Archiviazione di periodici elettronici

L'approccio di Portico, un progetto nato negli Stati Uniti

Amy Kirchhoff
Archiving Service Product Manager
Portico
Eileen Gifford Fenton*
Executive Director of Portico
eileenfenton@portico.org

Esigenze crescenti di preservazione digitale

Secondo la più recente (2006) indagine continuativa di Pew Internet & American Life Project, "il 73% degli intervistati … sono utenti di Internet, in crescita rispetto al 66%" nel gennaio 2005[1] e al 47% nel giugno 2000.[2] Con l'incremento degli utenti è cresciuto anche il contenuto disponibile. Nel 2000, Lyman e Varian hanno stimato che la "surface Web" (le pagine statiche dei siti) ammontava a circa 50 terabyte[3] ed entro il 2003 tale stima era salita a 170 terabyte.[4] I siti web attivi sono cresciuti da circa 8,2 milioni nel maggio 2000 a quasi 41 milioni nel maggio 2006.[5]
L'uso e la creazione di contenuto digitale da parte della comunità accademica ha rispecchiato questa tendenza più vasta, come illustrato da una varietà di punti di rilevamento. Per esempio, nel settembre 2001, la Directory of Electronic Journals, Newsletters and Academic Discussion Lists dell'Association of Research Libraries (ARL, Associazione delle biblioteche di ricerca) ha catalogato 3.887 periodici elettronici e newsletter.[6] Entro maggio 2006, il numero di periodici elettronici e newsletter catalogati era salito a 17.392.[7] Anche l'utilizzo di contenuto elettronico accademico è cresciuto enormemente, come è dimostrato da una varietà di dati. Per esempio, l'uso di JSTOR, un archivio online dei numeri arretrati di periodici accademici, è cresciuto da 3,9 milioni di accessi significativi nell'aprile 2000 a quasi 40 milioni di accessi significativi nell'aprile 2006.[8] Anche la percentuale dei budget delle biblioteche accademiche assegnata a risorse elettroniche è in crescita e per l'anno accademico 2003-2004 le biblioteche ARL hanno speso per le risorse elettroniche in media il 31% delle spese totali per i materiali bibliotecari (mediamente 2.718.015 dollari per biblioteca all'anno).[9] Alla luce di altre tendenze, si può concludere che oggi la percentuale è perfino maggiore e che probabilmente continuerà a crescere.
Nel corso dell'ultimo decennio, i docenti ritengono di essere diventati maggiormente dipendenti dalle risorse elettroniche. Un'indagine del 1995 condotta fra docenti di chimica, fisica, sociologia, psicologia, inglese e storia ha concluso che gli intervistati stavano iniziando a utilizzare risorse di rete, ma nutrivano una "mancanza di fiducia" nei confronti dei periodici elettronici.[10] Un sondaggio del 1999-2001 condotto fra docenti e studenti nei campi degli affari internazionali, scienze ambientali e scienze politiche da parte dell'Electronic Publishing Initiative at Columbia (EPIC) ha rilevato che:

> il 40% [degli intervistati] concorda parzialmente o completamente con il fatto che preferirebbe contentarsi di quello che riesce a trovare online, anche se non è esattamente quello che stava cercando, per evitare di recarsi in biblioteca. Per il 20,5% degli intervistati, l'uso di risorse elettroniche è diventato così diffuso che ammette perfino di cercare raramente informazioni oltre le risorse elettroniche.[11]

Nel 2000, anno in cui JSTOR ha svolto un'indagine fra più di 4.000 docenti di scienze sociali e studi classico-umanistici, più del 60% dei docenti intervistati attribuiva alle basi di dati elettroniche un valore inestimabile.[12] Un sondaggio di verifica condotto nel 2003 fra più di 7.000 docenti di scienze sociali e studi classico-umanistici ha riscontrato che "l'uso di risorse elettroniche, e la dipendenza da esse, è aumentato rispetto al 2000".[13]
Il crescente affidamento sulle risorse elettroniche ha contribuito a una crescente preoccupazione per la fragilità del contenuto elettronico. Il primo motore di ricerca full-text, Web Crawler, ha fatto il suo debutto nel 1994 e ha indicizzato circa 72.000 pagine. Nessuna delle venticinque pagine principali allora elencate esiste oggi.[14] Nell'ottobre 2003, "Science" ha pubblicato i risultati di un'indagine che ha rilevato che tre mesi dopo la pubblicazione il 3,8% dei riferimenti a Internet nei tre periodici medici di maggiore impatto non erano attivi; dopo quindici mesi, tale percentuale era salita al 10% ed è poi passata al 13% ventisette mesi dopo la pubblicazione.[15] In modo analogo, uno studio del 2005 pubblicato in "The Serials Librarian" ha mostrato che tre anni dopo la

pubblicazione, metà dei riferimenti a Internet nei tre principali periodici di comunicazione non si collegava a contenuto attivo.[16]
Le preoccupazioni per la fragilità delle risorse elettroniche, soprattutto dei periodici elettronici accademici, sono aggravate dalla:

> natura del sistema di concessione delle licenze nell'ambito del quale questi periodici sono attualmente distribuiti. Quando le biblioteche accademiche e di ricerca ottengono la licenza di periodici elettronici, non entrano effettivamente in possesso di una copia nello stesso modo in cui avveniva con la versione stampata. Piuttosto, utilizzano contenuto archiviato su sistemi remoti controllati da editori… Per i periodici elettronici, il mondo accademico non ha ancora un equivalente funzionale per la manutenzione a lungo termine e il controllo sui documenti accademici come lo aveva tramite il possesso di una copia dei periodici stampati.[17]

Questa dichiarazione tratta da *Urgent action needed to preserve scholarly electronic journals* riecheggia nella relazione del 2006 della Commissione europea Study on the Economic and Technical Evolution of the Scientific Publication Markets in Europe, che affermava che:

