DigItalia

Anno III, Numero 1 - **2008**

Rivista del digitale nei beni culturali

ICCU-ROMA

Il "documento digitale": analisi di un concetto in evoluzione

Mario Sebastiani
ICCU

Che cos'è un "documento digitale"? La risposta presuppone a sua volta una qualche definizione di "documento digitale". Ma questo è un concetto mutevole che cambia a seconda delle sollecitazioni provenienti dal campo dell'interoperabilità tra sistemi e della digital preservation.
Lo stato dell'arte oggi è tale che possiamo parlare dei "documenti digitali" come se fossero oggetti ben conosciuti, ma appena tentiamo di darne una definizione rigorosa, restiamo letteralmente senza parole. L'impulso spontaneo, nel tentare di definire questo concetto, è quello di equiparare il documento digitale ad un file di computer. Questo rende il documento digitale simile, sotto il profilo dell'unitarietà, al documento tradizionale. Ma, in realtà, un singolo documento digitale, strutturato in maniera adeguata per ottemperare ai vincoli connessi all'interoperabilità e alla preservation, è composto da una molteplicità di file distinti che assolvono a molteplici compiti differenti.
In primo luogo vi sono i file di contenuto e i file di metadati. Ma anche i file di contenuto, a loro volta, possono essere molteplici: ad esempio testo e immagini, di uno stesso documento digitale, possono essere contenuti in file separati. Vi sono poi metadati per i contenuti e metadati per gli aspetti tecnici dei file (formati, ecc.). La complessità, come è evidente, è notevole. Eppure la propensione a considerare il documento digitale come sostanzialmente analogo al documento tradizionale, persiste tenacemente. Probabilmente, contribuisce a ciò lo sforzo, lodevole, di rendere quanto più possibile "amichevoli" le tecnologie dell'informazione.
Quella che segue è un'indagine non sistematica e non esaustiva nel calderone degli standard e dei progetti attuali in materia di digitalizzazione, sufficiente però a produrre l'impressione che il concetto di documento digitale, lungi dall'essere definito una volta per tutte, sia tuttora un concetto in evoluzione.

Che cos'è un "documento digitale"? Difficilmente chi è coinvolto in progetti di digitalizzazione rimane indifferente alla risposta da dare a questa domanda. In particolare, chi opera in campo bibliotecario e culturale in genere, potrebbe chiedersi se i documenti digitali siano o meno assimilabili, da un punto di vista astratto, ai tradizionali documenti cartacei. Una risposta a questo interrogativo presuppone a sua volta una qualche definizione di documento digitale. Al riguardo, oggi, ve ne sono molte. Ma un'indagine sommaria su questo punto, condotta tra progetti e proposte varie di biblioteca digitale, *digital repositories*,

formati di dati, strutture di metadati e quant'altro, senza nessuna pretesa di esausitività e senza un piano di indagine preciso, ha prodotto un insieme esteso e confuso di varie definizioni. Nel tentare di mettere un po' di ordine in questo calderone, infarcito di neologismi tecnici inglesi, se ne è ricavata l'impressione che il concetto di documento digitale, lungi dall'essere definito una volta per tutte, sia tuttora un concetto in evoluzione. È, per così dire, un'idea sottoposta a una forte pressione evolutiva. Da dove viene questa pressione? Principalmente, a nostro avviso, dalle difficoltà connesse alla interoperabilità tra sistemi e alla *digital preservation.*

Come è noto, con interoperabilità si intendono questioni come il "Web semantico" e l'accesso alla "copia più adeguata". Il Web semantico non è altro che quell'insieme di procedure gestionali e di strutturazione dell'informazione – in gran parte ancora di là da venire – le quali dovrebbero consentire di accedere alle pagine Web oltre che mediante i tradizionali motori di ricerca, basati sulla presenza o meno di determinate parole in una data pagina, anche con sistemi in grado di tener conto, in qualche modo, del significato delle parole stesse. Con "copia più adeguata", invece, si indicano quei sistemi di accesso all'informazione sensibili al contesto, cioè in grado di discernere, tra formati digitali differenti di uno stesso documento, quello più adeguato alle particolari necessità di un dato utente.

Infine, con il neologismo *digital preservation*, ci si riferisce, come è noto, al fatto che, a breve termine, rischiano di scomparire tutti i materiali digitali che non siano stati trattati e gestiti secondo tecniche e procedure finalizzate alla conservazione a lungo termine. In altri termini, le informazioni digitali che non siano state strutturate e che non siano gestite secondo procedure adeguate alla conservazione a lungo termine, rischiano, per così dire, di essere gettate via insieme agli hardware e ai software obsoleti (i quali, come è noto, hanno cicli di vita computabili nell'ordine dei mesi piuttosto che degli anni).

Per affrontare quest'insieme di problemi, sono state lanciate numerose proposte e iniziative. Ma questo settore di studio è caratterizzato da una notevole confusione. Uno dei motivi è certamente il grande numero delle proposte e la loro estrema varietà. Ma ve ne sono anche altri. Ad esempio, vi sono iniziative che tentano di rispondere a più problemi contemporaneamente. Altre, invece, che sono state studiate per risolvere determinati tipi di problemi, vengono poi adattate alla soluzione di problemi di tutt'altro genere. Insomma, grande è la confusione sotto il cielo del digitale. Tuttavia, di progetto in progetto, l'analisi dei problemi si va focalizzando con sempre maggior precisione. Nuovi concetti, via via elaborati, permettono di fronteggiare con maggior cognizione di causa i problemi sul tappeto. A questo contribuisce anche la riformulazione dei concetti già utilizzati. Come quello di "documento digitale". L'analisi che segue tenta di ricostruire sommariamente le vicende più recenti di questo termine e dei suoi sinonimi.

Prima di tutto, i metadati

Il punto di partenza è costituito dalla nozione di risorsa elettronica (*electronic resource*). Questo termine, sul finire degli anni '90, è stato introdotto dall'International Federation of Library Associations (IFLA) nell'insieme delle norme catalografiche ISBD (International Standard Bibliographic Description) per fronteggiare l'emergere di nuove realtà quali i prodotti multimediali interattivi, lo sviluppo della tecnologia ottica, la disponibilità di risorse elettroniche remote, il World Wide Web, ecc. L'IFLA, che cura appunto le norme ISBD, posta di fronte a queste nuove realtà, aveva ritenuto necessario coniare un nuovo termine che differenziasse, ai fini pratici, queste risorse dai tradizionali documenti cartacei. Il termine prescelto fu appunto "risorsa elettronica", giudicato più appropriato del precedente "archivio elettronico"[1]. Le risorse elettroniche vennero poi suddivise, ai fini catalografici, tra risorse ad "accesso locale" (ad esempio CD-ROM) e risorse ad "accesso remoto" (ad esempio siti Web), ma senza fornire una precisa definizione formale di che cosa si intendesse per "risorsa elettronica", in particolare per "risorsa elettronica remota". Questo approccio è ancora considerato soddisfacente in svariati contesti. Ad esempio lo ritroviamo utilizzato nelle nuove *Regole italiane di catalogazione (REICA)* le quali distinguono appunto tra documenti elettronici disponibili "su supporti materiali" oppure "accessibili a distanza". In queste regole i documenti elettronici figurano come un caso particolare entro una varia tipologia di materiali documentari come testi, musica scritta, documenti grafici, documenti cartografici, audioregistrazioni, videoregistrazioni, filmati[2].

In altri contesti però troviamo definizioni più approfondite. È il caso ad esempio del manuale per i MAG (Metadati Amministrativi e Gestionali), edito dall'ICCU[3]. È un luogo comune ormai, ma ripeterlo non nuoce, che i documenti – "elettronici" o "digitali" che dir si voglia – per poter essere fruiti devono essere corredati dai cosiddetti "metadati". Questi sono letteralmente "dati sui dati", cioè informazioni che illustrano caratteristiche salienti di altri dati. In pratica, sono il frontespizio elettronico dei documenti digitali. In altri termini sono informazioni strutturate relative a un documento digitale – testo, audio, immagine o video – le quali consentono di utilizzare e gestire il documento secondo varie finalità. In realtà anche la nozione di metadati, come quella di documento digitale, soffre di margini consistenti di ambiguità: basti considerare che tanto le informazioni contenute in un

[1] *SBD(ER): International Standard Bibliographic Description for Electronic Resources*, edizione italiana a cura dell'Istituto centrale per il catalogo unico delle biblioteche italiane e per le informazioni bibliografiche, Roma: ICCU, 2000, p. 6.

