Giornata della documentazione di fonte pubblica: vent'anni di evoluzione dell'informazione nel settore pubblico

Roma, 4 dicembre 2017

DFP
Documentazione di Fonte Pubblica in rete

Progetti di conservazione del patrimonio di fonte pubblica digitale

Giovanni Bergamin
giovanni.bergamin@gmail.com
Chiara Storti
chiara.storti@beniculturali.it

La DFP e le raccolte delle biblioteche

Un percorso di 20 anni con punti fermi:

- la parte più rilevante dell'informazione "pubblica" è in rete
- sono necessari strumenti che facilitino l'accesso
- anche in questo campo il ruolo di mediazione delle biblioteche è fondamentale

dfp.aib.it > Chi siamo

La DFP e l'ecosistema dell'accesso all'informazione

- catalogare (o creare metadati) per facilitare l'accesso non basta (soprattutto se l'informazione nasce e vive in rete)
- fornire l'indirizzo di una risorsa in rete vs. garantire l'accesso nel tempo (due cose ben diverse)
- il deposito legale è un elemento fondamentale nell'ecosistema dell'accesso all'informazione (anche e soprattutto nel mondo del digitale)

Il deposito legale del digitale nativo (1)

D.P.R. 03/05/2006, n. 252, art 37
Regolamento deposito legale

1. Le modalità di deposito dei documenti *diffusi tramite rete informatica* sono definite con successivo regolamento
2. Il Ministero promuove forme volontarie di sperimentazione del deposito

depositolegale.it

(p.s.: nei primi mesi del 2018 dovrebbe uscire il Regolamento per il digitale nativo)

...

Il deposito legale del digitale nativo (2)

D.P.R. 03/05/2006, n. 252, art 37
Regolamento deposito legale

Priorità raccomandate nella fase di sperimentazione (comma 3):

"a) documenti che assicurino la continuità delle collezioni già avviate, anche su supporti e mediante tecnologie tradizionali;
b) documenti concernenti la produzione scientifica delle università, dei centri di ricerca e delle istituzioni culturali;
c) documenti elaborati e messi in rete da soggetti pubblici"

depositolegale.it

Magazzini digitali per il deposito legale (1)

MD è l'infrastruttura nazionale che offre il servizio coordinato di conservazione e accesso a lungo termine per le pubblicazioni digitali italiane acquisite attraverso il deposito legale (L. 106/2004, DPR 252/2006).

Il prototipo di MD è stato implementato alla fine del 2006 con il coinvolgimento delle Biblioteche Nazionali Centrali di Firenze e di Roma (BNCF, BNCR) e della Fondazione Rinascimento Digitale (FRD)

depositolegale.it

Magazzini digitali per il deposito legale (2)

2010 - Lettera d'intenti (19.1. 2010) tra BNCF/BNCR (soggetti depositari), Biblioteca Marciana (per il "dark archive") e FRD per la realizzazione dell'infrastruttura per il servizio nazionale di deposito legale delle pubblicazioni digitali, con particolare riguardo a:

- deposito delle tesi di dottorato in formato digitale
- istituzione del servizio National Bibliography Number (NBN) per le pubblicazioni digitali
- sperimentazione ex art 37 (deposito volontario digitale nativo)

Magazzini digitali per il deposito legale (3)

Qualche dato a fine 2016

- Tesi dottorato
  - 37 università (90 k tesi)
- Riviste accademiche ad accesso aperto:
  - 116 (27 k articoli)
- Editoria commerciale deposito volontario
  - 3k e-book
- NBN generati
  - 22k

depositolegale.it

Magazzini digitali per il deposito legale (4)

2016-2017

Progetto “ARCUS” per l’evoluzione di MD

- Nuova infrastruttura hardware e software
- Sviluppo di nuove procedure e strumenti per il deposito (es. BookDeposit) e l’accesso ai documenti digitali (es. Browser remoto)
- Interfaccia “catalogo” di Magazzini Digitali (sapere sempre cosa c’è in MD anche se non sempre - per questioni di copyright - è possibile l’accesso)

Magazzini digitali per il deposito legale (5)

2017

- Preparazione del progetto “raccolta (harvesting) dei siti web delle istituzioni culturali italiane” finanziato dalla L. 190/2014 (nome interno *Progetto 190)*
- DFP di fondamentale importanza per la preparazione del progetto