> l'era elettronica ha portato ad un cambiamento paradigmatico nel fornire accesso ai numeri arretrati dei periodici: nell'era della stampa, le biblioteche acquisivano periodici stampati ed erano responsabili della loro preservazione in modo che rimanessero accessibili alla loro comunità di utenti a lungo termine. Nell'era digitale, alle biblioteche e alle loro comunità di utenti è concesso in licenza l'accesso online a periodici elettronici per una durata determinata e limitata.[18]

Sforzi emergenti di preservazione digitale

Così come è cresciuta la consapevolezza della vulnerabilità delle risorse elettroniche accademiche, sono aumentati anche gli sforzi di concettualizzare e sviluppare un'infrastruttura di preservazione. A partire dal 1994-1996, la Task Force on Archiving of Digital Information (Task force per l'archiviazione delle informazioni digitali) ha indagato sugli elementi necessari per garantire "un accesso continuato a tempo indeterminato in futuro ai documenti archiviati in formato elettronico digitale".[19] Nel 1996 la task force ha pubblicato la sua relazione finale *Preserving digital information: report of the Task Force on Archiving of Digital Information*, che sottolinea le sfide tecniche e istituzionali inerenti all'archiviazione di informazioni digitali e i ruoli e le responsabilità delle organizzazioni di archiviazione. Nel gennaio 2002, il Consultative Committee of Space Data Systems (CCSDS, Comitato consultivo per i sistemi dei dati spaziali) ha pubblicato il *Reference model for an Open Archival Information System* (OAIS),[20] che descrive un modello tecnologico e organizzativo per la preservazione a lungo termine di oggetti digitali (nel 2003 questo modello di riferimento è divenuto uno standard ISO, ISO 14721:2003). Fondandosi sul lavoro di OAIS e *Preserving digital information*, il Digital Archive Attributes Working Group (Gruppo di lavoro sugli attributi degli archivi digitali) si è riunito nel 2000 e gli è stato "affidato l'incarico di 'definire le caratteristiche di servizi di archiviazione affidabili per collezioni di ricerca eterogenee'". Nel 2002 il gruppo di lavoro ha pubblicato *Trusted digital repositories: attributes and responsibilities*.[21] Questo rapporto è stato ampliato da una task force di RLG e del National Archives and Records Administration (NARA, Amministrazione degli archivi e dei documenti nazionali) per sviluppare uno strumento di verifica, *An audit checklist for the certification of trusted digital libraries*, per certificare gli archivi digitali.[22] Al momento della redazione del presente contributo, il Center for Research Libraries (CRL, Centro per le biblioteche di ricerca) sta utilizzando questo strumento per creare dei procedimenti di certificazione per gli archivi digitali.[23]
Nel corso dell'ultimo decennio, le biblioteche nazionali e altri gruppi hanno stabilito dei programmi sia per implementare archivi digitali, sia per stimolare lo sviluppo di pratiche di archiviazione per archivi digitali, fra cui:
– l'iniziativa Preserving Access to Digital Information (PADI) della Biblioteca nazionale dell'Australia;
– l'e-Depot e la ricerca di preservazione digitale della Koninklijke Bibliotheek (KB), la Biblioteca nazionale dei Paesi Bassi;
– la Digital Preservation Coalition (DPC, Coalizione di preservazione digitale) nel Regno Unito;
– la National Digital Information Infrastructure and Preservation Program (NDIIPP) della Biblioteca del Congresso americano.

La storia di Portico in breve

Per rispondere alla crescente esigenza di soluzioni incentrate sulla preservazione a lungo termine dei periodici accademici, Portico è stata lanciata nel 2005. La missione di Portico è preservare la letteratura accademica pubblicata in formato elettronico e garantire che questi materiali rimangano accessibili in futuro per accademici, ricercatori e studenti. Portico è stata avviata come Electronic-Archiving Initiative (Iniziativa di archiviazione elettronica), un progetto lanciato da JSTOR nel 2002 con una sovvenzione da parte della Andrew W. Mellon Foundation. L'iniziativa si fondava sull'E-journal Archiving

Program (Programma di archiviazione di periodici elettronici) del 1999 della Mellon Foundation, che finanziava diversi programmi per indagare i requisiti di un'infrastruttura tecnologica e le opzioni di modelli economici per sostenere un archivio di periodici elettronici.[24] L'obiettivo iniziale dell'iniziativa era quello di progettare e generare un prototipo di sistemi di archiviazione e di gestione di contenuto, creare modelli di potenziali servizi di archiviazione, collaudare questi possibili modelli di servizi con le biblioteche e gli editori e preparare una bozza di modello imprenditoriale per supportare un impegno di archiviazione a lungo termine. Per oltre due anni, il personale del progetto ha condotto discussioni approfondite con editori e biblioteche per sviluppare un modello imprenditoriale sostenibile e un approccio tecnologico che potesse equilibrare le esigenze di tutte le comunità interessate.

Durante il 2004, il progetto è stato trasferito a Ithaka, un'organizzazione senza scopo di lucro, con l'obiettivo di implementare nuove iniziative e di renderle sostenibili nell'ambito della sua missione più vasta di accelerare gli usi produttivi delle tecnologie informatiche a beneficio dell'istruzione superiore nel mondo. Una volta stabilitasi presso Ithaka, Portico ha avviato una serie di conversazioni con un'ampia rete informale di bibliotecari appartenenti a più di cinquanta istituzioni accademiche di ogni tipo e dimensione. Portico ha inoltre coinvolto formalmente dieci editori che hanno acconsentito a partecipare alla fase pilota del progetto, fra cui piccole società accademiche, una casa editrice universitaria e grossi editori commerciali.[25]

Il nostro lavoro iniziale si fondava su alcune ipotesi chiave relative a caratteristiche critiche per intraprendere l'archiviazione a lungo termine. Tali ipotesi provenivano da diverse fonti, fra cui l'E-journal Archiving Program, l'esperienza operativa di JSTOR in qualità di archivio di terzi dal 1995 e la relazione RLG/OCLC del maggio 2002 *Trusted digital repositories: attributes and responsibilities*.[26] Avvalendoci di questa base collettiva di esperienze, abbiamo presupposto che un archivio affidabile di terzi a lungo termine avrebbe avuto necessità di almeno cinque elementi fondamentali:

1) una missione istituzionale che abbia come valore fondamentale la preservazione;
2) un modello economico capace di sostenere il processo di archiviazione;
3) un'infrastruttura tecnologica solida e in evoluzione, sufficiente per soddisfare le complessità delle risorse elettroniche;
4) rapporti di collaborazione con le biblioteche responsabili della preservazione delle loro collezioni;
5) rapporti di collaborazione con gli editori, creatori del contenuto che deve essere preservato.