[2] Commissione RICA, *Regole italiane di catalogazione (REICA): bozza complessiva: appendice D - Designazioni specifiche del materiale*, aggiornamento 10 febbraio 2008, Roma: ICCU, 2008, p. 352, http://www.iccu.sbn.it/upload/documenti/ReicaFeb2008.pdf?l=it.

[3] Comitato MAG, *Metadati amministrativi e gestionali: manuale utente*, a cura di Elena Pierazzo, versione 2.0.1, marzo 2006, Roma: ICCU, 2006, p. 7, http://www.iccu.sbn.it/upload/documenti/Manuale.pdf.

catalogo di biblioteca, quanto i permessi di accesso di un *file system*, sono, a tutti gli effetti, "metadati" [4].

METS (Metadata Encoding&Transmission Standard)

Quello che ci interessa, in questa sede, è che qualsiasi forma di interazione con un documento digitale presuppone che siano stati compilati i relativi metadati. L'importanza dei metadati, ai fini della fruizione dei documenti digitali, è tale da farli includere nella definizione stessa di documento digitale. Un esempio in questo senso è rappresentato appunto dal Manuale MAG nel quale si legge che i metadati devono essere considerati come:

> «parte costituente della definizione stessa di oggetto digitale: una risorsa digitale è inseparabilmente composta dal contenuto informativo (una sequenza di bit) e da una serie di informazioni (metadati) tali da rendere quella sequenza di bit significante, individuabile, e accessibile per la fruizione, l'archiviazione, la conservazione, la disseminazione e le altre operazioni gestionali»[5].

Qui, chiaramente, la locuzione "oggetto digitale" può essere considerata sostanzialmente sinonima di "documento digitale". Una nota, poi, spiega che l'origine di questa nozione di "oggetto digitale" risale a una definizione formulata nel 2001 dalla California Digital Library (CDL), secondo la quale per oggetto digitale si doveva intendere:

> «un qualcosa (ad esempio un'immagine, una registrazione audio, un documento testuale) che è stato codificato in modo digitale e integrato con metadati tali da supportarne l'individuazione, l'uso e l'immagazzinamento»[6].

Però, dalla formulazione primitiva a quella attuale, la definizione di "oggetto digitale" della CDL è mutata considerevolmente. Al centro di questa evoluzione vi è il rapporto tra dati e metadati. Come infatti si può leggere oggi nel glossario degli oggetti digitali (*Digital Objects Glossary*) della CDL, per "oggetto digitale" deve intendersi un'entità nella quale uno o più file di contenuto (*content files*) e i loro corrispondenti metadati, sono legati stabilmente tra loro, fisicamente e/o logicamente, mediante l'uso di un "connettore digitale" (*digital wrapper*)[7].

[4] In informatica con *file system* si intende quella componente del sistema operativo deputata alla archiviazione dei file su un supporto di memorizzazione non volatile come un hard-disk o un CD-ROM.

[5] Comitato MAG, *Metadati* cit., p. 7.

[6] Comitato MAG, *Metadati* cit., p. 7, nota 4.

[7] California Digital Library, *Digital Objects Glossary*, a cura di Jennifer Colvin, aggiornamento 15 maggio 2006, http://www.cdlib.org/inside/diglib/glossary/?field=glossary&action=search&query=oac.

Per file di contenuto, conformemente alla nozione comune, si intendono i file audio, immagine, video o testo, sia nati in formato digitale, sia catturati da documenti analogici mediante applicazioni software. Uno stesso file di contenuto può essere distribuito secondo differenti formati file (*file format*); ad esempio da una immagine TIFF si possono derivare immagini JPEG e GIF per finalità di distribuzione o tutela dei diritti d'autore. Il glossario CDL distingue poi tra "oggetti digitali semplici", cioè composti da un singolo file di contenuto (*content file*) e "oggetti digitali complessi", cioè composti da più file di contenuto (*content files*).
Un connettore digitale invece è definito come un file di testo strutturato il quale lega insieme tra loro i file di contenuto e i metadati associati. Dal punto di vista operativo si possono distinguere due casi:

– file di contenuto e metadati sono inglobati tutti insieme all'interno del connettore digitale;
– file di contenuto e metadati sono archiviati separatamente dal connettore digitale.

Nel primo caso si parlerà di connessione fisica (*physical wrapping*) nel secondo caso di connessione logica (*logical wrapping*). In questo secondo caso viene demandato al connettore fisico il compito di mantenere le connessioni, mediante puntatori logici, tra i file di contenuto e i metadati. Un oggetto digitale può condividere entrambe queste modalità di *wrapping*. Da notare, al riguardo, che nelle regole *REICA* citate prima, viene preso in considerazione solo il *wrapping* fisico. Qui, infatti, la fonte primaria dalla quale acquisire la descrizione bibliografica delle pubblicazioni elettroniche accessibili a distanza è costituita unicamente dalle «informazioni presentate formalmente al principio di un file, i metadati inclusi o collegati con il contenuto»[8].
Uno standard per il *wrapping* dei materiali delle biblioteche digitali – leggiamo ancora sul glossario CDL – è costituito dal METS (Metadata Encoding & Transmission Standard). Questo standard costituisce un'iniziativa della Digital Library Federation ed è mantenuto dalla Library of Congress[9]. L'obiettivo è quello di realizzare uno schema di documento in formato XML per la codifica dei metadati necessari alla gestione e allo scambio degli oggetti digitali contenuti in sistemi per la gestione e l'archiviazione dei documenti digitali. A questo riguardo, il cuore di un documento METS è costituito dalla cosiddetta "mappa strutturale". Questa componente del documento METS ha il compito di mettere in evidenza le gerarchie della struttura informativa nella quale è inclusa l'oggetto digitale e di fornire i col-

[8] Commissione RICA, *Regole italiane* cit., p. 53.
[9] Metadata Encoding & Transmission Standard, http://www.loc.gov/standards/mets/mets-home.html.

legamenti tra gli elementi di questa struttura, i file di contenuto e i metadati specifici di ogni elemento[10].
Da notare infine che, per quanto riguarda la struttura di fondo, un documento METS si conforma abbastanza fedelmente al modello OAIS. Quest'ultimo oggi è forse il modello di riferimento più importante per impostare una gestione adeguata dei documenti digitali ai fini della *digital preservation*.

OAIS (Open Archival Information System)

Lo standard Sistema informativo aperto per l'archiviazione (OAIS – Open Archival Information System) è nato nel 1997 ad opera del Comitato consultivo per i sistemi di dati spaziali (Consultative Committee for Space Data Systems – CCSDS), un ente fondato nel 1982 come organo di coordinamento per le agenzie spaziali interessate allo sviluppo di standard per la gestione dei dati relativi alla ricerca spaziale[11]. L'oggetto fondamentale in questo modello è rappresentato dal cosiddetto oggetto informativo (*information object*) che, a sua volta, risulta composto da due elementi: l'oggetto digitale vero e proprio (*data object*), cioè un oggetto composto da un insieme di sequenze di bit, e le informazioni su tale oggetto, dette informazioni sulla rappresentazione (*representation informations*), cioè i metadati. Il modello OAIS si caratterizza, rispetto agli altri, per il fatto che l'elemento fondamentale, cioè l'oggetto informativo, non è dato una volta per tutte, ma cambia a seconda delle funzioni che è chiamato via via ad assolvere. Le funzioni previste sono tre: l'immissione dei dati all'interno del sistema; la conservazione a lungo termine delle informazioni; la distribuzione delle informazioni agli utenti. Per ognuna di queste tre funzionalità sono previste tre diverse configurazioni dell'oggetto informativo chiamate "pacchetti di informazioni" (*information packages*). Avremo perciò:

– il "pacchetto di versamento" (SIP – Submission Information Package) per l'immissione dei dati;
– il "pacchetto di archiviazione" (AIP – Archival Information Package) per l'archiviazione;
– il "pacchetto di distribuzione" (DIP – Dissemination Information Package) per distribuire le informazioni agli utenti.

