

14. Seminario Angela Vinay.
bibliotECONOMIA: la frontiera digitale, 4-5 aprile 2003.

Vademecum on-the-fly **per progetti di digitalizzazione**

Venezia, 5 aprile 2003

Antonella De Robbio
Centro di Ateneo per le Biblioteche
Università di Padova

Abstract

Il presente intervento, strutturato in forma di semplice *vademecum on-the-fly* per progetti di digitalizzazione, si basa sulla comparazione di tre progetti: il piano di digitalizzazione della BEIC; Biblioteca Europea di Informazione e Cultura, il modulo copyright della Biblioteca Digitale Italiana e la proposta per il progetto europeo UE DML Digital Math Library, presentato entro il 6. Framework Program per la costruzione di una biblioteca digitale mondiale per la matematica.

Vuole essere un semplice strumento per un approccio “al volo” semplice, ma progettuale, sulla questione della digitalizzazione.

I contenuti dell'intervento sono:

1. Tre esperienze a confronto per tre progetti diversi
2. Issues comuni emersi dal benchmarking tra progetti: i punti chiave da considerare
3. Strategie per progetti di digitalizzazione locali, nazionali ed europei
4. Conclusioni

Premessa

In occasione di questo mio intervento alla tavola rotonda del 14.mo Seminario Angela Vinay mi è riemerso dalla memoria, quasi fosse rimasto nascosto forzatamente per venticinque anni, il ricordo di Angela Maria Vinay Pietra che conobbi quando ero appena ventenne e bibliotecaria di fresca nomina. Era l'aprile del 1978 e mi trovavo ospite a Roma da un'amica comune dell'età di Angela Vinay, anche lei scomparsa oramai da qualche anno. Angela, che era stata da poco nominata direttrice dell'Istituto centrale per il catalogo unico delle biblioteche italiane e per le informazioni bibliografiche, mi propose una visita all'allora modernissima biblioteca nazionale centrale Vittorio Emanuele II di Roma e mi condusse personalmente tra labirinti e sale di lettura, illustrandomi l'organizzazione e le fasi di lavoro. Il suo entusiasmo era trascinate e avvertivo ancora vivo il timbro della sua voce nel descrivere i punti di innovazione della biblioteca e i vari problemi organizzativi che all'epoca avevano incontrato. Il nostro fu un incontro breve, ma intenso. Non la incontrai mai più nel corso della mia carriera. Ringrazio Chiara Rabitti per avermi dato l'opportunità di partecipare al Seminario in onore di Angela Vinay e di poter ricordare, così, a mio modo, questa nostra grande maestra.

1. Tre esperienze a confronto

Negli ultimi due anni sono stata coinvolta in tre distinti progetti di digitalizzazione. Il presente lavoro si basa su queste distinte esperienze e tenta di mettere a confronto i tre modelli di riferimento, individuando le comunanze e i punti chiave presenti in ciascuno di essi.

I tre progetti a cui mi sono dedicata sono stati i seguenti:

1. Sviluppo del Terzo Modulo “*Digitalizzazione e Copyright*” dello Studio di fattibilità della BEIC, nell’aprile 2002 [1]. Oltre al coinvolgimento come responsabile del piano di digitalizzazione ho partecipato anche allo sviluppo del Progetto delle collezioni BEIC coordinato da Giovanni Solimine.
2. Stesura del Modulo “*Copyright*” per la Biblioteca Digitale Italiana in qualità di consulente nel marzo 2003.
3. Coinvolgimento (gennaio-aprile 2003) nella predisposizione della proposta di Progetto UE DML Digital Math Library [2] per la costruzione di una biblioteca matematica digitale europea, e in asse coordinato con il progetto internazionale, nell’ambito del Sesto programma quadro dell’Unione Europea, con la partecipazione del CAB Centro di Ateneo per le Biblioteche dell’Università di Padova tra i partner europei del progetto.

1.1. Il piano di digitalizzazione della BEIC.

Lo studio di fattibilità per il piano di digitalizzazione della BEIC presenta un business plan di cinque anni. Il piano descrive strumenti e servizi necessari alla digitalizzazione delle collezioni di documenti della BEIC su supporto analogico ed è allineato con la mission della BEIC stessa. Lo scopo principale del piano è quello di rendere accessibile attraverso la conversione in un formato digitale un "nucleo centrale" del patrimonio fisico della BEIC, ancora non costituito e che si basa sul modulo di sviluppo delle collezioni coordinato da Giovanni Solimine. Lo scopo è lo sviluppo di un insieme di collezioni digitali da costruirsi attorno al nucleo centrale in collaborazione con le istituzioni del territorio e in asse cooperativo con le altre iniziative italiane nascenti e con le biblioteche digitali europee già esistenti. Il piano sarà composto da progetti compatibili con gli obiettivi del piano, i quali avranno durata minima di diciotto mesi, massima di trentasei mesi e prevede dei criteri, correlati agli insiemi di documenti da digitalizzare, che ogni progetto dovrà osservare nell’ambito di una pianificazione generale.