Nel corso dei primi mesi del progetto di Portico abbiamo messo alla prova queste supposizioni, collaborando con dieci editori che facevano parte della nostra fase pilota e con varie biblioteche. Gli editori hanno fornito un'ampia gamma di dati campione di periodici elettronici che ci hanno permesso di valutare le sfide tecnologiche che il nostro archivio di periodici elettronici dovrà affrontare; inoltre hanno condiviso il loro punto di vista sulla sfida di archiviazione dei periodici elettronici e le preoccupazioni e le esigenze che dovevano affrontare come editori di contenuto. Le nostre conversazioni con gli editori hanno rispecchiato quelle che abbiamo condotto nella comunità delle biblioteche. I bibliotecari hanno condiviso le loro preoccupazioni in materia di preservazione, le aspettative nei confronti di un archivio di periodici elettronici e le loro opinioni in merito a varie possibilità di modelli economici.

L'impegno di questa comunità e il nostro lavoro per costruire un prototipo di archivio e definire un modello imprenditoriale sostenibile per un servizio di archiviazione di terzi hanno confermato le nostre ipotesi iniziali, ma hanno offerto anche diverse lezioni importanti.

In primo luogo, la preservazione di periodici elettronici presenta molte sfide tecnologiche significative. La qualità dei dati e il formato dei periodici elettronici variano non solo all'interno del settore, ma spesso anche fra le pubblicazioni di uno stesso editore e talvolta fra un numero e l'altro di una pubblicazione. Qualsiasi infrastruttura di preservazione a lungo termine, fra cui software, hardware e personale munito delle competenze appropriate, deve rispondere alle sfide di questa diversità.

In secondo luogo, la sola portata del materiale digitale da preservare impone che le biblioteche e gli editori costruiscano meccanismi molteplici per uno sforzo di cooperazione. La collaborazione della comunità editoriale accademica è essenziale perché, in molti casi, le biblioteche affittano periodici elettronici piuttosto che possederli. Garantire questa cooperazione è complicato dal fatto che la preservazione non è un'attività primaria per la maggioranza degli editori.

In terzo luogo, l'accesso alla letteratura archiviata è una questione chiave sia per gli editori che per le biblioteche, ma essi vedono la questione da prospettive molto diverse. Gli editori sono desiderosi di garantire che l'accesso alla letteratura archiviata non riduca il valore delle loro offerte attuali di prodotti; l'obiettivo delle biblioteche è quello di assicurare che l'accesso alla letteratura fondamentale per i loro campus sia affidabile e tempestivo.

Infine, tutte le parti che abbiamo interpellato hanno riconosciuto l'importanza di sviluppare un modello economico robusto, in grado di sostenere un archivio a lungo termine.
Partendo da queste supposizioni e risultati, il servizio di archiviazione elettronica di Portico è stato lanciato nel 2005.

Il modello di Portico

Il servizio di archiviazione di Portico è aperto all'elenco completo di periodici di un editore accademico, compresi quei titoli che possono essere pubblicati solo in formato elettronico, quelli pubblicati congiuntamente in formato elettronico e cartaceo e quelli digitalizzati da una pubblicazione stampata originale. L'approccio di archiviazione di Portico per i periodici elettronici è quello di preservazione gestita incentrata sulla migrazione iterativa a lungo termine dei file sorgente dei periodici elettronici degli editori (come discusso più dettagliatamente di seguito).

Supporto finanziario

Un supporto finanziario diversificato è fondamentale per l'impegno di preservazione a lungo termine di Portico. In generale occorrono due tipi di supporto: fondi per lo sviluppo iniziale dell'infrastruttura tecnologica e delle attività iniziali, e fondi continui che possono supportare le attività dell'archivio nel corso della vita dei materiali preservati. Portico ha ottenuto sovvenzioni da JSTOR, dalla Biblioteca del Congresso e dalla Andrew W. Mellon Foundation per coprire i costi dello sviluppo iniziale, con un supporto supplementare da parte di Ithaka. Man mano che l'archivio di Portico si espanderà, l'organizzazione coprirà i suoi costi operativi tramite fonti diversificate per evitare di dipendere da una fonte di reddito unica. I principali beneficiari dell'archivio, gli editori e le istituzioni accademiche, forniranno le fonti di finanziamento primarie. Anche le fondazioni di beneficenza e gli enti governativi saranno tenuti a erogare sovvenzionamenti periodici.
Agli editori è chiesto di effettuare un contributo annuale a favore dell'archivio per supportare i costi continui di ricezione, normalizzazione, archiviazione e migrazione dei file sorgente degli articoli. I contributi si basano sui ricavi totali dei periodici da parte degli editori (compresi l'abbonamento, la pubblicità e la concessione di licenze) e variano fra i 250 e i 75.000 dollari all'anno. Anche le biblioteche affiliate sono tenute a emettere un pagamento annuale per supportare il lavoro in corso dell'archivio. I pagamenti annuali di supporto dell'archivio da parte delle biblioteche sono suddivisi in livelli compresi fra i 1.500 e i 24.000 dollari all'anno, e variano in base alle spese totali della biblioteca per i materiali bibliotecari, riflettendo il valore di Portico di preservazione di quella parte crescente delle collezioni bibliotecarie che è il contenuto digitale. I dettagli relativi ai livelli contributivi attuali delle biblioteche e degli editori sono disponibili sul sito web di Portico, all'indirizzo <www.portico.org>.