[10] *Tutorial METS: quadro generale*, http://www.loc.gov/standards/mets/METSita.html.
[11] Consultative Committee for Space Data Systems, *Reference Model for an Open Archival Information System (OAIS): Recommendation for Space Data System Standards (Blue Book), CCSDS 650.0-B-1*, 2002, http://public.ccsds.org/publications/archive/650x0b1.pdf. Traduzione italiana: *OAIS: sistema informativo aperto per l'archiviazione*, a cura di Giovanni Michetti, Roma: ICCU, 2007.
Il modello OAIS è stato recepito anche dall'International Standards Organization (ISO) ed è identificato come ISO 14721.

Per ognuna di queste tre diverse tipologie di pacchetti informativi il modello prevede poi ulteriori informazioni finalizzate a garantirne la conservazione; si chiamano appunto "informazioni sulla conservazione" (PDI – Preservation Description Information). Queste informazioni sono articolate secondo quattro aree distinte, ognuna delle quali assolve a una specifica esigenza rilevante ai fini della conservazione:

- "informazioni sull'identificazione" (*reference information*): informazioni rappresentate da un identificatore univoco per il contenuto, ad esempio ISBN o URL;
- "informazioni sul contesto" (*context information*): informazioni che descrivono le relazioni tra il contenuto del pacchetto informativo e i contenuti di altri pacchetti;
- "informazioni sulla provenienza" (*provenance information*): informazioni che documentano l'iter del contenuto, cioè la creazione, le trasformazioni subite, i passaggi di possesso e di proprietà, ecc.;
- "informazioni sull'integrità" (*fixity information*): informazioni sull'integrità del contenuto, cioè informazioni mediante le quali è possibile verificare che il contenuto non abbia subito alterazioni non autorizzate (ad esempio *checksum*, impronte e firme digitali, ecc.)[12].

Riguardo ai rapporti con lo standard OAIS, il modello METS, citato prima, può essere utilizzato per codificare altrettanto bene sia i pacchetti di versamento (SIP), sia i pacchetti di archiviazione (AIP), sia i pacchetti di distribuzione (DIP)[13].

Lo spazio dei nomi

Vi è però un aspetto importante nella gestione dei documenti digitali che nel modello OAIS viene affrontato solo incidentalmente: l'identificazione stabile degli oggetti digitali. Il modello OAIS colloca la soluzione di questo problema al di fuori del suo ambito di applicazione. Questo probabilmente perché l'identificazione stabile degli oggetti digitali è più importante ai fini dell'interoperabilità tra sistemi che non ai fini della *digital preservation.* Due distinti archivi di oggetti digitali possono avere identificatori uguali per gli oggetti digitali a condizione di non interagire tra loro. Se però devono interagire allora è necessario che gli oggetti digitali scambiati tra di loro abbiano degli identificatori assoluti. Cioè non vi deve essere confusione tra gli identificatori di un sistema e quelli dell'altro. Il modello OAIS affronta il tema dell'identificazione stabile degli oggetti digitali nella sezione dedicata alla creazione di cosiddetti "consorzi" di archivi OAIS.

[12] *OAIS* cit., p. XI-XXX.
[13] *Tutorial METS* cit.

Nel caso che due o più archivi OAIS formino un "consorzio" sarà necessario che vengano predisposti "identificatori univoci" per i pacchetti di archiviazione (AIP) distribuiti all'interno del consorzio. Lo standard suggerito a questo scopo nel modello OAIS è «la norma ISO X.500 per la denominazione dei servizi di directory»[14]. L'X.500 però non può essere esteso fino a risolvere il problema dell'identificazione permanente e globale degli oggetti digitali su una rete universale quale è oggi Internet. Per realizzare ciò sarebbe necessario implementare, all'interno dello standard, uno spazio dei nomi (*namespace*) universale. Questo è una sorta di identificatore il quale assegna un significato univoco a elementi i quali, in contesti diversi, assumono significati diversi. Ad esempio, nel linguaggio corrente, il termine "libreria" può essere inteso tanto come "negozio di libri" (quindi appartenente allo spazio dei nomi, ipotetico, denominato "Commercio") sia come "insieme dei programmi elementari" di un'applicazione software o di un sistema operativo (quindi appartenente allo spazio dei nomi, anch'esso ipotetico, denominato "Informatica"). Formalmente, in un documento XML, mediante l'operatore xmlns, questi due diversi usi del termine "libreria" verrebbero espressi così:

<libreria xmlns="namespaceCommercio>
<libreria xmlns="namespaceInformatica>

In altri termini, in un documento XML, l'operatore xmlns permetterebbe di indicare a quale dei due possibili usi del termine "libreria" ci si vuole riferire. Contestualmente, "Commercio" e "Informatica" assolverebbero al ruolo di identificatori di due distinti spazi dei nomi all'interno dei quali lo stesso termine "libreria" viene utilizzato con significati diversi. Analogamente, in una rete globale, dove ogni sistema locale collegato può avere un proprio spazio dei nomi, per evitare conflitti e ambiguità è necessario che, al livello più alto, tutto sia ricondotto nell'ambito di un unico complessivo spazio dei nomi. Originariamente lo spazio dei nomi globale dell'X.500 doveva essere costituito dal sistema di identificazione dei paesi a due lettere codificato nello standard ISO 3166[15]. Ma questa è rimasta finora solo un'ipotesi. Per cui oggi X.500, o per meglio dire la versione di X.500 adattata al protocollo TCP/IP (Transmission Control Protocol/Internet Protocol), nota con l'acronimo LDAP (Lightweight Directory Access Protocol), utilizza come spazio dei nomi globale il comune spazio dei nomi di Internet, vale a dire il DNS (Domain Name System).

[14] *OAIS* cit., p. 150. In realtà, quella alla quale si fa riferimento nel documento OAIS, è una serie di standard emessi separatamente dall'International Telecommunication Union-Telecommunication Standardization Bureau (ITU-T), da una parte, e dall'International Standards Organization-International Electrotechnical Commission (ISO/IEC) dall'altra. Fino al 1993 l'ITU era noto come Comité consultatif international téléphonique et télégraphique (CCITT).

[15] David W. Chadwick, *Understanding X.500 – The Directory*, 1994, http://sec.cs.kent.ac.uk/x500book.

Il DNS, però, ha delle limitazioni intrinseche che lo rendono poco adatto all'identificazione globale permanente degli oggetti digitali[16].
In questa sede, per ciò che concerne la nostra indagine, interessa sottolineare che differenti soluzioni al problema dell'identificazione degli oggetti digitali riflettono concezioni diverse della nozione di "documento digitale". Gli indirizzi fondamentali, al riguardo, sembrano essere due: uno che assimila i documenti digitali ai tradizionali documenti cartacei; l'altro invece che tratta i documenti digitali come entità assolutamente originali e peculiari. Riconosciamo il primo indirizzo nel RDF (Resource Description Framework), progetto ispirato alle idee di Tim Berners-Lee, noto come ideatore del Web; riconosciamo invece il secondo indirizzo nel sistema di identificazione persistente Handle System, ideato da Robert Kahn, creatore, insieme a Vinton Cerf, dei protocolli TCP/IP. L'Handle System costituisce il punto di riferimento intorno al quale è costruito il sistema di archiviazione digitale, o meglio il *repository* digitale DSpace. Vediamolo.

DSpace

Nel 1995 Robert Kahn, nell'articolo intitolato *Un'infrastruttura di servizi per oggetti digitali diffusi*, scritto con Robert Wilensky, della University of California a Berkeley scriveva:

> «Formalmente un oggetto digitale costituisce la manifestazione concreta di un tipo astratto di dati che ha due componenti "dati" e "metadati-chiave". I dati sono scritti come specificato in seguito. I metadati-chiave includono un *handle* vale a dire un identificatore globalmente unico per l'oggetto digitale. […] Un *repository* è un sistema di archiviazione accessibile in rete nel quale gli oggetti digitali possono essere archiviati per utilizzazioni successive. […] I *repositories* hanno nomi unici ufficiali assegnati o approvati da una *naming autority* globale in modo da garantirne l'unicità»[17].

