JLIS.it
Italian Journal of Library and Information Science
Rivista italiana di biblioteconomia, archivistica e scienza dell'informazione
Pubblicazione periodica semestrale (esce in giugno e in dicembre)
ISSN: 2038-5366 (print) – ISSN: 2038-1026 (online)
Website: http://jlis.it – Email: info@jlis.it

ABBONAMENTO 2012:
Italia € 50,00; Estero € 60,00.

SUBSCRIPTION 2012:
Italy € 50,00; Overseas € 60,00.

Ledizioni – LediPublishing – Via Alamanni 11
20141 - Milano - Italia
Tel. +39-0245071824 – Fax +39-0242108107 – IVA/VAT: IT04627080965

Vol. 4, n. 2 (Luglio/July 2013)

JLIS.it

Italian Journal of Library and Information Science
Rivista italiana di biblioteconomia,
archivistica e scienza dell'informazione

Università di Firenze
Dipartimento di Storia, Archeologia, Geografia, Arte e Spettacolo (SAGAS)

JLIS.it è edita dall'Università di Firenze, Dipartimento di Scienze dell'Antichità, Medioevo e Rinascimento e Linguistica.
PROPOSTE DI PUBBLICAZIONE & PEER-REVIEW: Le submission, compiute tramite il sito web della rivista, verranno inizialmente esaminate da un editor e, superata la prima valutazione, saranno inviate a due revisori per il processo di peer review, al termine del quale verrà notificata l'accettazione o meno del contributo, o l'eventuale richiesta di modifiche.

JLIS.it is published by the University of Florence, Department of studies on the Antiquities, Middle Age, the Renaissance and Linguistics.
SUBMISSION & PEER-REVIEW: Papers submitted via the journal website will be checked by one of the editors: if these pass the first step, the papers will be sent to two reviewers for the peer-review process. After this, the author will be notified on the acceptance of his paper, or will be given suggestions on how to improve it.

Sommario
Table of Contents
Vol. 4, n. 2 (Luglio/July 2012)

Saggi = Essays

Fare il punto = Making the Point on

Reports & Reviews

Lectio magistralis in biblioteconomia

JLIS.it

L'uso del sistema Nuovo Soggettario per l'indicizzazione semantica di risorse web: problemi e proposte

Elisa Bianchi, Maria Clotilde Camboni, ElenaLazzarini

1 Il contesto

L'obiettivo di questo contributo è illustrare alcune questioni e spunti di riflessione relativi all'indicizzazione semantica di risorse web svolta nell'ambito del progetto "Panoramafirb".[1] Il progetto Panora-

[1]Il progetto Panoramafirb (FIRB RBNE07C4R9), http://www.panoramafirb.it, è finanziato dal Ministero dell'Istruzione, dell'Università e della Ricerca (Decreto n. 190/Ric., 12 marzo 2009), e si concluderà nel giugno 2013. Le autrici del presente contributo hanno collaborato alla definizione della struttura e alla costruzione del catalogo dei siti web Panoramafirb descritto nei paragrafi seguenti. Partecipano al progetto i Dipartimenti di Studi Italianistici, Storia delle arti e Informatica dell'Università di Pisa, il Dipartimento di Italianistica e spettacolo dell'Università degli studi di Roma 'La Sapienza', il Consorzio ICoN - Italian Culture on the Net, la Direzione Generale per i Beni Librari e gli Istituti Culturali del Ministero per i beni e le attività culturali e Cap s.p.a. Nell'ambito del progetto, è stata avviata una collaborazione con il Sistema Bibliotecario d'Ateneo (SBA) dell'Università di Pisa e, attraverso di esso, con la Biblioteca Nazionale Centrale di Firenze (d'ora in poi, BNCF). Per l'aiuto preziosissimo che hanno fornito, desideriamo ringraziare le bibliotecarie del Sistema Bibliotecario d'Ateneo Cinzia Bucchioni, Francesca Cecconi, Anna Colotto, Daniela Fiaschi, Anna Delogu, Chiara Garzetti, Maria Picciani, Cinzia Romagnoli, Elisabetta Soldati, Paola Spinesi, Simona Turbanti.