Accesso all'archivio

L'accesso in tutto il campus al contenuto archiviato sarà concesso alle biblioteche che sostengono Portico quando vengono soddisfatte condizioni specifiche che fanno sì che certi titoli non siano più disponibili dall'editore o da una qualsiasi altra fonte. Queste condizioni specifiche sono spesso denominate eventi d'innesco, e comprendono:
– la cessazione dell'attività da parte di un editore;
– la cessazione della pubblicazione di un titolo da parte di un editore;
– la cessazione dell'offerta di numeri arretrati da parte di un editore;
– il caso di guasto catastrofico e sostenuto della piattaforma di distribuzione di un editore.
Portico fornisce anche un mezzo sicuro per garantire accesso continuo, se gli editori affiliati scelgono di designare Portico come un fornitore di accesso post-annullamento.
Oltre all'innesco dell'accesso da parte degli eventi suddetti, ad un massimo di quattro bibliotecari presso ogni istituzione affiliata sarà concesso l'accesso controllato da password all'archivio solo a scopi di controllo e verifica. Anche agli editori è concesso l'accesso controllato da password al loro contenuto nell'archivio.

Approccio di archiviazione

Oltre a sviluppare un modello di accesso ed economico, Portico ha formulato una serie di principi guida che determinano il nostro approccio alla preservazione di contenuto elettronico. Questi principi sono emersi dalla comprensione attuale dei problemi di preservazione digitale da parte della comunità accademica e dalle nostre discussioni iterative con le biblioteche e la comunità degli editori.
a) *L'integrità dei documenti accademici deve essere preservata*
L'archivio di Portico accetta il contenuto nel formato in cui è stato pubblicato originariamente. Le versioni originali e migrate di file sorgente componenti sono preservate e, una volta depositato, il contenuto archiviato sarà preservato nell'archivio in modo permanente.
b) *L'obiettivo di Portico è l'archiviazione di contenuto accademico*
Il ruolo di Portico consiste nel preservare il contenuto intellettuale delle risorse accademiche, a cominciare dai periodici elettronici

accademici. Non è nostro obiettivo preservare i sistemi aziendali degli editori, i dati di produzione di contenuto correlati o le piattaforme di distribuzione.

c) *I file sorgente possono acquisire correttamente il contenuto intellettuale dei periodici elettronici accademici*

Portico preserva i periodici accademici mediante l'acquisizione e la migrazione dei file sorgente che comprendono i periodici elettronici. Questi file includono la copia principale dell'articolo e i componenti utilizzati nelle riproduzioni web e stampate del periodico. I file sorgente possono includere:

– file PDF utilizzati nella versione stampata e online;
– file a testo integrale SGML o XML;
– file di intestazione SGML o XML di metadati bibliografici;
– molteplici immagini in risoluzioni diverse per ogni immagine nell'articolo;
– immagini per le equazioni o tabelle nell'articolo e altri dati supplementari, come ad esempio video, audio e insiemi di dati.

Siamo consapevoli che alcune informazioni potrebbero non essere acquisite nei file sorgente. Fra queste, vi sono di solito materiali che esistono solo nei sistemi aziendali dell'editore o che sono generati tramite la infrastruttura di distribuzione dell'editore, come per esempio copertine e altre parti introduttive, pubblicità stampate o online e sommari stampati o online. Anche i materiali supplementari che esistono al di fuori dei limiti del periodico accademico e che non sono mantenuti dall'editore potrebbero non essere disponibili per la preservazione dei file sorgente.

d) *La preservazione del contenuto dei periodici elettronici è realizzabile tramite la migrazione*

Portico ha optato per una strategia di migrazione tramite la transizione dei file sorgente componenti da un formato di file a un altro man mano che la tecnologia cambia e i formati di file diventano obsoleti. Portico potenzia e supporta la politica di migrazione con preservazione di byte, archiviando i file sorgente originali insieme a tutte le versioni migrate.

Il nostro interesse e la nostra attenzione di preservazione sono rivolti alla comprensione dei componenti dei file sorgente che costituiscono gli articoli e non alla tecnologia per la distribuzione del contenuto. Portico si concentra sulla gestione del formato dei file. Monitoriamo la tecnologia informatica e le comunità di archiviazione per capire l'accettazione da parte della comunità dei formati specifici di file all'interno dell'archivio di Portico. In base alle necessità, aggiorneremo i file in archivio con nuovi formati al fine di mantenere i contenuti dell'archivio al passo con i progressi tecnologici.

La prima azione di preservazione di Portico consiste nel normalizzare i file a testo integrale o di intestazione SGML o XML degli editori in conformità al Journal Archiving and Interchange DTD[27] creato dal National Center for Biotechnology Information (NCBI, Centro nazionale per le informazioni di biotecnologia) della National Library of Medicine (NLM, Biblioteca nazionale di medicina). Questo processo di normalizzazione limita il numero di formati che dobbiamo monitorare e ci consente inoltre di sviluppare una riproduzione HTML uniforme di tutto il contenuto in archivio. I file SGML o XML originali e i file XML basati sul DTD della NLM sono tutti preservati in archivio. In conformità al modello OAIS, l'archivio include anche tutti i DTD, gli schemi e la documentazione correlata necessari per capire il contenuto archiviato.[28]

Il fare affidamento su standard accettati e la partecipazione allo sviluppo di progetti comunitari migliora l'affidabilità di archiviazione. Di seguito sono riportati alcuni degli standard e dei progetti che hanno influenzato la pratica di archiviazione di Portico:

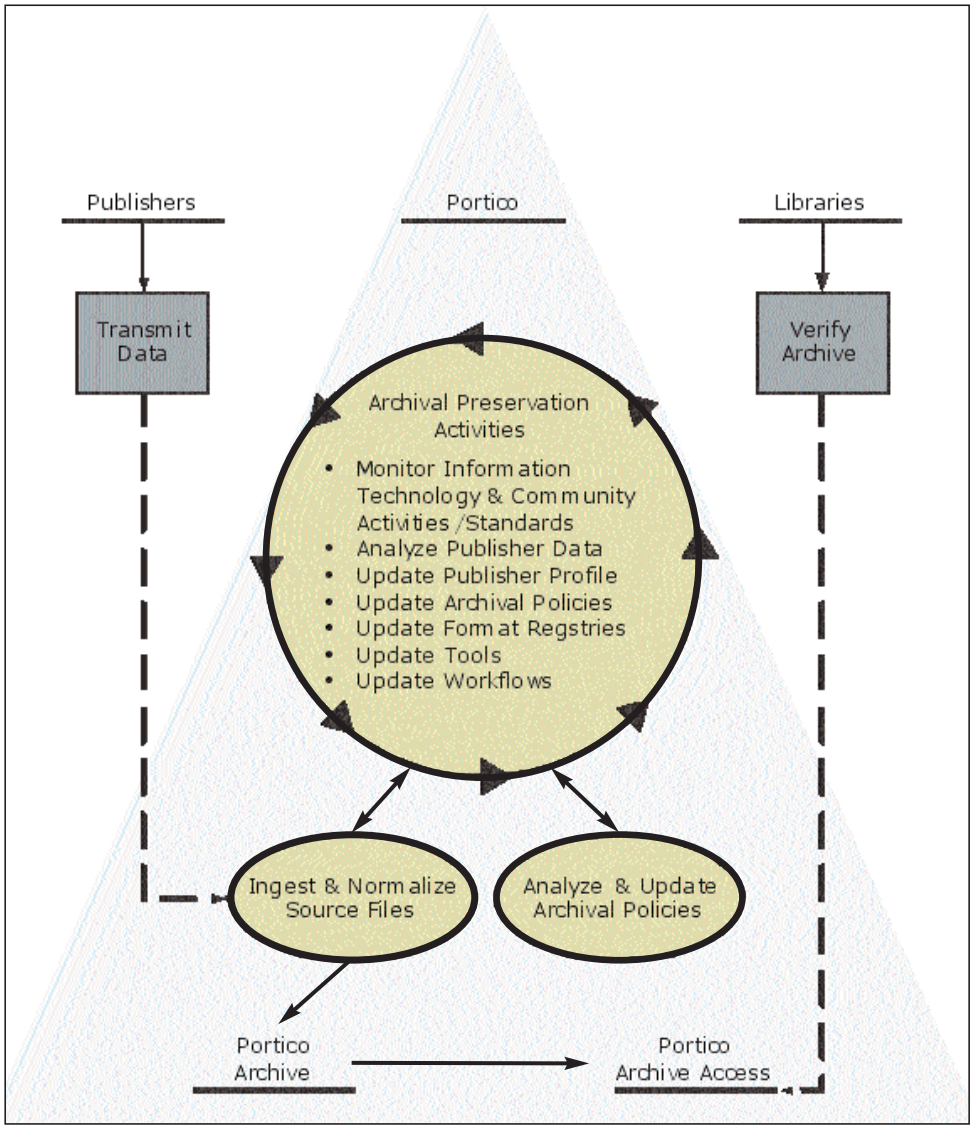
– Digital Item Declaration Language (DIDL) di MPEG-21: DIDL è il linguaggio di formattazione per "l'impacchettamento" di dati di MPEG-21, una struttura multimediale in via di sviluppo.[29] Il modello del contenuto di Portico si basa sui concetti di "impacchettamento" DIDL.
– Metadata Encoding and Transmission Standard (METS): METS è uno schema XML per la codifica di metadati descrittivi, amministrativi e strutturali relativi a oggetti all'interno di una biblioteca digitale.[30] Portico utilizza una versione modificata di METS per rappresentare gli articoli nell'archivio e per tenere traccia delle misure di preservazione adottate sui file sorgente di ogni articolo.
– Journal Archiving and Interchange DTD creato dalla National Library of Medicine (DTD della NLM): Il DTD della NLM fornisce un formato comune in cui gli editori e gli archivi possono scambiare e preservare contenuto di periodici. L'intento della suite DTD è quello di "preservare il contenuto intellettuale di periodici a prescindere dal formato in cui" sono stati distribuiti in origine.[31] Tutti i file a testo integrale o di intestazione degli editori forniti a Portico sono trasformati in un file XML basato sul DTD della NLM.
– Reference model for an Open Archival Information System (OAIS): OAIS è una raccomandazione relativa ai requisiti su cui si dovrebbe fondare un archivio che fornisce la preservazione permanente di informazioni digitali.[32] L'archivio di Portico è stato progettato per conformarsi a OAIS.
– Global Digital Format Registry (GDFR): GDFR è un progetto di creazione di un registro distribuito sui formati per archiviare, scoprire

e distribuire informazioni relative ai formati digitali.[33] Il registro sui formati di Portico si basa sul lavoro GDFR e sarà sostituito da GDFR al momento opportuno.
– JSTOR/Harvard Object Validation Environment (JHOVE): JHOVE è un progetto di collaborazione per sviluppare una struttura espandibile per convalidare i formati di file.[34] JHOVE è utilizzato per convalidare il formato di file nell'archivio e stiamo sviluppando moduli aggiuntivi in base alle esigenze.
– Preservation Metadata: Implementation Strategies (PREMIS): il gruppo di lavoro PREMIS ha sviluppato un dizionario di dati e uno schema XML per la preservazione dei metadati.[35] Il personale di Portico ha partecipato al gruppo di lavoro di PREMIS e ha influenzato i metadati di preservazione acquisiti nel procedimento operativo di Portico.