[16] Su «Dig*Italia*» sono già state illustrate, sommariamente, queste limitazioni. In particolare sono stati descritti i protocolli TCP/IP e DNS e le relazioni tra questi e i principali sistemi di identificazione persistente degli oggetti digitali.
Mario Sebastiani, *Identificatori persistenti per gli oggetti digitali*, «Dig*Italia*», 2005, n. 0, p. 62-82, http://digitalia.sbn.it/upload/documenti/digitalia20050_SEBASTIANI.pdf.
Un'esposizione esaustiva e aggiornata degli identificatori persistenti attualmente disponibili si trova in: Hans-Werner Hilse – Jochen Kothe, *Implementing Persistent Identifiers: Overview of concepts, guidelines and recommendations*, London: Consortium of European research Libraries, Amsterdam: European Commission on Preservation and Access, 2006, http://www.cerl.org/, http://www.knaw.nl/ecpa.

[17] Robert Kahn – Robert Wilensky, *A framework for distributed digital object services*, 13 maggio 1995, http://www.cnri.reston.va.us/k-w.html, doi:cnri.dlib/tn95-01; 13 marzo 2006, http://www.doi.org/topics/2006_05_02_Kahn_Framework.pdf, doi:10.1007/s00799-005-0128-x; «International Journal on Digital Libraries», 6 (2006), n. 2, p. 115-123 («Formally, a digital object is an instance of an abstract data type that has two components, "data" and "key-metadata". The data is typed, as is described below. The key-metadata includes a "handle", i.e., an identifier globally unique to the digital object. […] A repository is a network-accessible storage system in which digital

In realtà Kahn è universalmente conosciuto per aver messo a punto, dal 1973 al 1978, insieme a Vinton Gray Cerf, l'insieme dei protocolli di trasmissione TCP/IP, oggi a base delle più comuni reti domestiche e aziendali e della stessa Internet. Attualmente Kahn presiede la Corporation for National Research Initiatives (CNRI) una organizzazione *no-profit*, fondata nel 1986, che ha come missione quella di promuovere ricerche e progetti nell'ambito dell'infrastruttura informativa nazionale USA[18]. Il CNRI contribuisce, tra l'altro, nell'ambito della D-Lib Alliance, alla pubblicazione della rivista online «D-Lib Magazine»[19].
Lo Handle System è stato sviluppato dal CNRI sulla base di un finanziamento della Defense Advanced Research Projects Agency (DARPA). La sua prima implementazione è avvenuta nel 1994. Il sistema si articola, sostanzialmente, in un insieme di protocolli, uno spazio dei nomi e un software di indirizzamento. Tra i requisiti principali previsti per il sistema, figurano:

– unicità dei nomi, vale a dire unicità degli *handles*;
– persistenza, vale a dire sussistenza di un indirizzo programmatico di mantenimento di una connessione tra lo *handle* e l'entità identificata;
– istanze multiple, vale a dire capacità degli *handles* di indirizzare istanze multiple di una data risorsa digitale.

Tecnicamente parlando, la nota qualificante di questo sistema è che, sebbene compatibile con i *root server* del sistema DNS – vale a dire quella dozzina di server sparsi per il mondo dedicati alla gestione dei domini *top-level* della rete Internet – esso dispone anche di un proprio *root server*, il Global Handle System, mantenuto dal CNRI. Questo consente allo Handle System di accedere alle risorse disponibili in rete sia tramite le comuni URL basate sul protocollo HTTP, sia con un proprio particolare protocollo di identificazione che è esente da varie limitazioni connesse allo standard DNS (in primo luogo la non-persistenza delle URL). Per aderire allo Handle System bisogna registrare una propria *naming authority* presso il CNRI il quale, a fronte del pagamento di una modica quota, rilascia una licenza e il software gestionale. Tra i vari progetti che attualmente si avvalgono dello Handle System, figurano:

– la Defense Virtual Library, un progetto pilota di biblioteca digitale a cura del DARPA e del CNRI[20];

objects may be stored for possible subsequent access or retrieval. […] Repositories have official, unique names, assigned or approved to assure uniqueness by a global naming authority»).
[18] Corporation for National Research Initiatives, http://www.cnri.reston.va.us/about_cnri.html.
[19] *D-Lib Alliance Participants*, http://www.dlib.org/dlib/alliance-participants.html.
[20] *Defense Virtual Library*, http://www.dtic.mil/dtic/prodsrvc/rdi/dvl.html.
Nell'ambito della Defense Virtual Library il Defense Technical Information Center (DTIC) si avvale dell'Handle System, http://www.dtic.mil/dtic/handles/index.html.

– il DOI (Digital Object Identifier)[21];
– DSpace, un *repository* open source per oggetti digitali[22].

DSpace è stato sviluppato nel 2000 nell'ambito di un progetto congiunto del Massachussetts Institute of Technology (MIT) con la Hewlett-Packard. Il progetto era finalizzato alla realizzazione di un software applicativo per il deposito, la conservazione e la distribuzione degli oggetti digitali. Il sistema è una piattaforma open source, disponibile gratuitamente per tutti, in grado di riconoscere e gestire un gran numero di formati di file e di tipi di dati[23]. Tra questi: Adobe PDF, documenti Word, immagini JPEG e TIFF, video MPEG, ecc. Per ogni file, il sistema garantisce la cosiddetta *bit integrity*, assicura cioè l'integrità di tutti i bit, uno per uno, tramite controllo delle *checksum* MD5[24]. Una *history* consente di ricostruire tutte le modificazioni, legittime, riportate da un dato file e dai suoi metadati associati. Il deposito del file è facilitato da un'interfaccia Web. DSpace si impegna a offrire funzioni di preservazione che consentano di mantenere accessibile il file anche di fronte dell'evoluzione dei formati tecnologici, dei media e dei paradigmi nel corso del tempo[25].
Questa piattaforma, in altri termini, non è un mero sistema di gestione di dati, bensì un vero e proprio *repository* orientato alla *digital preservation*. Recentemente, Sally Hubbard, del Getty Research Institute of Los Angeles, ha così riassunto, su «Dig*Italia*», le funzionalità fondamentali che un *repository* non può fare a meno di garantire:

– identificazione persistente: cioè mantenimento lungo tutto l'arco di vita di un dato oggetto digitale della capacità del sistema di identificare tale oggetto;
– validazione dell'oggetto digitale: capacità di determinare il formato di file dell'oggetto e di convalidarlo a fronte di specificazioni standard;
– "agnosticismo" rispetto ai metadati: cioè capacità di supportare, registrare e gestire qualsiasi insieme di metadati;
– "agnosticismo" rispetto ai formati: ovvero capacità di gestire qualsiasi formato di file;

[21] Cfr. Hilse, *Implementing Persistent Identifiers* cit., p. 21-25; Digital Object Identifier, http://www.doi.org.
[22] Hilse, *Implementing Persistent Identifiers* cit., p. 17-20.
[23] DSpace può essere scaricato liberamente da Sourceforge, un *repository* di software open source (http://sourceforge.net/projects/dspace) sotto licenza BSD (Berkeley Software Distribution, http://www.opensource.org/licenses/bsd-license.php).
[24] MD5 è un algoritmo crittografico suscettibile di varie applicazioni tra cui la verifica dell'integrità di un file dopo il suo trasferimento su un altro supporto. La verifica avviene mediante il raffronto di stringhe di controllo (*checksum*) generate prima e dopo il trasferimento. MD5 è descritto nella RFC 1321 dell'aprile 1992, http://tools.ietf.org/html/rfc1321.
[25] DSpace, http://www.dspace.org.

- *fixity checking*: capacità di controllare, nel corso del tempo, l'integrità dei flussi di bit;
- capacità di normalizzazione e migrazione: possibilità di far fronte alle necessità di trasformazione dei dati imposte dall'obsolescenza tecnologica.

Naturalmente, come specifica Sally Hubbard, questo elenco, come ogni altro del genere, va considerato come altamente speculativo: nessuno infatti, finora, ha fatto concretamente l'esperienza di conservare oggetti digitali per decenni o più[26]. Tuttavia è certamente significativa, in questo elenco, la priorità assegnata all'identificazione permanente degli oggetti digitali. DSpace, a questo riguardo, è stato implementato in modo da garantire prima di tutto questa priorità. Lo strumento utilizzato a questo scopo è appunto l'Handle System. La cosa interessante è che, nell'architettura che ne consegue, risultano del tutto assenti tanto la nozione di "oggetto digitale" quanto quella di "documento digitale". Infatti un *repository* DSpace appare, da un punto di vista logico, come una struttura piramidale articolata, partendo dal basso, nei seguenti componenti:

- i flussi di bit (*bitstream*);
- gli atomi d'archivio (*archival atoms*), più semplicemente *item*;
- le collezioni (*collections*);
- le comunità (*communities*).