I criteri di progetto, in linea con quelli del piano, si riferiscono ad un livello inferiore e prevedono una "selezione" delle opere da trattare sulla base di "linee guida" stabilite.

E’ compreso nello studio anche il modulo sul copyright il quale suggerisce fin dove possibile, l’utilizzo delle speciali clausole che riguardano le eccezioni alla riproduzione di materiale soggetto a tutela.

1.2. Il modulo copyright della Biblioteca Digitale Italiana

Il modulo sul copyright per la BDI Biblioteca Digitale Italiana è focalizzato sui principi normativi fondamentali che regolano la proprietà intellettuale con riferimento particolare all’ambiente delle biblioteche digitali/ibride.

Con l’aumentare delle diverse tipologie di materiali all’interno delle raccolte che costituiranno la BDI e con le regole specifiche da applicare ad ogni medium, la BDI dovrà assicurare che ogni procedura sia posta in atto in modo da trattare correttamente la complessa materia del copyright, per esempio quando si acquisisce nuovo materiale, quando si digitalizza quello esistente, quando lo si riproduce per l’utenza.

L'applicazione di griglie, che aiutino nella regolamentazione della proprietà intellettuale in ambiente multimediale, deve tener conto di fattori connessi alla natura stessa dei documenti testuali, sonori, visivi o multimediali, ma anche all'ambito normativo cui necessariamente ci si riferisce.

Il modulo tiene conto che esistono ad oggi differenti corpi normativi che regolano il mondo della proprietà intellettuale prevalentemente riferibili al contesto analogico. Il concetto di proprietà intellettuale, riconducibile ai vari sistemi normativi (copyright o diritto d'autore o sistemi ancora differenti) deve necessariamente estendersi oltre i confini di ogni singolo stato e includere nelle sue possibili eccezioni alle tutele o limitazioni dei diritti, ogni entità connessa, compresi tutti i cittadini.

Tutelare i diritti di proprietà intellettuale deve essere visto come fattore imprescindibile, parimenti deve però essere trovata qualsiasi via alternativa al fine di garantire:

- un libero accesso ai contenuti digitalizzati, per quanto possibile, da parte dell'utenza
- la possibilità di duplicare o effettuare copie degli oggetti digitalizzati nel rispetto dei diritti, nei casi in cui il materiale non sia di pubblico dominio
- la possibilità di provvedere ad una disseminazione del materiale scientifico di ricerca e per la didattica, in un'ottica quanto più possibile vicina al concetto del FairUse statunitense.

1.3. Il progetto DML Digital Math Library

Il progetto mondiale DML Digital Mathematical Library si pone l'obiettivo di rendere disponibili e liberamente accessibili, in formato digitale, la totalità dei documenti di matematica, stimati in circa 50 milioni di pagine. La DML comprenderà sia collezioni di documenti retro-digitalizzati chiamate *past literature* sia collezioni di documenti in formato nativo digitale. Numerosi saranno i soggetti coinvolti, tra cui biblioteche, università, editori, progetti e le società matematiche nazionali e internazionali. Il tempo previsto per la realizzazione del progetto è di dieci anni, il budget totale di 100 milioni di dollari. Il progetto mira anche a creare un coordinamento tra le esperienze di digitalizzazione già esistenti, attraverso una rete che poggia su un modello distribuito.

I tre scopi primari di questo progetto mondiale sono:

1. digitalizzare una parte preponderante della letteratura matematica scientifica che non esiste già in formato digitale e che viene definita "past literature", la quale è collocata a livello temporale prevalentemente prima del 1997
2. predisporre un insieme di standard tecnici per rendere i documenti di matematica digitalizzati accessibili online a chiunque
3. negoziare un protocollo per rendere la letteratura matematica digitale disponibile anche in futuro

Ciò significa servire le comunità scientifiche, gli studenti e, più generalmente, tutti i cittadini del mondo, offrendo un accesso facilitato ed efficace ad un largo corpus di documenti matematici, ma in particolare all'importante patrimonio del materiale a stampa. I dati grezzi entro la DML dovrebbero essere liberamente disponibili a tutti via Internet, regole del copyright permettendo.

Numerosi tuttavia sono i nodi da sciogliere tra cui il copyright, i formati, l'archiviazione, gli standard da adottare. Molti sono gli aspetti su cui si dovrà decidere, tra cui la selezione dei contenuti, gli standard tecnici e i formati di metadata.

Accanto al progetto mondiale che nasce alla Cornell University e che è finanziato dalla National Foundation of Science, l'European Mathematical Society EMS si è attivata per presentare, nell'ambito del Sesto Programma Quadro per i finanziamenti alla ricerca europea, una proposta di progetto DML europeo. Il Centro di Ateneo per le Biblioteche dell'Università di Padova è uno dei partner coinvolti nel progetto DML europeo, ora ancora allo stato di proposta per il finanziamento entro il Sesto Programma Quadro europeo. Un altro partner italiano è l'Unione Matematica Italiana con la quale il nostro ateneo di Padova collabora attivamente attraverso la Biblioteca del Seminario Matematico. In

seno all'UMI sta decollando il progetto italiano ItalDML che vedrà la partecipazione delle biblioteche di matematica italiane e che confluirà entro il progetto europeo.