Procedimenti e preservazione in corso di realizzazione

Portico riceve i file sorgente dagli editori e tramite un procedimento automatizzato assembla questi file in un pacchetto di archiviazione indipendente dal formato originale dell'editore. Tale pacchetto è successivamente incorporato nell'archivio. Il procedimento di incorporazione è guidato da una serie di regole specifiche all'editore e da strumenti costruiti tramite la nostra analisi dei file sorgente dell'editore e discussioni con l'editore. La maggior parte di tale analisi avviene prima che iniziamo a elaborare i dati; tuttavia, essa è rivista in base alle esigenze quando riceviamo un contenuto che non combacia con l'insieme di regole corrente. Il procedimento di incorporazione verifica la struttura e le relazioni dei file sorgente, identifica i formati dei file sorgente e convalida la conformità dei file alle specifiche del loro formato. I file di intestazione o a testo integrale SGML o XML sono normalizzati secondo il DTD della NLM nel corso del procedimento di incorporazione. In seguito, il procedimento crea i metadati di preservazione richiesti da Portico per tenere traccia delle azioni perpetrate sui dati. L'intero procedimento di incorporazione è gestito dal personale operativo, che esegue inoltre controlli di qualità su dati campione. L'ultima fase del procedimento di incorporazione consiste nel depositare l'articolo nell'archivio di Portico.
Tutti gli articoli nell'archivio e tutti i loro file associati sono considerati un'unità di archiviazione e sono rappresentati da un file XML leggibile dall'uomo in una versione modificata dello schema METS, conosciuta come Portico METS o PMETS.[36] Questo file PMETS fa riferimento a tutti i file sorgente componenti necessari per ri-

Fig. 1 – Procedimento di preservazione di Portico



produrre l'articolo. L'intero archivio può essere ricreato dai file sorgente componenti e dai file PMETS senza fare riferimento ad alcuna struttura di dati proprietaria, ad eccezione dei file PMETS XML leggibili dall'uomo. Per motivi di sicurezza, Portico ha sviluppato un calendario trimestrale per replicare l'archivio su molteplici supporti, all'interno di vari sistemi di gestione dell'archivio e in aree geografiche distinte.

L'elaborazione operativa dei file sorgente assicura la nostra comprensione dello stato di file specifici (per esempio, quali file sono ben formati e quali sono corrotti) e ne cattura le relazioni in modo uniforme fra tutti gli editori e i titoli. Effettueremo dei controlli sull'archivio con regolarità e monitoreremo l'accettazione di formati di file specifici nella comunità. Le politiche di preservazione e dei formati di Portico saranno riviste regolarmente e aggiornate in base alle esigenze e il contenuto sarà migrato se appropriato. Come mostrato nella figura 1, l'incorporazione e la normalizzazione di file sorgente e l'analisi e l'aggiornamento delle politiche di archiviazione sono compiti continui su cui si fondano le nostre attività di preservazione.

Organizzazioni e progetti citati

Andrew W. Mellon Foundation, <http://www.mellon.org>.
Association of Research Libraries (ARL), <http://www.arl.org/>.
Biblioteca del Congresso, <http://www.loc.gov/>.
Center for Research Libraries (CLR), <http://www.crl.edu/>.
The Consultative Committee for Space Data Systems (CCSDS), <http://public.ccsds.org/>.
Creative Archiving at Michigan and Leeds Emulating the Old On the New (CAMiLEON), <http://www.si.umich.edu/CAMILEON>.
Digital Library Federation (DLF), <http://www.diglib.org/>.
Digital Preservation Coalition (DPC), <http://www.dpconline.org/>.
Electronic Publishing Initiative at Columbia (EPIC), <http://www.epic.columbia.edu/>.
Ithaka, <http://www.ithaka.org>.
JSTOR, <http://www.jstor.org>.
Koninlijke Bibliotheek (Biblioteca nazionale dei Paesi Bassi, conosciuta anche come "KB"), <http://www.kb.nl/index-en.html>.
The National Archives (NARA), <http://www.archives.gov/>.
National Center for Biotechnology Information (NCBI), <http://www.ncbi.nih.gov/>.
National Digital Information Infrastructure and Preservation Program (NDIIPP, presso la Biblioteca del Congresso), <http://www.digitalpreservation.gov/>.
National Library of Medicine (NLM), <http://www.nlm.nih.gov/>.
Online Computer Library Center (OCLC), <http://www.oclc.org>.
Organizzazione internazionale di standardizzazione (ISO), <http://www.iso.org/>.
Pew Internet & American Life Project, <http://www.pewinternet.org/>.
Portico, <http://www.portico.org>.
Preserving Access to Digital Information (PADI, presso la Biblioteca nazionale dell'Australia), <http://www.nla.gov.au/padi/>.
The Research Libraries Group (RLG), <http://www.rlg.org>.