JLIS.it. Vol. 4, n. 2 (Luglio/July 2013).
DOI: 10.4403/jlis.it-8828

mafirb ha preso le mosse dalla constatazione che è molto difficile, soprattutto per un utente non esperto, reperire in rete risorse di qualità relative alla lingua e linguistica, letteratura e arte italiane.[2] Il web infatti si presenta oggi come un enorme repertorio non organizzato di siti, pagine e materiali di varia natura, in cui contenuti rilevanti e qualitativamente validi sono mescolati a tanti altri estremamente scadenti. Ai fattori qualitativi e quantitativi si aggiunge la dimensione temporale, lungo la quale le risorse web possono essere ordinate secondo un progressivo grado di permanenza vs. instabilità dei contenuti, in alcuni casi intrinsecamente legato al tipo di sito (i contenuti di un blog, per esempio, saranno molto meno stabili di quelli di un sito istituzionale o di una rivista online).[3] L'obiettivo principale del progetto era favorire l'orientamento dell'utente "non esperto" nella selezione e valutazione di risorse in rete relative a lingua e linguistica, letteratura e arte italiana, attraverso due strumenti: un catalogo di siti e un metamotore di ricerca sul web. In questo contributo desideriamo impostare una riflessione sull'esperienza (in parte ancora in corso) di indicizzazione per soggetto delle risorse web attraverso

[2]Sull'argomento non esistono studi sistematici ma, limitatamente al dominio di lingua e linguistica italiana, un quadro sintetico ma esaustivo dei contenuti disponibili è fornito dai due contributi di Mirko Tavosanis pubblicati nella sezione "Lingua italiana" del Magazine online di Treccani: "La lunga marcia attraverso il web" http://www.treccani.it/magazine/lingua_italiana/speciali/divulgazione/Tavosanis.html e "L'italiano (e la grammatica) nel web" http://www.treccani.it/magazine/lingua_italiana/speciali/grammatica/Tavosanis.html.

[3]La questione della persistenza, autorevolezza e affidabilità dei contenuti web e dei problemi di catalogazione ad essa connessi è già stata trattata da Lunghi et al, che ne discutono in relazione al passaggio dai documenti ai linked data, ed è una questione molto complessa, cui in questa sede è opportuno solo accennare. Esiste inoltre una consistente bibliografia sulla definizione di standard catalografici per le risorse elettroniche, tra cui segnaliamo le due corpose monografie di Stefano Gambari e Mauro Guerrini (*Definire e catalogare le risorse elettroniche*; *Le risorse elettroniche. Definizione, selezione e catalogazione*).

il sistema Nuovo Soggettario,[4] svolta nell'ambito dell'allestimento del catalogo: in particolare, ci proponiamo di tracciare un quadro dei termini del Thesaurus NS usati, dei nuovi termini che abbiamo proposto di inserire e delle criticità e spunti di riflessione emersi nella selezione dei contenuti da indicizzare e nella costruzione delle stringhe di soggetto.

Il catalogo, che allo stato attuale consiste di circa 1000 schede, adotta con alcune modifiche (v. infra, § 2) il modello di dati del progetto europeo MICHAEL (Multilingual Inventory of Cultural Heritage in Europe[5]) e prevede tre tipi di schede: Istituzione, Servizio e Collezione. L'indicizzazione semantica dei contenuti web riguarda le due schede Servizio e Collezione, e ad essa è dedicato il campo "Tema", dove si trovano una o più stringhe di soggetto costruite in base al metodo pre-coordinato. Ciascuna stringa corrisponde a un'unità di contenuto del sito o della collezione catalogata, intendendo con "unità di contenuto" l'intero sito, se le informazioni e tipi di dati ivi contenuti sono caratterizzati da una certa omogeneità, oppure una sottosezione di esso, o ancora un insieme di dati o materiali che possano essere contrastivamente descritti come un insieme distinto rispetto agli altri contenuti presenti (v. § 2).