2. Issues comuni ai tre progetti

La mia esperienza nei tre progetti si è differenziata soprattutto in relazione al tipo di apporto dato al progetto, alla materia da me trattata e sviluppata per ciascuno dei progetti, e al contesto di riferimento. Ogni biblioteca digitale, al pari di ogni biblioteca tradizionale, ha le sue peculiarità, le sue caratteristiche, il suo modello economico, la sua struttura architettonica, le sue regole, le proprie offerte di servizi, una propria utenza fidelizzata... Ma come in ogni biblioteca tradizionale vi sono principi e regole che tracciano, come un filo rosso conduttore, le comunanze, e che tentano di condurre ad approcci metodologici standard, nel rispetto del concetto di "interoperabilità" tra sistemi di cui tanto si parla oggi. Utilizzo il termine "sistema" proprio per identificare la biblioteca come un network polifunzionale, come una rete comunicante attraverso l'interoperabilità intesa come scambio di beni "comuni".

Tra le comunanze individuate tra progetti di digitalizzazione diversi, emergono con prepotenza alcuni issues comuni che potremmo definire i "punti chiave" di ciascun progetto di digitalizzazione.

Semplificando, potremmo individuarne sette, sebbene sicuramente ve ne siano altri la cui trattazione richiederebbe uno spazio assai più ampio rispetto al mio intervento odierno.

I **sette punti chiave** che ho individuato come comuni sono:

1. L'impostazione del piano di digitalizzazione
2. L'approccio metodologico
3. Gli strumenti
4. La costruzione del registro
5. I progetti entro il piano
6. Il piano finanziario e le risorse
7. I costi

2.1. Impostazione di un piano di digitalizzazione

La descrizione del piano deve prevedere alcuni issues fondamentali tra i quali: mission, scopi e obiettivi, durata (che deve essere necessariamente per un piano a medio-lungo termine).

Deve inoltre essere analizzato con attenzione il background. Gli elementi di contesto sono fondamentali, particolare attenzione andrà posta all'ambiente di riferimento ove dovrà agire il servizio di digitalizzazione. E' essenziale stabilire a priori o quanto meno avere un'idea di quanto materiale delle collezioni sarà possibile digitalizzare, delle tipologie delle collezioni e della tipologia di materiale, qualità dei supporti, digitalizzazione massiccia o selettiva della collezione individuata ...

Tutto ciò è strettamente correlato ai vincoli: di contesto, di budget, normativi, interni o esterni.

Nel piano andrà individuato il modello economico dove dovranno essere considerati i presupposti base di scenario, gli attori in gioco o *players*, la struttura generale e i finanziamenti. Il budget è il riferimento chiave, senza un budget stabilito non è possibile far partire nessun progetto di digitalizzazione.

Al fine di salvaguardare la letteratura detta di *legacy* sarà importante creare e pianificare una struttura che possa attrarre i fornitori e possessori di contenuto che detengono i diritti o che producono e

distribuiscono il materiale oggetto delle collezioni. Questi attori possono essere società, organizzazioni o editori no-profit e editori commerciali.

Va considerato che la maggioranza degli editori è piuttosto scettica nel trasferimento della proprietà dei materiali o nell'assegnazione del copyright elettronico a favore delle biblioteche digitali. Inoltre gli editori commerciali preferirebbero ospitare i contenuti sui propri siti di modo da integrare questi materiali con i contenuti digitali esistenti, per aumentare la loro visibilità in termini di business.

Non tutti gli editori commerciali sono disposti inoltre a digitalizzare materiale considerato letteratura di *legacy* in quanto non si assumono i costi che possono anche essere rilevanti a fronte di scarsi profitti. Sarà necessario quindi trovare finanziamenti adeguati per la digitalizzazione di questo patrimonio storico.

Le biblioteche possono essere l'agente naturale per archiviare questi materiali, tenendo conto delle opportune questioni connesse all'archiviazione e alla conservazione del digitale e nel rispetto di standard e protocolli (per esempio OAIS).

2.2. Approccio metodologico

La metodologia consente di delineare delle strategie mirate alla creazione di strumenti per l'attuazione di "servizi", per esempio per la creazione dello stesso modello di business. La cooperazione con gli editori per i testi già digitalizzati dai produttori di contenuto è un fattore strategico come pure il coordinamento con gli altri progetti di digitalizzazione presenti nel territorio sia a livello locale, sia nazionale, e sia in relazione alle grandi esperienze europee e internazionali.

L'approccio metodologico dovrà consentire al progetto di delineare delle strategie previe le quali condurranno alla creazione di strumenti per l'attuazione di "servizi" applicabili alla biblioteca digitale.

La biblioteca digitale è prima di tutto un modello economico e la sua fruibilità deve basarsi su un consenso allargato da parte dell'utenza che vi accede.

