNM;

- le biblioteche saranno autorizzate a depositare i documenti presso *data center* esterni, certificati ISO 27001;
- gli archivi digitali saranno conformi ad OAIS (ISO 14721-2003) e saranno certificati come affidabili;
- BNCF, BNCR e BNM assicureranno l'accessibilità e la conservazione permanente dei documenti depositati, e manterranno traccia di qualunque loro modifica, fornendo un rendiconto trimestrale all'editore;
- BNCF, BNCR e BNM saranno autorizzate ad effettuare tutte le operazioni necessarie al mantenimento dell'accessibilità e della conservazione permanente dei documenti depositati (duplicazioni, migrazioni, etc.);
- i documenti depositati soggetti a licenza saranno resi consultabili solo da utenti registrati su postazioni multiple prive di stampanti, porte USB etc. poste sulle reti locali di BNCF e BNCR; tutte le operazioni svolte dagli utenti saranno tracciate, secondo le normative vigenti;
- la stampa o il *download* di file sarà soggetto a specifici accordi, e sarà previsto ove necessario un sistema di remunerazione del diritto d'autore (es. per i documenti protetti non disponibili sul sito dell'editore);
- l'accessibilità e la consultazione dei documenti depositati sarà consentita anche alle biblioteche regionali di deposito, con le stesse modalità, ma limitatamente ai documenti prodotti dagli editori la cui sede si trova nella stessa regione della biblioteca regionale di deposito.

Inoltre, al fine di estendere la base della sperimentazione, il progetto prevede il deposito anche dei seguenti tipi di risorse elettroniche, da regolare anch'esso tramite specifici accordi:

1. risorse digitali native prodotte dalle università e soggette anch'esse a deposito, con particolare riferimento alle tesi di dottorato;
2. risorse digitali risultanti dai progetti di digitalizzazione di materiali analogici finanziati dalla Biblioteca Digitale Italiana[34], soprattutto nell'ambito delle istituzioni della memoria e limitatamente alle copie master.

Per quanto riguarda il primo di questi punti è stato attivato sperimentalmente il sito[35] del deposito legale, in cui compare una pagina informativa sulle procedure di deposito legale delle tesi di dottorato presso le biblioteche nazionali centrali.
Il servizio è stato realizzato e sperimentato in collaborazione con il Gruppo Open Access della CRUI (Conferenza dei Rettori delle Università italiane)[36].

[34] http://www.iccu.sbn.it/genera.jsp?s=18&l=it.
[35] http://www.depositolegale.it/oai.html.
[36] La raccolta automatica (*harvesting*) dei dati e dei metadati delle tesi di dottorato di ricerca ai fini del deposito legale è prevista dalla Circolare MiUR n. 1746 del 20 luglio 2007.

La procedura di raccolta automatica (*harvesting*) delle tesi consente a «tutte le Università italiane che raccolgono le tesi di dottorato in formato digitale in un archivio aperto, secondo le raccomandazioni contenute nelle *Linee guida per il deposito delle tesi di dottorato negli archivi aperti*[37] approvate dalla Conferenza dei Rettori delle Università Italiane del 2007, di ottemperare agli obblighi di legge (senza ricorrere all'invio di documentazione cartacea)». Sono previste anche la raccolta e la conservazione delle tesi soggette ad embargo, che saranno consultabili solo all'interno delle reti locali delle biblioteche nazionali depositarie su postazioni prive di apparecchiature periferiche. La pagina informativa precisa inoltre:

1. il formato preferito per il deposito ovvero il PDF(A)[38];
2. raccomandazioni per i metadati esposti dai *repository* delle Università mediante il protocollo OAI-PMH per facilitare l'interoperabilità sintattica e semantica;
3. specifiche tecniche per la raccolta di tesi digitali suddivise in più file, e indicazioni per la configurazione di EPrints3 e DSpace 1.5.