2 Il modello di dati

Ai fini della costruzione del catalogo, è stato adottato il MICHAEL Data Model, ma è stato necessario apportarvi alcune modifiche, che lo rendessero più adatto a descrivere le risorse di interesse per il progetto. Infatti nel MICHAEL Data Model il principale oggetto di interesse per l'utente, e quindi del catalogatore, è la "collezione

[4] http://thes.bncf.firenze.sbn.it e Biblioteca Nazionale Centrale di Firenze.
[5] http://www.michael-culture.org.

digitale",[6] ma solo una parte dei contenuti online da censire ai fini del progetto Panoramafirb poteva essere considerata tale (al massimo tra un quarto e un quinto delle risorse oggetto di catalogazione). Inoltre, mentre il progetto MICHAEL si propone principalmente di censire il patrimonio culturale digitale, non di darvi accesso (tanto è vero che sono catalogati DVD e altre risorse accessibili solo in loco), uno degli scopi principali di Panoramafirb era permettere all'utente finale l'immediato reperimento delle risorse in rete, fornendogli l'indirizzo preciso (URL con link) del sito web in cui era possibile rintracciare i contenuti desiderati. La necessità di associare le risorse catalogate ad un indirizzo web era inoltre legata a vincoli tecnici connaturati alle esigenze di sviluppo del progetto (in particolare ai meccanismi di indicizzazione del metamotore di ricerca sviluppato nell'ambito del progetto stesso), che determinavano anche i requisiti per la scelta di una certa URL tra le diverse che spesso rinviano alla stessa risorsa. L'oggetto "risorsa online", su cui doveva focalizzarsi la catalogazione, finiva quindi con l'avvicinarsi molto al "sito web".

Nel MICHAEL Data Model, i siti web sono solo uno dei possibili servizi che possono dare accesso a una collezione, e di conseguenza sono catalogati seguendo un modello non adatto agli scopi di Panoramafirb. I problemi più rilevanti scaturivano dal fatto che nel MICHAEL Data Model le schede dei servizi non hanno una sezione dedicata al tema. È evidente che ciò, nel catalogo Panoramafirb, avrebbe pregiudicato la possibilità degli utenti di rintracciare in base ai contenuti di loro interesse la maggior parte delle risorse catalogate.

D'altro canto, il "sito web" come oggetto di catalogazione pone problemi almeno in parte già noti.[7] Il concetto viene correntemente

[6] Va qui osservato che le collezioni digitali di MICHAEL sono molto spesso l'esito della digitalizzazione di collezioni tradizionali.

[7] Tali problemi vengono affrontati nel quadro dei diversi progetti che si propongono di studiare soluzioni per l'archiviazione del Web, tra i quali si può citare in

usato e compreso senza apparenti difficoltà, ma non ha una definizione formale univoca. Intuitivamente, si può dire che un sito web è un insieme di pagine Internet che si trovano nello stesso dominio web: perlopiù è vero, ma non sempre.

Alcuni siti infatti occupano più di un dominio: ad esempio, il sito della rivista online Bollettino '900[8] permette di leggere il numero in corso nel dominio web http://www3.unibo.it, ma dalla ricerca nel sito si arriva agli articoli dello stesso numero tramite il dominio http://www.boll900.it (la situazione è ancor più complessa, ma in questa sede è inutile approfondire). Molto più spesso accade che all'interno dello stesso dominio si trovino contenuti assai eterogenei, più o meno distinguibili: http://www.maldura.unipd.it/italianistica/ALI ospita una bibliografia sulle autrici italiane dei sec. XIX-XX; http://www.maldura.unipd.it/ami/php corrisponde all'Archivio metrico italiano (database di versi con marcatura degli accenti metrici, da opere dei secoli dal XIII al XVI); http://www.maldura.unipd.it/masters/italianoL2/Lingua_nostra_e_oltre rimanda a una rivista che si occupa di aspetti teorici e applicativi dell'apprendimento e insegnamento dell'italiano come lingua seconda; http://www.maldura.unipd.it/alci è una pagina dove vengono riuniti i lemmari di più repertori di neologismi.[9]