2.3. Gli strumenti

Alcuni strumenti dovrebbero essere sviluppati a livello nazionale e mi riferisco per esempio a quanto proposto dal progetto europeo MINERVA o dagli stessi sforzi intrapresi dalla BDI e dovrebbero essere messi a disposizione di qualsiasi progetto di digitalizzazione nazionale.

Mi riferisco in particolare alla creazione di una banca dati dei progetti e delle biblioteche digitali e alla creazione di un MetaOpac per la ricerca nelle biblioteche digitali già esistenti.

Adeguati strumenti messi a disposizione a livello nazionale potrebbero aiutare nell'evitare duplicazioni di lavoro, sovrapposizioni e digitalizzazioni fuori standard. E' da considerare che solo con adeguati strumenti è possibile attuare un'attività di benchmarking tra progetti ed esperienze al fine di ricavare dati su: analisi degli altri progetti esistenti, benchmarking su standard tecnici, buone prassi, qualità dei metadati, qualità dei dati, modelli di costo, ...

Di essenziale utilità è la creazione nazionale di un registro delle opere digitalizzate e da digitalizzare.

2.4. Creazione del registro digitale

L'impostazione di un registro delle opere digitali è lo strumento centrale per le attività di ogni progetto di digitalizzazione. In DML si è pensato di strutturare il registro utilizzando l'architettura di OAI Open

Archives Iniziative, suddivisa in data provider e service provider. Questo perché il registro dovrà essere strutturato sulla base di metadati standard nell'ottica dell'interoperabilità tra archivi.

Usare metadati significa preparare le basi per un recupero efficace dei testi digitalizzati da parte dell'utenza. Significa anche essere in grado di sapere se quel dato volume è stato digitalizzato da altri oppure se il "master" esistente è di qualità inferiore agli standard che la BEIC si è prefissata. Coi metadati si creerà il registro comune delle opere digitalizzate, di modo da condividere gli sforzi nell'ottica della cooperazione.

Questo significa aggiungere valore ad un lavoro di per sé assai costoso e quindi rendere visibile ciò su cui si è investito. Conviene quindi investire un po' di più nel versante dei metadati e dell'indicizzazione degli oggetti trattati, al fine di offrire un modello interattivo di assistenza all'utenza che possa fornire strumenti e metodi per raggiungere i contenuti digitalizzati in modo efficace.

Attuare una politica di economia di scala in merito alla digitalizzazione dei documenti, attraverso la tenuta di un "registro delle opere digitalizzate" avrà lo scopo di razionalizzazione costi e dei tempi.

Il registro è uno strumento di lavoro per lo staff che fornisce informazioni sull'accesso e disponibilità e dovrà avere connessioni entro un quadro locale, nazionale, europeo, internazionale.

2.5. I progetti entro il piano

Ogni progetto dovrà rispondere ai criteri e ai principi del piano in cui esso è incardinato, ma dovrà anche sottostare alla carta dei criteri del progetto. I criteri di progetto sono fondamentali e si basano su un'attività di pianificazione. La pianificazione di ciascun progetto è ciò che rende stabile il piano nel suo insieme. Questo è maggiormente valido in progetti di piccola dimensione i quali necessariamente dovrebbero essere parte di un piano più ampio entro un quadro di biblioteca digitale che prevede oltre al servizio di digitalizzazione, anche altri servizi tra cui costruzione di archivi aperti per il deposito dei materiali di ricerca o per la didattica o di materiale di altro tipo, accesso ai periodici elettronici, accesso alle banche dati, documenti di archivio...

Oltre ai criteri di progetto dovranno essere previste le linee guida di progetto dove vengono raccolte le cose utili, ciò che è bene prevedere, ciò che va assolutamente evitato, consigli su processi e metodologie, come conviene procedere ...

2.6. Il piano finanziario e le risorse

Il piano finanziario è in stretta correlazione con modello di business. Contiene i dati utili ai fini dei costi. Il modello di business delinea le attività di profitto per l'offerta dei servizi.

Le azioni del piano finanziario sono le seguenti:

- Individuazione del budget a disposizione
- Vincoli di budget
- Analisi delle categorie di costo e stime per la digitalizzazione (es. per DML sono stati previsti 2 Euro a pagina)
- Analisi dei requisiti e dei costi per Hardware e Software
- Analisi dei costi per il personale
- Correlazione con il modello di business

Nell'analisi dei servizi da offrire all'utenza, sarà opportuno individuare ciò che può rientrare nel modello di business in quanto "conveniente", per esempio un servizio di Print-On-Demand per alcune collezioni ad alta frequenza di consultazione.

2.7. I costi

Le stime del costo di attuazione di queste fasi in un'operazione a grande scala sono variabili, ma per il progetto DML con una grossa approssimazione si è stabilito un costo di \$ 2 per pagina.