I flussi di bit corrispondono, approssimativamente, all'usuale nozione di file; quindi vi possono essere tanto flussi di bit di dati, cioè di contenuti, che flussi di bit di metadati. Più precisamente: un flusso di bit è semplicemente una sequenza di bit riproducibile mediante un determinato formato. A ogni flusso di bit, quindi, sono associate le informazioni relative al formato tecnico.
La nozione più interessante però è quella di atomo d'archivio, ovvero *item*. Questo è il termine che di fatto, in DSpace, sostituisce i consueti termini "oggetto digitale" e "documento digitale". L'*item* è un raggruppamento di vari contenuti (*bitstream* di contenuti) e metadati (*bitstream* di metadati) che sono correlati tra di loro secondo varie modalità. Gli *item*, a loro volta, possono essere raggruppati in collezioni (*collections*) le quali, a loro volta, possono essere raggruppate in comunità (*communities*)[27].
DSpace quindi, fornisce un sistema complessivo di gestione degli oggetti digitali che garantisce la loro conservazione duratura e, con essa, la loro identificazione persistente. L'architettura conseguente a tali finalità, però, è così peculiare che, all'interno di essa, risulta più opportuno riferirsi agli oggetti trattati, mediante un termine affatto nuovo, cioè il termine *item* (ovvero "atomo d'archivio").

[26] Sally Hubbard, *Getting to the web*, «Dig*Italia*», 2007, n. 2, p. 11-19.
[27] Diagramma DSpace, http://www.dspace.org/images/stories/dspace-diagram.pdf.

RDF (Resource Description Framework)

Finora abbiamo visto iniziative e progetti il cui baricentro gravita nell'area della *digital preservation.* Con il RDF (Resource Description Framework) prendiamo ora in considerazione una iniziativa focalizzata principalmente sul tema dell'interoperabilità tra sistemi. RDF è una infrastruttura per descrivere risorse e metadati presenti sul Web. RDF costituisce anche un'infrastruttura per lo scambio interoperabile dei dati, con particolare riguardo al Web semantico. Secondo il World Wide Web Consortium (W3C) l'RDF è un linguaggio disegnato per supportare il Web semantico in maniera analoga a come l'HTML (HyperText Markup Languge) ha reso inizialmente possibile l'evoluzione del Web. Nel 2001, in un articolo su Scientific American, Tim Berners-Lee ha scritto a proposito del Web semantico:

> «Il Web semantico darà struttura al contenuto significativo delle pagine Web, creando un ambiente dove gli agenti software possano svolgere velocemente compiti complessi per i loro utenti. Uno di questi agenti che arrivi alla pagina Webdel Centro fisioterapico non saprà solo che quella pagina contiene parole chiave come "trattamento, medicina, fisico, terapia", ma anche che il dottor Hartmann effettua visite il lunedì, mercoledì e venerdì, e che il programma per gli appuntamenti accetta le date nel formato gg/mm/aaaa e presenta un elenco degli orari disponibili»[28].

RDF persegue la finalità di realizzare il Web semantico fornendo un metodo generale per decomporre la conoscenza in piccoli pezzi, insieme ad alcune regole concernenti la semantica o il significato di questi pezzi. RDF assomiglia a XML. La conoscenza, secondo la metodologia RDF, può essere decomposta in piccoli mattoni di base, detti triple, costituiti ognuno da tre elementi: soggetto, oggetto, predicato. RDF fornisce una sorta di grammatica per combinare questi mattoni, le triple, e rappresentare con esse sul Web conoscenze complesse. Dato che ognuno degli elementi di una tripla deve poter essere utilizzato senza equivoci e ambiguità, vi sono degli identificatori persistenti globali per ognuno di questi elementi detti URI (Uniform Resource Identifier)[29]. Lo standard URI, in pratica, è lo spazio dei nomi dell'RDF. Alla base di questo standard vi è però una curiosa definizione di "risorsa". Infatti, nel 1998, questo standard veniva così definito: «Un URI è una stringa compatta di caratteri che identifica una risorsa fisica o astratta»[30]. Ma, a sua volta la nozione di risorsa era definita così: «Una

[28] Tim Berners-Lee – James Hendler – Ora Lassila, *Il Web semantico: quando Internet diventa intelligente: agenti software e rappresentazioni condivise per un'automazione dei servizi in rete… e a casa vostra,* «Le Scienze», 2001, n. 393, p. 78.
[29] Shelley Powers, *Practical RDF*, Beijing; Sebastopol: O'Reilly, 2003, p. 14-28.
[30] Tim Berners-Lee – Roy Fielding – Larry Masinter, *RFC 2396: Uniform Resource Identifiers (URI): Generic Syntax,* agosto 1998, http://gbiv.com/protocols/uri/rfc/rfc2396.html.

risorsa è qualsiasi cosa abbia identità»[31]. Per cui un "identificatore" URI, infine, risultava essere semplicemente «un oggetto che può agire come riferimento a qualcosa che ha identità»[32].
Naturalmente una definizione del genere non può non sollevare problemi. A cominciare dal fatto se esistano o meno eventuali risorse prive di identità. Questo, in realtà, sembra poco plausibile. Infatti, come osservava Quine, grande filosofo contemporaneo, già nel 1969, il nostro senso comune è governato da un precetto molto rigoroso: «Nessuna entità senza identità!»[33].
Ma l'identità degli oggetti digitali è problematica anche perché non possediamo ancora criteri certi per stabilire quando due distinti oggetti digitali sono espressioni dello "stesso" documento digitale. Come infatti si legge in un interessante documento della Conferenza Dublin Core del 2003:

> «Stranamente ci sono state poche ricerche sullo sviluppo formale delle condizioni di identità per un'entità che è fondamentale nella *library information science*: il documento, nel senso di una espressione simbolica astratta che può essere concretizzata fisicamente ripetutamente e tramite diversi media. Il risultato è che non solo questo concetto critico appare sotto-teorizzato, ma i progressi in una gran varietà di importanti questioni – come la *preservation*, la conversione, la sicurezza dell'integrità, il recupero, la collaborazione, i metadati, gli identificatori – sono stati ostacolati»[34].

Non sorprende quindi che nella nuova definizione dello standard URI, del 2005, la definizione di "risorsa" risulti leggermente modificata. Ora, risorsa è qualsiasi cosa possa essere identificata mediante un URI. Più precisamente:

> «il termine "risorsa" è utilizzato in un senso generale per qualsiasi cosa possa essere identificata da un URI. Esempi tipici sono un documento elettronico, un'immagine, una fonte di informazione con una finalità coerente (ad esempio "il bollettino meteorologico odierno per Los Angeles"), un servizio (ad esempio un *gateway* da HTTP a SMS) e una collezione di altre risorse.

[31] Berners-Lee, *RFC 2396*, cit. p. 2.
[32] Berners-Lee, *RFC 2396*, cit. p. 3.
[33] Willard Van Orman Quine, *La relatività ontologia e altri saggi*, Roma: Armando, 1986, p. 55 (Tit. orig. *Ontological Relativity and other Essays*, New York: Columbia, 1969).
[34] Allen Renear – David Dubin, *Towards Identity Conditions for Digital Documents*, 2003, http://www.dublincore.go.kr/dcpapers/pdf/2003/503_Paper71.pdf («Surprisingly then, there has been little research on developing formal identity conditions for an entity which is fundamental in library information science: the document, in the sense of an abstract symbolic expression that may be physically instantiated repeatedly and in various media. As a result, not only is this critical concept under-theorized, but progress on a number of important problems — including preservation, conversion, integrity assurance, retrieval, federation, metadata, identifiers — has been hindered»).

> Una risorsa non deve essere necessariamente accessibile attraverso Internet; ad esempio anche gli esseri umani, gli enti, i libri raccolti in una biblioteca possono essere considerati risorse»[35].

Mentre "identificatore" è ora definito così:

> «Un identificatore ingloba l'informazione necessaria per distinguere ciò che viene identificato da tutte le altre cose incluse nel proprio ambito di identificazione»[36].