Alla sperimentazione della procedura di deposito legale delle tesi in formato digitale via *harvesting* automatico hanno partecipato l'Alma Mater Studiorum – Università di Bologna, l'Università Federico II di Napoli e l'Università di Trieste. Hanno contribuito alla definizione delle specifiche tecniche l'Alma Mater Studiorum – Università di Bologna per il software EPrints e l'Università di Trieste per il software DSpace. Hanno successivamente aderito alla sperimentazione la LUISS (Libera università internazionale degli studi sociali Guido Carli di Roma), l'Università di Parma, l'Università Cattolica di Milano, l'Università degli studi di Milano-Bicocca e l'Università di Venezia[39].
L'ultimo aspetto da affrontare brevemente è relativo alla sostenibilità finanziaria del progetto Magazzini digitali: come è noto, l'accesso ai periodici elettronici è di norma soggetto ad una licenza. Una tipica clausola di tali licenze riguarda l'"accesso perpetuo" ai contenuti, che l'editore si impegna a garantire. Si tratta di una clausola di grande importanza sia per le biblioteche che per i loro utenti, e costituisce l'unico modo in cui le biblioteche possono garantire nel tempo la disponibilità di contenuti per i quali hanno sostenuto dei costi. Allo stesso tempo è una clausola che può essere rispettata solo attraverso la predisposizione di un'infrastruttura tecnica ed organizzativa dedicata (o terza), cioè un archivio digitale affidabile (o fidato); un archivio che è improbabile che gli editori abbiano interesse a gestire. Questo tipo di servizio potrebbe dunque essere affidato alla rete delle biblioteche di deposito legale, ed il suo costo potrebbe essere parte delle negoziazioni con gli editori per le licenze elettroniche[40].

[37] http://www.crui.it/HomePage.aspx?ref=1149#.
[38] http://www.iso.org/iso/catalogue_detail?csnumber=38920.
[39] Alla data di chiusura di questo articolo (18 maggio 2010).
[40] Terry Morrow – Neil Beagrie – Maggie Jones – Julia Chruszcz, *A comparative study of e-journals archiving solutions: A JISC funded investigation: Final report*, 2008, http://www.slainte.org.uk/news/archive/0805/jiscejournalreport.pdf.

deposito legale delle tesi di dottorato presso le biblioteche nazionali centrali. Il servizio è stato realizzato e sperimentato in collaborazione con il Gruppo Open Access della CRUI (Conferenza dei Rettori delle Università italiane). La procedura di raccolta automatica (*harvesting)* delle tesi consente a «tutte le Università italiane che raccolgono le tesi di dottorato in formato digitale in un archivio aperto, secondo le raccomandazioni contenute nelle *Linee guida per il deposito delle tesi di dottorato negli archivi aperti* approvate dalla Conferenza dei Rettori delle Università Italiane del 2007, di ottemperare agli obblighi di legge (senza ricorrere all'invio di documentazione cartacea)». Sono previste anche la raccolta e la conservazione delle tesi soggette ad embargo, che saranno consultabili solo all'interno delle reti locali delle biblioteche nazionali depositarie su postazioni prive di apparecchiature periferiche. La pagina informativa precisa inoltre:

1. il formato preferito per il deposito ovvero il PDF(A),
2. raccomandazioni per i metadati esposti dai *repository* delle Università mediante il protocollo OAI-PMH per facilitare l'interoperabilità sintattica e semantica;
3. specifiche tecniche per la raccolta di tesi digitali suddivise in più file, e indicazioni per la configurazione di EPrints3 e DSpace 1.5.

Alla sperimentazione della procedura di deposito legale delle tesi in formato digitale via *harvesting* automatico hanno partecipato l'Alma Mater Studiorum – Università di Bologna, l'Università Federico II di Napoli e l'Università di Trieste. Hanno contribuito alla definizione delle specifiche tecniche l'Alma Mater Studiorum – Università di Bologna per il software EPrints e l'Università di Trieste per il software DSpace. Hanno successivamente aderito alla sperimentazione la LUISS (Libera università internazionale degli studi sociali Guido Carli di Roma), l'Università di Parma, l'Università Cattolica di Milano, l'Università degli studi di Milano-Bicocca e l'Università di Venezia.
L'ultimo aspetto da affrontare brevemente è relativo alla sostenibilità finanziaria del progetto Magazzini digitali: come è noto, l'accesso ai periodici elettronici è di norma soggetto ad una licenza. Una tipica clausola di tali licenze riguarda l'"accesso perpetuo" ai contenuti, che l'editore si impegna a garantire. Si tratta di una clausola di grande importanza sia per le biblioteche che per i loro utenti, e costituisce l'unico modo in cui le biblioteche possono garantire nel tempo la disponibilità di contenuti per i quali hanno sostenuto dei costi. Allo stesso tempo è una clausola che può essere rispettata solo attraverso la predisposizione di un'infrastruttura tecnica ed organizzativa dedicata (o terza), cioè un archivio digitale affidabile (o fidato); un archivio che è improbabile che gli editori abbiano interesse a gestire. Questo tipo di servizio potrebbe dunque essere affidato alla rete delle biblioteche di deposito legale, ed il suo costo potrebbe essere parte delle negoziazioni con gli editori per le licenze elettroniche.