Ci sono molti modi di digitalizzare la *letteratura passata* (cioè, letteratura che è già in forma digitale) ma l'unico modo conveniente in termini di costi/benefici è combinare la scansione con una parziale lettura ottica dei caratteri, creando una combinazione di immagini di pagine scandite e file di testo associati, che rendano possibile la ricerca per parola o stringa. Naturalmente il processo risulta costoso. Si devono catturare, di solito battendoli manualmente, i dati bibliografici rilevanti di ciascun documento; i documenti devono essere analizzati per individuarne le varie parti (articoli, capitoli, ecc.); deve essere effettuata un'attenta correzione dei dati critici nei file di testo.

In molti progetti i costi di conversione rappresentano soltanto un terzo dei costi dell'intero progetto.

La "cattura" degli oggetti rappresenta circa il cinquanta per cento delle spese di digitalizzazione.

Il fattore indicizzazione incide fortemente sui costi, come pure il tipo di formato utilizzato per la copia master digitale, la risoluzione delle immagini, l'utilizzo di software OCR per il riconoscimento dei caratteri. Vi sono inoltre "dettagli" che fanno lievitare il costo reale di tre o quattro volte quello previsto: l'assemblaggio del materiale d'origine, il copyright, il settaggio delle macchine, il controllo della qualità del prodotto, la fase post-produzione, la catalogazione con i metadati, la resa finale del lavoro, la gestione del progetto complessivo

I costi devono essere suddivisi in categorie.: hardware, software, personale ...

Le spese per le attrezzature, le macchine e gli strumenti hardware, le spese generali per attuare il piano, quelle per l'implementazione effettiva devono essere minuziosamente descritte, tenendo conto del deprezzamento delle macchine e dei costi di manutenzione. Sarà utile corredare il progetto di tabelle relative a modelli stilati sulla base del costo della singola pagina o immagine, per il costo del personale, ecc...

3. Strategie per progetti di digitalizzazione locali, nazionali ed europei

Accanto ai sette issues comuni vi sono altrettante sette aree di intervento in cui applicare strategie su cui effettuare un confronto serrato con i progetti di digitalizzazione locali, nazionali ed europei già avviati o in corso d'opera.

Le **strategie si riferiscono alle seguenti tematiche:**

1. Gestione del progetto
2. Archiviazione ed accessibilità a lungo termine
3. Individuazione e definizione di standard comuni
4. I formati
5. Copyright
6. Infrastruttura di linking
7. Attività di ricerca

3.1. La gestione del progetto

La gestione del progetto richiederà un corpo decisionale, delle commissioni per la selezione delle collezioni e delle opere da digitalizzare e dei gruppi di lavoro.

In particolare il corpo decisionale si occuperà del coordinamento con altri partners per una sostenibilità a lungo termine, delle questioni di budget, del reperimento delle fonti di finanziamento e dell'integrazione con i progetti nazionali e internazionali.

Attività tipiche sono:

- individuazione del modello economico e gestione del budget
- individuare le realizzazioni di progetti di digitalizzazione di ambito nazionale al fine di evitare duplicazioni, come purtroppo già avvenuto nel passato
- predisporre le linee guida per il progetto di digitalizzazione
- predisporre le politiche di copyright e di IPR (Intellectual Property Rights) che dovranno essere calate nelle realtà nazionali (leggi nazionali)
- stabilire standard e formati di metadati o recepire protocolli, formati e standard già adottati in altri progetti
- individuare i formati per la digitalizzazione e conservazione e quelli di delivery idonei a ciascuna categoria di documenti da trattare
- stabilire gli standard tecnici per la digitalizzazione, le modalità e i protocolli per l'archiviazione e per il recupero dei documenti

Le azioni delle commissioni si svilupperanno attorno a queste linee di pianificazione:

- Obiettivi di programma
- Contenuti
- Scopi
- Selezione dei materiali da digitalizzare
- Priorità per la digitalizzazione
- Specifiche tecniche
- Standard per la digitalizzazione
- Standard di metadati
- Standard per l'interoperabilità
- Standard per il reference linking
- Standard per l'archiviazione
- Copyright e proprietà intellettuale
- Project management
- Costi e modello economico
- Gestione e mantenimento

I gruppi di lavoro dovranno quindi costituirsi attorno a queste linee tematiche: contenuti, standard tecnici, metadati, copyright, archiviazione, modello economico.

3.2. Archiviazione ed accessibilità a lungo termine

Il problema dell'archiviazione non è tanto un problema per il progetto stesso quanto per chi manterrà la collezione dopo la chiusura del progetto. Ancora una volta, è un problema spesso frainteso, anche dagli esperti, esattamente perché non ci sono veri esperti in un'area come l'archiviazione del digitale, dove nessuno ha molta esperienza.

Fino a tempi recenti, non c'era un gran bisogno di archiviare in modo consapevole riviste o libri scientifici, l'archiviazione era quasi automatica perché molte copie erano distribuite nelle biblioteche di tutto il mondo e ubicate in diverse località. Ci si affidava alle leggi della probabilità per assicurarsi che almeno una copia avrebbe continuato a esistere, negli anni a venire. Quell'unica copia avrebbe potuto essere utilizzata come fonte per la riproduzione simultanea di molte copie, anni dopo la pubblicazione iniziale.