Come si vede secondo lo standard URI – alla base dell'RDF – tanto un libro che un documento elettronico sono identificabili secondo le stesse modalità. Inoltre l'applicazione dello standard non è limitata ai soli documenti distribuiti attraverso Internet. Qualsiasi cosa può essere suscettibile di essere identificata mediante un URI.
Da notare infine, al termine di questa breve escursione nel regno della metafisica, che la definizione originaria URI permane nello standard Dublin Core[37]. Qui, infatti, risorsa è ancora «qualsiasi cosa che abbia identità»[38]. Mentre, contestualmente, per identificatore si deve intendere:

> «L'elemento Dublin Core che costituisce un riferimento non ambiguo alla risorsa nell'ambito di un determinato contesto. Una buona pratica raccomandata è di identificare la risorsa mediante una stringa o un numero che si conforma ad un sistema formale di identificazione»[39].

Anche Dublin Core quindi, come OAIS, esclude dal proprio ambito di competenza il problema dell'identificazione persistente degli oggetti digitali.

[35] Tim Berners-Lee – Roy Fielding – Larry Masinter, *RFC 3986: Uniform Resource Identifiers (URI): Generic Syntax*, gennaio 2005, p. 4, http://gbiv.com/protocols/uri/rfc/rfc3986.html («the term "resource" is used in a general sense for whatever might be identified by a URI. Familiar examples include an electronic document, an image, a source of information with a consistent purpose (e.g., "today's weather report for Los Angeles"), a service (e.g., an HTTP-to-SMS gateway), and a collection of other resources. A resource is not necessarily accessible via the Internet; e.g., human beings, corporations, and bound books in a library can also be resources»).
[36] Berners-Lee, *RFC 3986* cit. p. 5 («an identifier embodies the information required to distinguish what is being identified from all other things within its scope of identification»).
[37] Dublin Core Metadata Iniziative, http://dublincore.org.
[38] Dublin Core Metadata Initiative, *Glossary*, aggiornamento 23 aprile 2004, http://dublincore.org/documents/usageguide/glossary.shtml.
[39] *Ibidem* («Identifier. The Dublin Core element that is an unambiguous reference to the resource within a given context. Recommended best practice is to identify the resource by means of a string or number conforming to a formal identification system»).

PREMIS (PREservation Metadata Implementation Strategies)

Abbiamo visto quindi che la complessità dei documenti elettronici, legata principalmente al fatto di dover distinguere, in tali documenti, tra "dati" e "metadati", comporta, nella loro gestione, una gran mole di difficoltà tecniche e procedurali. Per risolvere queste difficoltà, sono state proposte soluzioni che, come abbiamo visto, assimilano totalmente i "documenti elettronici" ai documenti tradizionali cartacei (RDF). Altre soluzioni, invece, considerano i "documenti elettronici" come oggetti del tutto particolari: al punto che, per evitare equivoci, non vengono nemmeno più chiamati "documenti" (DSpace). Da notare che RDF è un'iniziativa orientata verso l'interoperabilità mentre DSpace è un progetto finalizzato principalmente alla *preservation*. Scegliere tra questi orientamenti, per ora, non sembra possibile perché, come appare evidente da quanto detto fin qui, il concetto di "documento digitale" è un concetto tuttora in evoluzione. Questo comporta che la fisionomia finale che esso è destinato ad assumere sarà determinata soprattutto dalle realizzazioni concrete via via implementate e da come, in queste realizzazioni, saranno stati affrontati e risolti i problemi relativi alla gestione dei "documenti digitali".

Per farsi un'idea quindi di come questo concetto potrebbe evolvere occorre osservare quello che si fa concretamente, nel campo della gestione dei documenti digitali, da noi o altrove. Ad esempio, lo scorso dicembre 2007 la National Library of Australia ha registrato il proprio profilo METS presso la Library of Congress. L'obiettivo è quello di realizzare una modalità comune per l'impacchettamento e lo scambio dei contenuti digitali all'interno di un futura infrastruttura comune Australiana per la gestione dei dati di ricerca che possono essere condivisi e riutilizzati. Quale che sia il formato di impacchettamento che verrà adottato – leggiamo in un articolo su «D-Lib» del marzo scorso firmato da vari collaboratori della National Library of Australia – esso dovrà gestire modelli complessi di contenuti e operare attraverso molteplici scenari di acquisizione e distribuzione. Questo dovrà essere fatto conservando memoria della catena di custodia degli oggetti attraverso il tempo. Per queste finalità la National Library of Australia ha ritenuto opportuno adottare il modello METS, secondo la particolare estensione che ne è stata data mediante PREMIS[40].

Il progetto PREMIS (PREservation Metadata Implementation Strategies) è stato promosso nel 2003 dall'Online Computer Library Center (OCLC) e dal Research Libraries Group (RLG). Gli obiettivi del progetto, leggiamo nel rapporto finale, pubblicato nel 2005, sono quelli di:

[40] Judith Pearce – David Pearson – Megan Williams – Scott Yeadon, *The Australian METS Profile – A Journey about Metadata*, «D-Lib Magazine», 14 (2008), n. 3/4, doi:10.1045/march2008-pearce, http://www.dlib.org/dlib/march08/pearce/03pearce.html.

- definire e implementare un insieme di base di elementi di metadati per la *digital preservation*;
- redigere la bozza di un Data Dictionary per supportare l'insieme dei metadati di base;
- esaminare e valutare differenti strategie per la codifica, l'archiviazione e il mantenimento dei *preservation metadata* nell'ambito di un sistema di *digital preservation*, anche ai fini dello scambio dei metadati tra sistemi diversi;
- avviare programmi pilota per testare le raccomandazioni del gruppo in una varietà di sistemi;
- esplorare opportunità per la creazione in forma cooperata e la condivisione dei *preservation metadata*[41].

L'aspetto caratterizzante del modello PREMIS è la distinzione tra "oggetti digitali" ed "entità intellettuali". Un "oggetto digitale" è definito come una «unità discreta di informazione in formato digitale»[42]. Una "entità intellettuale" invece è definita come:

> «un insieme coerente di contenuti che può essere ragionevolmente descritto come una unità, ad esempio un libro particolare, una mappa, una fotografia o un database. Una entità intellettuale può includere altre entità intellettuali; ad esempio un sito Web può includere una pagina web, una pagina Web può includere una fotografia. Una entità intellettuale può avere una o più rappresentazioni digitali»[43].

Di fatto PREMIS crea una distinzione tra la rappresentazione del contenuto intellettuale e l'oggetto digitale vero e proprio (file, *bitstream*) in un modo che ricorda molto da vicino lo schema FRBR (Functional Requirements for Bibliographic Records)[44]. Certo non si può dire adesso su due piedi se questo è effettivamente l'approccio destinato ad affermarsi stabilmente nel prossimo futuro. Ma se così dovesse essere, allora dovrebbe diventare sempre più difficile, in futuro, riferirsi ai "documenti digitali" prescindendo totalmente dal loro effettivo contenuto intellettuale.

[41] Preservation Metadata: Implementation Strategies, *Data Dictionary for Preservation Metadata*, maggio 2005, p. 7, http://www.oclc.org/research/projects/pmwg/premis-final.pdf.
[42] Ivi, p. 1-1.
[43] *Ibidem* («An Intellectual Entity is a coherent set of content that is reasonably described as a unit, for example, a particular book, map, photograph, or database. An Intellectual Entity can include other Intellectual Entities; for example, a Web site can include a Web page, a Web page can include a photograph. An Intellectual Entity may have one or more digital representations»).
[44] *Requisiti funzionali per record bibliografici: rapporto conclusivo*, approvato dallo Standing Committe della IFLA Section on Cataloguing, edizione italiana a cura dell'Istituto centrale per il catalogo unico delle biblioteche italiane e per le informazioni bibliografiche, Roma: ICCU, 2000.

OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting)

Fondamentale poi, per la fruizione degli oggetti digitali, è la possibilità di poterli acquisire, conservare e utilizzare in maniera sistematica e coerente. Scandagliare la rete, in maniera casuale oppure accedendo direttamente a siti o *repositories* che si ritengono in qualche modo rilevanti, acquisendonde i metadati, si chiama, secondo uno dei tanti neologismi tecnici inglesi difficilmente traducibili, effettuare l'*harvesting* dei metadati. Una possibile traduzione è: "affastellare metadati".
L'OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) si è proposto da subito come protocollo di riferimento per l'*harvesting* dei metadati. È un progetto che si avvale di XML su HTTP. La versione più aggiornata è la 2.0 del 2002. È prodotto dalla Open Archives Iniziative (OAI), una organizzazione fondata nel 1999, insieme ad altri, da Herbert Van de Sompel, ora ricercatore presso i Los Alamos National Laboratori, ma più noto come ideatore del protocollo OpenUrl, le cui basi vennero gettate negli anni '80 quando lavorava come capoprogetto dell'automazione della biblioteca dell'Università di Ghent, in Belgio[45].
L'OAI ha come scopo quello di individuare gli strumenti fondamentali per allestire infrastrutture di rete che favoriscano l'interoperabilità, cioè lo scambio di contenuti digitali tra sistemi diversi, con particolare riferimento ai *repositories* istituzionali. Sostanzialmente il protocollo OAI-PMH definisce una infrastruttura autonoma, non dipendente da particolari applicazioni, per l'*harvesting* dei metadati. Il protocollo prevede due categorie di partecipanti alla infrastruttura: i *data providers*, i quali gestiscono sistemi che utilizzano OAI-PMH come un mezzo per "esporre" i metadati e i *service providers*, i quali utilizzano i metadati "affastellati" mediante il protocollo OAI-PMH come base sulla quale costruire servizi a valore-aggiunto. Una delle prescrizioni del protocollo è che i metadati siano conformi allo standard di metadati Dublin Core, già citato.
OAI-PMH distingue fra tre diverse entità fondamentali correlate ai metadati che vengono resi accessibili[46]:

- La risorsa. È ciò di cui trattano i metadati. Questioni come la natura di una risorsa, se fisica o digitale, oppure se essa sia archiviata nel *repository* o in un altro database non rientrano nell'ambito del protocollo OAI-PMH.
- Il record. I metadati secondo uno specifico formato di metadati costituiscono un record; in risposta a una specifica richiesta, viene inviato un record sottoforma di un flusso di byte (*byte stream*) codificato in XML.

[45] Open Archives Initiative, http://www.openarchives.org.
[46] Carl Lagoze – Michael Nelson – Herbert Van de Sompel – Simeon Warner, *The Open Archives Initiative Protocol for Metadata Harvesting*, versione 2.0 del 14 giugno 2002, documento versione 2004/10/12T15:31:00Z, http://www.openarchives.org/OAI/openarchivesprotocol.html#DefinionsConcepts.

– L'*item*. Un *item* è quell'elemento di un *repository* a partire dal quale i metadati possono essere disseminati. Concettualmente un *item* è un contenitore nel quale vengono archiviati o che genera dinamicamente, in molteplici formati, i metadati concernenti una singola risorsa. Ogni *item* ha un identificatore che, nell'ambito del *repository*, è unico.

A queste tre componenti di base si aggiunge:

– L'identificatore unico. Un identificatore unico identifica senza ambiguità un *item* dentro un *repository*. L'identificatore unico viene utilizzato nelle richieste OAI-PMH per estrarre i metadati dall'*item*. Tutti i *record* derivabili da un singolo *item* condividono il medesimo identificatore. Il formato dell'identificatore deve conformarsi alla sintassi URI.

Nel 2006 è stato avviato il progetto di un nuovo standard denominato OAI-ORE (Open Archives Initiative Protocol – Object Exchange and Reuse). Questo nuovo standard dovrà sviluppare specificazioni che permettano ai *repositories* non solo di scambiare i metadati ma anche informazioni relative agli oggetti digitali veri e propri ospitati al loro interno. Queste specificazioni, leggiamo nel sito ufficiale di questa iniziativa, includeranno indicazioni su come rappresentare gli oggetti digitali e i servizi connessi, in modo da facilitare l'accesso e l'acquisizione di queste rappresentazioni. Ci si aspetta che queste rappresentazioni rendano possibili una nuova generazione di servizi in grado di trasferire il valore intrinseco degli oggetti digitali anche al di là dei confini dei *repositories* nei quali sono collocati[47].

Conclusione

Come detto all'inizio, non si è trattato di un'analisi esaustiva. L'indagine è rimasta confinata principalmente alla letteratura professionale e ai progetti di area anglosassone, dove abbondano gli studi nel campo delle tecnologie dell'informazione, con particolare riferimento all'interoperabilità e alla *preservation*. Ma, anche in un campo d'indagine così delimitato, procedere è reso faticoso dalla grande quantità di progetti, proposte e iniziative attualmente in corso. Lo stato dell'arte oggi è tale che possiamo parlare discorsivamente dei "documenti digitali" come se fossero oggetti ben conosciuti, ma appena tentiamo di darne una definizione rigorosa, restiamo letteralmente senza parole. L'impulso spontaneo, nel tentare di definire questo concetto, è quello di equiparare i documenti digitali ai file di un computer. Questo induce ad assimilare il documento digitale al documento tradizionale. L'unitarietà del volume cartaceo può essere sostituita dall'unitarietà del file informatico: questa, in fondo, è l'aspettativa implicita e spontanea che contagia chi af-

[47] Open Archives Initiative – Object Reuse and Exchange, http://www.openarchives.org/ore.

fronta per la prima volta questi problemi. Ma, come abbiamo visto, così non può essere. Un singolo documento digitale che sia stato strutturato in maniera adeguata ai fini dell'interoperabilità o della *preservation* è composto in realtà da una molteplicità di file che devono assolvere molteplici compiti. In primo luogo vi sono i file di contenuto e i file di metadati. Ma anche i file di contenuto, a loro volta, possono essere molteplici: ad esempio testo e immagini, di uno stesso documento digitale, possono essere contenuti in file separati. Vi sono poi metadati per i contenuti e metadati per gli aspetti tecnici dei file (formati, ecc.). La materia è resa ancora più complessa dal fatto che le esigenze dell'interoperabilità e della *preservation* impongono l'adozione di standard comuni, oggi ancora in fase embrionale.
Eppure la propensione a considerare il documento digitale come sostanzialmente analogo al documento tradizionale, persiste tenacemente. Probabilmente, contribuisce a ciò lo sforzo messo in opera dai produttori – di per sé lodevole – di rendere quanto più possibile "amichevoli" le tecnologie dell'informazione. Vi sono software, ad esempio, che consentono di "sfogliare" un libro digitale, su un terminale, in maniera analoga ad un libro cartaceo. Ma questi "effetti speciali", sebbene da un lato facilitino l'interazione con i documenti digitali, dall'altro occultano, in particolare ai profani, la complessità che giace dietro ciò che si vede sul monitor. In effetti un cosiddetto "supporto informatico" che sia stato strutturato per affrontare efficacemente problemi di interoperabilità o di *preservation*, costituisce un prodotto tecnologico estremamente complesso. Gestire e controllare questa complessità è un problema che deve ancora trovare una risposta conclusiva. Per cui, a dispetto di tutte le definizioni finora viste, sembra proprio che, riguardo ai cosiddetti "documenti digitali", si possa affermare la stessa cosa che, più di quindici secoli or sono, Sant'Agostino diceva a proposito dell'oscura ed enigmatica natura del tempo: «Che cos'è il tempo? Se nessuno me lo chiede, lo so; se voglio spiegarlo a chi me lo chiede, non lo so più!»[48].
"Documento digitale", concetto in evoluzione: per quanto tempo ancora?

> *What constitutes a "digital record"? In order to answer such a question, it is first necessary to somehow define what a 'digital record' in itself is. This, however, is a changing notion that tends to be adjusted so as to accommodate the inputs arriving from the interoperability and digital preservation sectors.*
> *The current state of the art is such that we speak today of "digital records" as if they were well-known objects – yet as soon as we attempt to formulate a rigorous definition of what exactly they are, we literary remain speechless. The natural instinct, when trying to define the concept, is that of comparing a digital record to a computer file: this way the digital record becomes similar, in terms of units, to the traditional one. Yet in reality a single digital record, if adequate-*

[48] Sant'Agostino, *Confessioni*, libro XI, cap. XIV cit. in: Max Jammer, *Tempo*, in: *Enciclopedia Europea*, vol. 11, Milano: Garzanti, 1980, p. 147 («Quid est tempus? Si nemo a me quaerat, scio; si quaerenti explicare velim, nescio!»).

ly structured so as to satisfy a number of requirements connected to interoperability and preservation needs, is composed of multiple distinct files which serve multiple distinct purposes.
First of all, there are content files and metadata files. Secondly, the content files can themselves be multiple – as in the case of a digital record containing both text and images, which may be stored in separate files. And then there are the metadata dealing with content and those dealing with the technical aspects of the files (formats and so forth). The problem is hence clearly one of remarkable complexity, yet the tendency strenuously persists to consider a digital record as something basically equivalent to a traditional record. The praiseworthy efforts to make information technology as "friendly" as possible likely play a role in such an oversimplification.
What follows is a survey of the intricate jungle of standards and ongoing projects on digitisation issues. Such a survey does not have the ambition of being systematic and comprehensive, yet it is sufficient to gain the impression that the concept of digital record is far from having been defined for once and for all, but is rather still in evolution.