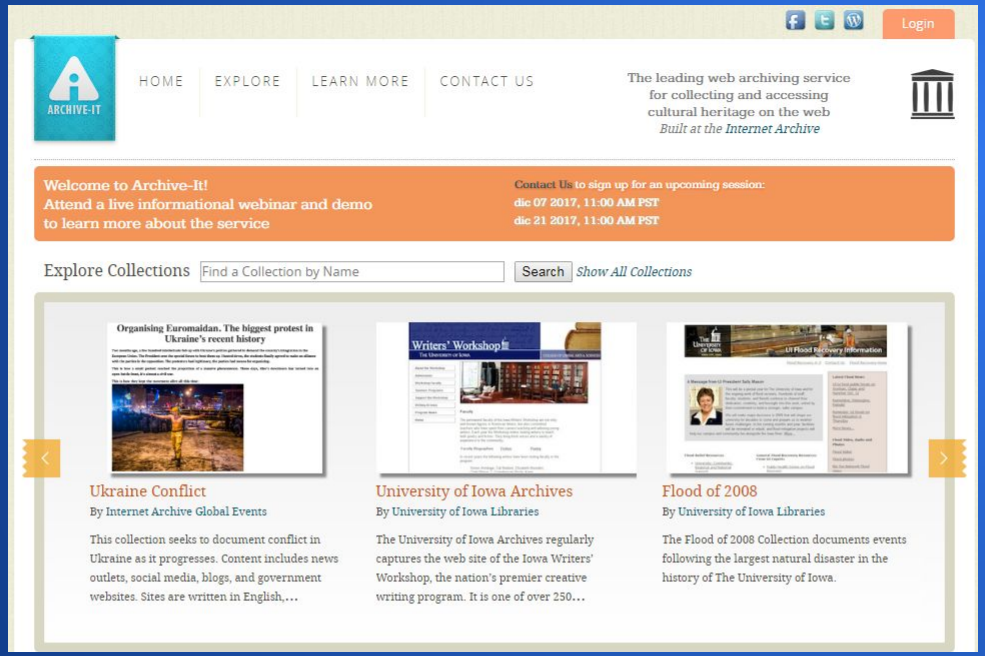
Il progetto 190 (1)

- *NON* harvesting dei siti in senso stretto
  (servizio sostanzialmente già offerto da Internet Archive)

- *MA* harvesting dei *documenti pubblicati nei siti* di interesse culturale (senza perdita del contesto di provenienza)
  - raccolta più “profonda” e selettiva (il *seed* di partenza può anche essere un sottodominio del dominio principale)
  - documenti che si configurano come la naturale prosecuzione delle collezioni delle biblioteche

  (la complementarietà al lavoro di DFP risulta evidente)

Il progetto 190 e la piattaforma Archive-it



Piattaforma *in cloud* di Internet Archive offerta a pagamento:

- raccolta (harvesting)
- possibile di arricchimento con metadati (manuale ...)
- l'accesso anche in modalità full-text
- organizzazione delle raccolta in collezioni (anche *private*)
- disponibilità del formato WARC

Il progetto 190:
aspetti giuridici-amministrativi

In attesa del regolamento tecnico attuativo della L. 252/2006:

- non c'è obbligo di deposito dei documenti diffusi tramite rete informatica
- occorre chiedere il permesso per la raccolta e il ri-utilizzo dei dati

Progetto 190: aspetti tecnici

(alcuni) limiti della raccolta automatica (harvesting):

- presenza di *robots.txt* (che "impediscono" ai motori di ricerca, compreso quello di Archive-it, di archiviare i siti) → occorre l'autorizzazione del gestore del sito
- siti o parti di siti ad accesso limitato → è possibile fornire ad Archive-it le credenziali di accesso ma non è possibile superare la presenza di eventuali *captcha* (o in generale quando l'accesso prevede una interazione con l'utente)
- Il controllo qualità può essere effettuato solo *a campione*

Progetto 190:
Selezione dei siti

Priorità nella selezione dei siti contenenti pubblicazioni ...

- ufficiali dello Stato
- degli organi centrali e periferici del MIBACT
- dei Ministeri
- di altri Enti e Istituzioni statali o a carattere nazionale
- delle Regioni e degli altri enti locali territoriali
- delle Università (*non* Tesi di dottorato e pubblicazioni OA già oggetto di raccolta)
- di altre istituzioni di interesse culturale

Progetto 190:
formati dei documenti

Per partire:

- pdf
- epub
- doc/docx
- html

Progetto 190:
aspetti biblioteconomici -- fase 1

Inserimento di metadati in formato Dublin Core (con l'aggiunta di qualche campo per l'*uso locale*)

- a livello di *seed*: minimali (servono soprattutto a ricostruire il contesto di provenienza dei documenti)
- a livello di singolo documento: minimali ma con riferimento alla compatibilità con SBN es. definizione della "natura bibliografica": monografia, fascicolo o spoglio

Progetto 190:
aspetti biblioteconomici -- fase 2

- Riversamento dei metadati in SBN (polo e indice)
  dovrà essere sviluppato uno strumento per facilitare il controllo delle voci di autorità in SBN (autori, titoli, collane) → esempio del Mix'n'match di Wikidata

- Conservazione dei dati in MD (in formato WARC)
  controllo delle licenze, gestione degli accessi e delle visualizzazioni

Progetto 190:
prospettive future - 1

Nel corso del progetto (tre anni) contiamo di prendere in conto anche l'accesso per soggetto ai documenti:

- occorre realisticamente mettere in primo piano considerazioni di *sostenibilità* ...
- vista la disponibilità del full-text è oggi possibile pensare all'*aiuto* di applicazioni basate sul *machine learning*
- *integrare* i documenti digitali sempre di più con il lavoro della BNI e del Nuovo Soggettario

Progetto 190:
prospettive future - 2

Interpretiamo la nostra presenza a questo evento come un invito alla collaborazione:

- con DFP/AIB (oltre che per l'individuazione dei *seed* anche per gli aspetti legati all'accesso per soggetto e più in generale alla categorizzazione dei documenti)

- con tutti coloro che hanno esperienze in questo dominio

Grazie