Bibliografia

Archiving electronic journals [website], Digital Library Federation, 2003, <http://www.diglib.org/preserve/ejp.htm> (accessed May 18, 2006).
ARL directory of electronic journals, newsletters and academic discussion lists, 7th Edition, (ARL, 2001), <https://db.arl.org/edir/> (accessed May 22, 2006).
Audit checklist for the certification of trusted digital repositories: draft for public comment, Mountain View, CA:RLG and NARA, 2005, <http://www.rlg.org/en/page.php?Page_ID=20769> (accessed May 22, 2006).
Budd, John M. – Connaway Lynn Silipigni, *University faculty and networked information: results of a survey, "Journal of the American Society for Information Science"*, 48 (1997), 9, p. 843-852, <http://www3.interscience. wiley.com/cgi-bin/abstract/39759/ABSTRACT> (accessed June 9, 2006).
Bugeja, Michael – Daniela V. Dimitrova, *The half-life phenomenon: eroding citations in journals*, "The Serials Librarian", 49 (2005), 3, p. 115-123, <http://www.haworthpress.com/store/EText/View_EText.asp?sa=3&s=J123&v=49&i=3&fn=J123v49n03%5F10> (accessed May 23, 2006).
CRL auditing and certification of digital archives [website], Center for Research Libraries, 2005, <http://www.crl.edu/content.asp?l1=13&l2=58&l3=142> (accessed June 19, 2006).
Dellavalle, Robert P. – Eric J. Hester – Lauren F. Heilig — Amanda L. Drake – Jeff W. Kuntzman – Marla Graber – Lisa M. Schilling, *Going, going, gone: lost internet references*, "Science", 302 (2003), 5646, p. 787-788, <http://www.sciencemag.org/cgi/content/full/302/5646/787> (accessed May 22, 2006).
Dewatripont, Mathias – Victor Ginsburgh – Patrick Legros – Alexis Walckiers – Jean-Pierre Devroey – Marianne Dujardin – Françoise Vandooren – Pierre Dubois – Jérôme Foncel – Marc Ivaldi – Marie-Dominique Heusse, *Study on the economic and technical evolution of the scientific publication markets in Europe*, European Commission, 2006, <http://europa.eu.int/comm/research/science-society/pdf/scientific-publication-study_en.pdf> (accessed May 22, 2006).
EPIC faculty survey [Power Point presentation], Electronic Publishing Initiative at Columbia, 2003, <http://www.epic.columbia.edu/eval/FacSurv.903.ppt> (accessed May 19, 2006).
Garrett, John – Don Waters, *Preserving digital information: report of the task force on archiving of digital information*, Task Force on Archiving of Digital Information, 1996, <http://www.rlg.org/ArchTF/> (accessed May 2, 2006).
Global digital format registry [website], Harvard University Library, 2005, <http://hul.harvard.edu/gdfr/> (accessed May 2, 2006).
Guthrie, Kevin, *What do faculty think of electronic resources*, presented at the JSTOR ALA Annual Conference Participants' Meeting, June 17, 2001,

<http://www.jstor.org/about/faculty.survey.ppt> (accessed May 23, 2006).
GUTHRIE, KEVIN – ROGER SCHONFELD, *What do faculty think of electronic resources? Findings from the 2003 Academic Research Resources study*, presented at the CNI Task Force Meeting, Alexandria, Virginia, April 16, 2004, <http://www.cni.org/tfms/2004a.spring/presentations/CNI_Guthrie_What.ppt> (accessed May 23, 2006).
JHOVE-JSTOR/Harvard object validation environment [website], Harvard University Library, 2006, <http://hul.harvard.edu/jhove/> (accessed May 2, 2006).
Journal archiving and interchange DTD [website], National Center for Biotechnology Information of the National Library of Medicine, 2004, <http://dtd.nlm.nih.gov/> (accessed May 18, 2006).
JSTOR usage statistics, JSTOR, <http://stats.jstor.org/> (accessed May 22, 2006).
LYMAN, PETER – HAL VARIAN, *How much information 2000*, University of California, 2000, <http://www2.sims.berkeley.edu/research/projects/how-much-info/index.html> (accessed May 22, 2006).
ID., *How much information 2003*, University of California, 2003, <http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/> (accessed May 22, 2006).
MADDEN, MARY, *Internet penetration and impact*, Pew Internet & American Life Project, 2006, <http://www.pewinternet.org/PPF/r/182/report_display.asp> (accessed June 21, 2006).
May 2006 Web server survey, Netcraft, 2006, <http://news.netcraft.com/archives/2006/06/04/june_2006_web_server_survey.html> (accessed May 22, 2006).
METS: Metadata Encoding & Transmission Standard [website], Library of Congress, <http://www.loc.gov/standards/mets/> (accessed May 18 2006).
MPEG-21 part 2: Digital Item Declaration Language (DIDL), in *Technology reports* [website], cover pages, 2004, <http://xml.coverpages.org/mpeg21-didl.html> (accessed June 9, 2006).
New Jour Archive [email list archive], University of California at San Diego Libraries, 2006, <http://gort.ucsd.edu/newjour/> (accessed June 21, 2006).
PREMIS (Preservation Metadata: Implementation Strategies) working group [website], OCLC, 2005, <http://www.oclc.org/research/projects/pmwg/> (accessed June 19, 2006).
RAINIE, LEE – DAN PACKEL – SUSANNAH FOX – JOHN HORRIGAN – AMANDA LENHART – TOM SPOONER – OLIVER LEWIS – CORNELIA CARTER, *More online, doing more*, Pew Internet & American Life Project, 2001, <http://www.pewinternet.org/PPF/r/30/report_display.asp> (accessed June 21, 2006).
Reference model for an Open Archival Information System (OAIS), National Aeronautics and Space Administration – Consultative Committee for Space Data Systems, 2002, <http://public. ccsds.org/publications/archive/650x0b1.pdf> (accessed May 2, 2006).
RUNNING, JORDAN, *The top 25 Web sites of 1994*, in *downloadsquad* [weblog], 2006, <http://www.downloadsquad.com/2006/04/14/the-top-25-web-sites-of-1994/> (accessed May 22, 2006).
Trusted digital repositories: attributes and responsibilities, Mountain View, CA:RLG-OCLC, 2002, <http://www.rlg.org/longterm/repositories.pdf> (accessed May 19, 2006).
Urgent action needed to preserve scholarly electronic journals, Digital Library Federation, 2005, <http://www.diglib.org/pubs/waters051015.htm> (accessed May 1, 2006).
YOUNG, MARK – MARTHA KYRILLIDOU, *ARL statistics 2003-2004*, Washington, Association of Research Libraries, 2005.

Note

*Alcuni elementi di questo contributo sono apparsi per la prima volta in articoli di Eileen Fenton, pubblicati in "Ariadne Web Magazine", (April 2006), 47, <http://www.ariadne.ac.uk/issue47/>, e "Serials Review", 32, (2006), 2, <doi:10.1016/j.serrev.2006.03.004>.

[1] MADDEN, *Internet penetration and impact* (2006), p. 3, <http://www.pewinternet.org/PPF/r/182/report_display.asp>.

[2] RAINIE et al., *More online, doing more* (2001), p. 2, <http://www.pewinternet.org/PPF/r/30/report_display.asp>.

[3] LYMAN – VARIAN, *How much information 2000* (2000), <http://www2.sims.berkeley.edu/research/projects/how-much-info/index.html>.