Qu'est-ce qu'un "document numérique"? La réponse sous-entend à son tour une définition quelconque de "document numérique". Mais ce dernier est un concept qui change selon les sollicitations qui proviennent du domaine de l'interopérabilité entre systèmes et de la digital preservation.
L'état de l'art actuel nous pousse à parler des "documents numériques" comme s'il s'agissait d'objets bien connus, mais dès que l'on essaye d'en donner une définition rigoureuse on ne trouve plus, littéralement, les mots.
L'élan spontané, dans la tentative de définir ce concept, est celui d'assimiler le document numérique à un fichier d'ordinateur. Ce qui rapproche le document numérique, sous l'aspect du caractère unitaire, du document traditionnel. Mais en réalité, un document numérique structuré de façon à respecter les limites liées à l'interopérabilité et à la preservation est composé d'une multiplicité de fichiers distincts qui accomplissent différentes tâches.
En premier lieu, il existe des fichiers de contenu et des fichiers de métadonnées. Mais les fichiers de contenu peuvent être variés eux-mêmes: par exemple le texte et les images du même document numérique peuvent être contenus dans des fichiers séparés. Il existe aussi des métadonnées pour les contenus et des métadonnées pour les aspects techniques du fichier (formats, etc.). La complexité, comme on le voit, est grande. La tendance à considérer le document numérique comme un analogue du document traditionnel persiste cependant. Probablement, l'effort, louable, de rendre les technologies de l'information le plus "amicales" possible y contribue.
Ce qui suit est une étude non systématique et non exhaustive dans la confusion des standards et des projets actuels concernant la numérisation. Elle suffit cependant à donner l'impression que le concept de document numérique est loin d'être cerné définitivement. Il s'agit d'un concept qui est encore en train d'évoluer.

RIFERIMENTI BIBLIOGRAFICI

Commissione RICA. *Regole italiane di catalogazione (REICA): bozza complessiva: appendice D – Designazioni specifiche del materiale*. Aggiornamento 10 febbraio 2008. Roma: ICCU, 2008, p. 352, http://www.iccu.sbn.it/upload/documenti/ReicaFeb2008.pdf?l=it.

Judith Pearce – David Pearson – Megan Williams – Scott Yeadon. *The Australian METS Profile: A Journey about Metadata*. «D-Lib Magazine». 14 (2008) n. 3/4, doi:10.1045/march2008-pearce, http://www.dlib.org/dlib/march08/pearce/03pearce.html.

Sally Hubbard. *Getting to the web*. «Dig*Italia*», 2007, n. 2, p. 11-19.

California Digital Library. *Digital Objects Glossary*, a cura di Jennifer Colvin. Aggiornamento 15 maggio 2006, http://www.cdlib.org/inside/diglib/glossary/?field=glossary&action=search&query=oac.

Hans-Werner Hilse – Jochen Kothe. *Implementing Persistent Identifiers: Overview of concepts, guidelines and recommendations*. London: Consortium of European research Libraries; Amsterdam: European Commission on Preservation and Access, 2006, http://www.cerl.org/, http://www.knaw.nl/ecpa/.

Robert Kahn – Robert Wilensky. *A framework for distributed digital object services*. 13 maggio 1995, http://www.cnri.reston.va.us/k-w.html, doi:cnri.dlib/tn95-01; 13 marzo 2006, http://www.doi.org/topics/2006_05_02_Kahn_Framework.pdf, doi:10.1007/s00799-005-0128-x; «International Journal on Digital Libraries», 6 (2006), n. 2, p. 115-123.

Comitato MAG. *Metadati amministrativi e gestionali: manuale utente*, a cura di Elena Pierazzo, versione 2.0.1, marzo 2006. Roma: ICCU, 2006, p. 7, http://www.iccu.sbn.it/upload/documenti/Manuale.pdf.

Tim Berners-Lee – Roy Fielding – Larry Masinter. *RFC 3986: Uniform Resource Identifiers (URI): Generic Syntax*. Gennaio 2005, http://gbiv.com/protocols/uri/rfc/rfc3986.html.

Preservation Metadata: Implementation Strategies. *Data Dictionary for Preservation Metadati*. maggio 2005, p. 7, http://www.oclc.org/research/projects/pmwg/premis-final.pdf.

Dublin Core Metadata Iniziative. *DCMI Glossary*. Aggiornamento 23 aprile 2004, http://dublincore.org/documents/usageguide/glossary.shtml.

Carl Lagoze – Michael Nelson – Herbert Van de Sompel – Simeon Warner. *The Open Archives Initiative Protocol for Metadata Harvesting*, versione 2.0 del 14-06-2002, documento versione 2004/10/12T15:31:00Z, http://www.openarchives.org/OAI/openarchivesprotocol.html.

Shelley Powers. *Practical RDF*. Beijing; Sebastopol: O'Reilly, 2003.

Allen Renear – David Dubin. *Towards Identity Conditions for Digital Documents*. 2003, http://www.dublincore.go.kr/dcpapers/pdf/2003/503_Paper71.pdf.

Consultative Committee for Space Data Systems. *Reference Model for an Open Archival Information System (OAIS): Recommendation for Space Data System Standards (Blue Book), CCSDS 650.0-B-1*. 2002, http://public.ccsds.org/publications/archive/650x0b1.pdf. Traduzione italiana: *OAIS: sistema informativo aperto per l'archiviazione*, a cura di Giovanni Michetti. Roma: ICCU, 2007.

Tim Berners-Lee – James Hendler – Ora Lassila. *Il Web semantico: quando Internet diventa intelligente: agenti software e rappresentazioni condivise per un'automazione dei servizi in rete… e a casa vostra*. «Le Scienze», 2001, n. 393, p. 77-84.

ISBD(ER): International Standard Bibliographic Description for Electronic Resources, edizione italiana a cura dell'Istituto centrale per il catalogo unico delle biblioteche italiane e per le informazioni bibliografiche. Roma: ICCU, 2000.

Requisiti funzionali per record bibliografici: rapporto conclusivo, approvato dallo Standing Committe della IFLA Section on Cataloguing, edizione italiana a cura dell'Istituto centrale per il catalogo unico delle biblioteche italiane e per le informazioni bibliografiche. Roma: ICCU, 2000.

Mario Sebastiani. *Identificatori persistenti per gli oggetti digitali*. «Dig*Italia*», 2005, n. 0, p. 62-82, http://digitalia.sbn.it/upload/documenti/digitalia20050_SEBASTIANI.pdf.
Tim Berners-Lee – Roy Fielding – Larry Masinter. *RFC 2396: Uniform Resource Identifiers (URI): Generic Syntax*. Agosto 1998, http://gbiv.com/protocols/uri/rfc/rfc2396.html.
David W. Chadwick. *Understanding X.500 – The Directory*. 1994, http://sec.cs.kent.ac.uk/x500book/.
Willard Van Orman Quine. *La relatività ontologica e altri saggi*. Roma: Armando, 1986, (Tit. orig. *Ontological Relativity and other Essays*. New York: Columbia, 1969).

Siti Web

CNRI, http://www.cnri.reston.va.us/about_cnri.html.
D-Lib Alliance, http://www.dlib.org/dlib/alliance-participants.html.
Defense Virtual Library, http://www.dtic.mil/dtic/prodsrvc/rdi/dvl.html.
DOI, http://www.doi.org/.
DSpace, http://www.DSpace.org/.
DTIC Handle Service, http://www.dtic.mil/dtic/handles/index.html.
DTIC OAI-PMH service, http://www.dtic.mil/dtic/prodsrvc/oai.html.
Dublin Core Metadata Iniziative, http://dublincore.org/.
METS, http://www.loc.gov/standards/mets/mets-home.html; http://www.loc.gov/standards/mets/METSita.html.
OAI-ORE, http://www.openarchives.org/ore/.
OAI-PMH, http://www.openarchives.org/.