Con la pubblicazione elettronica sono cambiate due cose. Primo, le copie possono non essere a larga distribuzione, ma molto spesso risiedono in forma elettronica in uno o due siti. Questo è il problema della "robustezza", ed è l'argomento a cui più si pensa discutendo di archiviazione. Secondo, anche se una copia di un file persiste molti anni nel futuro, può succedere che non sia possibile produrre copie dell'"opera", cioè copie pienamente funzionali che siano identiche a quelle in esistenza anni prima. Questo perché le riviste e i libri elettronici sono spesso costituiti di file inseriti in un sistema più grande che fa uso di programmi, file ausiliari e anche hardware specifico per presentare l'opera. In breve, il contesto in cui l'opera è inserita è spesso essenziale per realizzare una copia fedele, e l'archiviazione richiede la capacità di riprodurre quel contesto. Spesso si parla di questo problema come del problema del "formato", ma tale espressione lo banalizza, riducendolo a un mero problema di presentazione. Di fatto, è il problema centrale non è quello del formato, bensì quello dell'archiviazione.

Per sostenere un progetto di digitalizzazione e fare in modo che esso si mantenga nel tempo, si deve trovare il modo di coprire i costi potenzialmente elevati dell'aggiornamento dei formati nel futuro, così come di prendere decisioni ragionevoli su quali informazioni mandare avanti. Mantenendo collezioni su molti siti, ciascuno con interessi professionali o finanziari al materiale, si crea la sicurezza che un vasto gruppo sarà motivato a condividere tali costi elevati. Sarà nell'interesse di tutti di rendere certo che al cambiamento dei formati vengono prese decisioni affidabili. Nondimeno, questi sono problemi che si estendono per lunghi periodi di tempo, andando spesso oltre la durata della carriera delle persone coinvolte: ci deve essere quindi un meccanismo per garantire che i problemi dell'archiviazione vengono affrontati su una base di continuità.

3.3. Individuazione degli standard

E' fondamentale individuare protocolli standard per l'interoperabilità. Per esempio l'adozione del protocollo OAI-PMH Open Archives Initiative Protocol Metadata Harvesting per quanto concerne lo scambio di metadati. OAI-PMH prevede metadati Dublin Core ai fini dello scambio, ma supporta numerosi altri formati di metadati molto più ricchi del Dublin Core. Per la digitalizzazione sarà necessario prevedere formati di metadati ricchi, soprattutto per il registro come pure per la conservazione dei metadati stessi e dei contenuti.

Nella digitalizzazione sono numerose le tipologie di metadati che possono essere utilizzate: i metadati descrittivi, gli amministrativi-gestionali detti MAG, i tecnici e di utilizzo, i metadati per la conservazione.

Sarà opportuno riferirsi al protocollo OAIS Open Archival Information Systems orientato oltre che al trattamento dei documenti elettronici, alle finalità della conservazione permanente. indirizzato alle questioni dell'archiviazione per la conservazione a lungo termine. Nel medio termine e verso il lungo termine i gestori degli archivi dei progetti di digitalizzazione possono bene volere l'applicazione dei principi OAIS nelle funzionalità dei loro archivi.

Lo sviluppo di linee guida per gli standard tecnici per la digitalizzazione dei contenuti sarà un altro aspetto fondamentale che comprenderà anche la definizione degli standard di accesso e la distribuzione del materiale digitalizzato.

Alcuni degli item da prevedere dovranno riguardare:

- ❑ la qualità della scansione
- ❑ i formati per l'archiviazione
- ❑ le convenzioni dei nomi dei file
- ❑ i formati di distribuzione
- ❑ le unità download
- ❑ le modalità tecniche sui server

3.4. I formati

Naturalmente, la definizione di standard per contenuti che sono già in forma digitale è un problema ben noto, se non proprio ben compreso. Questo richiederà un duro lavoro e una consistente negoziazione, ma anche il problema apparentemente semplice di decidere il formato del materiale scandito è estremamente difficile. Non molto tempo fa, molti avrebbero indicato di usare qualche forma di file TIFF compresso incapsulato in file Adobe PDF. Ma, sebbene PDF abbia al momento un'ampia disponibilità di strumenti di supporto, per certi sistemi operativi, per esempio Unix, questo supporto è diventato problematico. Ancora più importante, ci sono formati nuovi, estremamente efficienti per le immagini scandite, che riducono la dimensione dei file di un fattore da tre a otto o più. Fra questi, il più rappresentativo è DjVu, un formato sviluppato negli AT&T Labs, che usa la tecnica delle ondine per una maggiore compressione e un algoritmo progressivo per la decompressione delle immagini, il quale presenta immediatamente un'immagine grossolana, in progressivo miglioramento. DjVu è un formato web-centrico e una piattaforma software per distribuire documenti e immagini. DjVu può tranquillamente rimpiazzare i formati PDF, PS, TIFF, JPEG, e GIF nella distribuzione di documenti scanditi, documenti digitali o immagini ad alta risoluzione, utilizzando minori risorse client quindi risultando più veloce e leggero e con un più piacevole layout visivo.