[4] ID., *How much information 2003* (2003), <http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/>.

[5] *May 2006 Web server survey* (2006), <http://news.netcraft.com/archives/2006/06/04/june_2006_web_server_survey.html>.

[6] *ARL directory of electronic journals, newsletters and academic discussion lists*, 7th edition, ARL, 2001, <https://db.arl.org/edir/>.

[7] New Jour Archive [email list archive] (University of California at San Diego Libraries, 2006), <http://gort.ucsd.edu/newjour/>.

[8] *JSTOR usage statistics*, <http://stats.jstor.org/>. Un "accesso significativo" si verifica quando un utente accede a un oggetto intellettuale, come per esempio la visualizzazione di un volume o di un indice di numeri, la visualizzazione del sommario di un numero, la visualizzazione di una pagina, l'esecuzione di una ricerca oppure la stampa di un articolo.

[9] YOUNG – KYRILLIDOU, *ARL statistics 2003-2004* (Washington, Association of Research Libraries, 2005).

[10] BUDD – CONNAWAY, *University faculty and networked information: results of a survey*, "Journal of the American Society for Information Science", 48 (1997), 9, p. 850, <http://www3.interscience.wiley.com/cgi-bin/abstract/39759/ABSTRACT>.

[11] *EPIC faculty survey* [Power Point presentation] (Electronic Publishing Initiative at Columbia, 2003), p. 17, <http://www.epic.columbia.edu/eval/FacSurv.903.ppt>.

[12] GUTHRIE, *What do faculty think of electronic resources* (JSTOR ALA Annual Conference Participants' Meeting June 17, 2001), p. 17, <http://www.jstor.org/about/faculty.survey.ppt>.

[13] GUTHRIE – SCHONFELD, *What do faculty think of electronic resources? Findings from the 2003 Academic Research resources study* (CNI Task Force Meeting April 16, 2004), p. 16, <http://www.cni.org/tfms/2004a.spring/presentations/CNI_Guthrie_What.ppt>.

[14] RUNNING, *The top 25 Web sites of 1994* [weblog] (2006), <http://www.downloadsquad.com/2006/04/14/the-top-25-web-sites-of-1994/>.

[15] Dellavalle et al., *Going, going, gone: lost Internet references*, "Science", 302 (2003), 5646, <http://www.sciencemag.org/cgi/content/full/302/5646/787>.
[16] Bugeja – Dimitrova, *The half-life phenomenon: eroding citations in journals* ("The Serials Librarian"), 49 (2005), 3, p. 117, <http://www.haworthpress.com/store/E-Text/View_EText.asp?sa=3&s=J123&v=49&i=3&fn=J123v49n03%5F10>.
[17] *Urgent action needed to preserve scholarly electronic journals* (2005), <http://www.diglib.org/pubs/waters051015.htm>.
[18] Dewatripont et al., *Study on the economic and technical evolution of the scientific publication markets in Europe* (2006), p. 75, <http://europa.eu.int/comm/research/science-society/pdf/scientific-publication-study_en.pdf>.
[19] Garrett – Waters, *Preserving digital information: report of the task force on archiving of digital information* (1996), p. III, <http://www.rlg.org/ ArchTF/>.
[20] *Reference model for an Open Archival Information System* (OAIS) (2002), <http://public.ccsds.org/publications/archive/650x0b1.pdf>.
[21] *Trusted digital repositories: attributes and responsibilities* (2002), <http://www.rlg.org/longterm/repositories.pdf>.
[22] *Audit checklist for the certification of trusted digital repositories: draft for public comment* (2005), <http://www.rlg.org/en/page.php?Page_ID=20769>.
[23] *CRL auditing and certification of digital archives* [website] (Center for Research Libraries, 2005), <http://www.crl.edu/content.asp?l1=13&l2=58&l3=142>.
[24] *Archiving electronic journals* [website], (Digital Library Federation, 2003), <http://www.diglib.org/preserve/ejp.htm>.
[25] Fra gli editori partecipanti alla fase pilota di Portico si annoverano: American Economic Association, American Mathematical Society, American Political Science Association, Association of Computing Machinery, Blackwell, Ecological Society of America, National Academy of Sciences (PNAS), Royal Society, University of Chicago Press e John Wiley & Sons.
[26] *Trusted digital repositories: attributes and responsibilities*, cit.
[27] *Journal archiving and interchange DTD* [website], National Center for Biotechnology Information of the National Library of Medicine, 2004, <http://dtd.nlm.nih.gov/>.
[28] *Reference model for an Open Archival Information System* (OAIS), cit.
[29] *MPEG-21 part 2: Digital Item Declaration Language (DIDL)* [website] (cover pages, 2004), <http://xml.coverpages.org/mpeg21-didl.html>.
[30] *METS: Metadata Encoding & Transmission Standard* [website] (Library of Congress), <http://www.loc.gov/standards/mets/>.
[31] *Journal archiving and interchange DTD*, cit.
[32] *Reference model for an Open Archival Information System* (OAIS), cit.
[33] *Global digital format registry* [website] (Harvard University Library, 2005), <http://hul.harvard.edu/gdfr/>.
[34] *JHOVE-JSTOR/Harvard object validation environment* [website] (Harvard University Library, 2006), <http://hul.harvard.edu/jhove/>. Portico, al tempo in cui era l'Electronic-Archiving Initiative (Iniziativa di archiviazione elettronica) di JSTOR, ha collaborato con Harvard per creare questo strumento con codice aperto.
[35] *PREMIS (Preservation Metadata: Implementation Strategies) working group* [website] (OCLC, 2005), <http://www.oclc.org/research/projects/pmwg/>.
[36] *METS: Metadata Encoding & Transmission Standard*, cit.

Abstract

In response to the growing need for solutions focused on long-term preservation of scholarly journals, Portico was launched in 2005. Portico's mission is to preserve scholarly literature published in electronic form and to ensure that these materials remain accessible to future scholars, researchers and students. The article provides the history of this new electronic archiving service and describes its archival approach and preservation process.