Risulta essere un'alternativa superiore in termini di qualità al PDF e PostScript per i documenti digitali, al TIFF e PDF per i documenti bitonali scanditi, al JPEG per le fotografie e al GIF per le immagini ampie. E' adatto per la distribuzione di documenti a colori scansiti ad alta risoluzione.

I prodotti che implementano DjVu sono attualmente di proprietà e in vendita da Lizard Tech. Come PDF, DjVu richiede un software speciale per vedere le immagini nei browser, ma la tecnologia è open source e ci sono notevoli vantaggi sulla tecnologia più tradizionale.

Selezionare il giusto formato iniziale, eventualmente un formato proprietario, in un ambiente che è in costante cambiamento, per un progetto che dura più di dieci anni, è un lavoro praticamente impossibile. Questo è strettamente connesso al problema dell'archiviazione, ma non è lo stesso: il formato iniziale indicato per la presentazione può non essere quello giusto per l'archiviazione.

3.5. Copyright

La questione copyright è spesso sconosciuta e forse anche sottostimata persino nei progetti di digitalizzazione più importanti, sebbene recentemente si stia incominciando a parlare di diritti di proprietà tra gli addetti ai lavori nelle varie comunità coinvolte in tali progetti. Il copyright è il capitolo centrale di ogni biblioteca digitale e questo non dovrebbe mai essere sottovalutato.

La digitalizzazione dei contenuti e la loro messa in rete per un accesso all'utenza richiederà il permesso dei possessori di copyright per le opere non di dominio pubblico.

In alcuni casi può essere possibile ottenere il permesso attraverso negoziazioni relativamente semplici con i detentori dei diritti. In molti altri casi, tuttavia tali permessi possono essere difficili da ottenere. Tale permesso sarà garantito sotto forma di "licenze" preferibilmente basate su una licenza standard o

su un set di licenze standard preparate prima di fissare l'agenda delle negoziazioni con i detentori del copyright. In tal senso vi sarà molto lavoro da fare da parte delle diverse parti interessate per individuare la via da seguire. Questo è un aspetto fondamentale che deve essere tenuto presente da ogni progetto di digitalizzazione.

Le licenze e le linee guida, concordate tra i rappresentanti di biblioteca ed editori, possono essere usate per definire l'interpretazione dei permessi, o per la loro estensione.

I problemi di autorizzazione del copyright e l'identificazione di copie non autorizzate non devono essere sotto stimati. I costi possono essere molto alti, in particolare dovuti al fatto del tempo richiesto per le indagini e per la richiesta delle autorizzazioni.

I detentori di copyright possono essere moltissimi e aumentano in proporzione al numero e al tipo di lavori soggetti a copyright che vengono incorporati nel servizio di digitalizzazione.

Anche il costo delle trattative con ciascun detentore di diritti è individualmente proibitivo, senza tener conto del tempo che si perderebbe. Ciò che serve è piuttosto un approccio collettivo a favore di un intero settore che rende accettabile l'idea di un network gestito a servizio delle biblioteche in tutto il paese. Si dovrà lavorare molto a livello di negoziazione con le associazioni di categoria. Questo genere di approccio è valido però per grandi progetti nazionali, meno fattibile per piccoli progetti locali.

La cessione o trasferimento dei diritti dall'autore all'editore si attua attraverso dei contratti, ma si deve tenere in considerazione la questione delle diverse normative nazionali, delle leggi comunitarie (per quanto concerne l'Europa) e dei trattati internazionali. In sostanza digitalizzare un periodico comporta chiedere il permesso a migliaia di autori, o loro eredi, trattare con centinaia di editori, molti dei quali non facilmente identificabili o che non sono più presenti nel mercato. In tutto questo le leggi non aiutano la costruzione di Biblioteche Digitali che dovrebbero essere i templi del patrimonio scientifico da offrire alla comunità intera.

Questo aspetto fa aumentare notevolmente i costi amministrativi del progetto, come ben descrive Clifford Lynch nel suo articolo "*Convertire i vecchi volumi in formato digitale*".

Una possibile soluzione al problema del copyright è decidere di includere solo letteratura che è già di dominio pubblico o quella il cui permesso è facilmente ottenibile. Da una stima effettuata a grandi linee risulta però che dei 50 milioni di pagine di matematica da digitalizzare, oltre il 90% è protetto da copyright, di questo una buona metà richiederebbe una ricerca e conseguente negoziazione sulla questione dei diritti.

Ciò che andrebbe fatto è un lavoro a monte, da parte dei grossi progetti, quali la BDI o la BEIC, che dovrebbero avere la funzione di "agenzie nazionali per la contrattazione del copyright per le biblioteche digitali" o anche un supporto legale da parte dei ministeri competenti, come è avvenuto in Francia dove si sono digitalizzate quattro prestigiose riviste di matematica con la copertura del Ministero della Ricerca francese. Per il materiale scientifico accademico, risulta molto più costoso chiedere i permessi e le autorizzazioni a centinaia di migliaia di autori, contattando gli eredi, piuttosto che procedere alla digitalizzazione comunque con il solo accordo dell'editore, contando sul fatto che gli eventuali "autori" che daranno causa considerandosi parte lesa saranno davvero pochi considerando che l'autore scientifico considera la "disseminazione" dei suoi lavori un notevole vantaggio e non certo una violazione ai suoi diritti.

E' quindi necessario che si muovano i governi nel tutelare le biblioteche digitali. I governi dovrebbero comprendere che gli interessi delle comunità scientifiche e della popolazione globale nell'accesso ai contenuti del patrimonio dell'umanità prevalgono sugli interessi dei singoli che, molto spesso, sono piuttosto gli interessi di alcune lobbyies di mercato.

3.6. Infrastruttura di linking

Apparato fondamentale di ogni progetto di digitalizzazione è l'infrastruttura di linking tra gli oggetti digitali entro l'archivio del progetto, ma anche fuori di esso verso archivi entro la biblioteca digitale del sistema o di altre biblioteche digitali o di altri sistemi.

Il linking tra gli oggetti darà luogo alla costruzione di una rete "comunicante" che, se attuata nel rispetto dei protocolli standard di interoperabilità si conetterà facilmente con ambienti diversi, ma compatibili.

Fanno parte di questo settore di indagine queste tre azioni che si configurano come settori di ricerca attualmente in fase di consolidamento:

- Identificazione degli oggetti e delle peculiarità specifiche: ambiti disciplinari, supporti, canali...
- Link bidirezionali da e verso gli oggetti differenti (OpenURL): link reciproci
- Citation linking: estrazione automatica delle citazioni dai materiali

3.7. Attività di ricerca

Impostare progetti di digitalizzazione significa anche rendersi conto che numerosi sono gli aspetti ancora tutti da indagare. Per questa ragione parlare di progetti di digitalizzazione presuppone anche aprire la strada ad attività di ricerca che vadano in direzioni precise, per esempio verso l'estrazione automatica di metadati, la trascrizione automatica di testi attraverso OCR (es. utilissimo per la matematica), il riconoscimento automatico di strutture, il multilinguismo e il multiscript. Un'altra frontiera oggetto di indagini è la questione dell'interlinking dei testi per la creazione di network costruito sul reference linking tra notizie correlate.

Anche l'accesso o meglio la presentazione delle risorse è oggetto di interessanti ricerche attraverso sistemi di organizzazione della conoscenza, volti a migliorare le attuali funzionalità di search e browsing in particolare creando delle "viste" per profili di utenza diversificate, o per esempio creando strutture d'accesso per un browsing e navigazione entro schemi di classificazione intercorressi e opportunamente mappati.

L'area della MKM Mathematical Knowledge Management [3] è la nuova frontiera tra la matematica e l'informatica che si occupa di tutti questi aspetti e di altri ancora.

MKM, eccitante nuovo campo tutto da indagare, intersezione tra la matematica e l'informatica, è volta allo sviluppo di strumenti basati su sofisticate nozioni matematiche e su avanzate tecnologie del software che consentono di organizzare e rendere efficacemente utilizzabile il vasto complesso di risultati matematici oggi in nostro possesso e di applicarlo al mondo delle biblioteche digitali.

4. Conclusioni

Un aspetto vitale nella costruzione di collezioni digitali è assicurare l'interoperabilità, attraverso utilizzo di standard comuni per l'accesso entro un quadro di biblioteca digitale. Molto del lavoro dovrà basarsi su attività di benchmarking tra differenti progetti già operativi o in corso di attivazione, ma fondamentale sarà il coordinamento da parte di progetti europei come MINERVA che faranno da punto di riferimento forte.

Qualsiasi progetto di digitalizzazione non può essere considerato un punto a sé, ma un nodo comunicante entro un sistema di biblioteche digitali che colloquiano.

Comunanze e differenze sono poste nell'ottica dello scambio informativo e la condivisione di esperienze e di strumenti è un arricchimento collettivo.

La filosofia portante poggia sulla cooperazione tra differenti figure istituzionali e di mercato, e sull'interpretazione di biblioteca digitale come luogo "aperto" all'accesso di numerose utenze.

Riferimenti bibliografici

[1] De Robbio, Antonella (2002) Terzo Modulo dello Studio Biblioteconomico per lo Studio di fattibilità della BEIC Biblioteca Europea di Informazione e Cultura: Digitalizzazione e Copyright". Aprile 2002 (pagg. 191).

<http://eprints.rclis.org/archive/00000085/>

[2] De Robbio, Antonella (2003) Digital Math Library DML: un progetto mondiale per la matematica. Technical Report, Biblioteca del Seminario Matematico, Università degli Studi di Padova.

<http://eprints.rclis.org/archive/00000099/>

[3] De Robbio, Antonella and Marini, Alberto (2003) MKM Mathematical Knowledge Management: Second International Conference on MKM Mathematical Knowledge Management 2003. Bertinoro (Forlì), 16-20 febbraio 2003 [Italian]. *AIDA Informazioni*.

<http://eprints.rclis.org/archive/00000119/>