Vista questa situazione, è stato inevitabile adottare una soluzione di compromesso che permettesse ai catalogatori il massimo della flessibilità. Sono state riprese le entità Collezione e Servizio del

particolare Archives de l'Internet della Bibliothèque Nationale de France, soprattutto perché ha sperimentato un approccio incentrato sul sito web e non sulla singola pagina (Abiteboul et al. 7). Sul concetto di sito web si interrogano anche gli storici della rete (Brügger). Vi sono anche studi tesi a trovare un metodo per identificare automaticamente le pagine appartenenti ad un dato sito web, ma ricadono ovviamente al di fuori dell'ambito di questo contributo.

[8] http://www.boll900.it.

[9] Esiste anche un problema legato ai cosidetti mirror sites: nella nostra prospettiva aveva però un'incidenza minore rispetto ai precedenti.

MICHAEL Data Model, ma con forti modifiche, soprattutto nel caso della seconda. Rispetto al progetto MICHAEL, il focus della catalogazione e conseguentemente dell'indicizzazione semantica si è infatti spostato verso l'entità Servizio (senza dubbio quella usata più di frequente dai catalogatori), non più considerata un semplice punto di accesso ai contenuti, ma un contenitore degli stessi, più o meno coincidente col sito web. Nel modello adottato, l'entità Servizio è stata quindi arricchita di diversi campi, tra i quali "Tema" (dedicato all'inserimento delle stringhe di soggetto), "Periodo storico", eccetera. Va notato che, visto l'ambito e gli scopi specifici del progetto, sia i record Collezione che i record Servizio sono stati inoltre dotati di un campo obbligatorio "Dominio tematico", in cui segnalare se la risorsa online catalogata era pertinente all'arte, la letteratura o la lingua italiana (con la possibilità di selezionare uno, due o tutti i domini). Per la delimitazione degli oggetti da catalogare – ovvero i siti web – è stato scartato un approccio di tipo strettamente informatico (come sopra visto, pressoché impossibile) a favore di uno che, pur tenendo in considerazione il più possibile le esigenze tecniche del progetto, le armonizzasse con quelle della catalogazione e quindi con l'usabilità del catalogo da parte degli utenti. In particolare, una delle soluzioni adottate per ridurre il problema dell'eterogeneità dei materiali all'interno di una risorsa è stata la catalogazione separata di parti rilevanti chiaramente individuabili e dedicate a un argomento specifico di alcuni siti web, nei casi in cui ciò venisse considerato utile. Le schede delle sezioni catalogate separatamente sono state collegate a quelle dei siti di cui fanno parte tramite le relazioni "è parte di" e "contiene", la cui applicazione all'entità Servizio è un'altra novità del modello adottato rispetto al MICHAEL Data Model. Un esempio di applicazione di questa soluzione è il sito Italica,[10] che è stato catalogato con schede diverse corrispondenti all'intero sito

[10] http://www.italica.rai.it.

e alle sezioni di esso dedicate a Dante,[11] alla narrativa italiana del Novecento,[12] alla storia della lingua italiana,[13] al Rinascimento,[14] eccetera.

3 La scelta del sistema NS

La scelta di utilizzare il sistema NS per l'indicizzazione semantica delle risorse web è stata il punto di arrivo di un percorso di valutazione di altre risorse lessicali disponibili: abbiamo infatti preso in considerazione, in particolare, Ital-Wordnet[15] e DMOZ.[16] Ital-Wordnet è un database semantico-lessicale organizzato secondo tassonomie e relazioni lessicali codificate, liberamente consultabile in rete. Non è una risorsa lessicale disciplinare, ma contiene parole dell'italiano generale. DMOZ è un progetto di classificazione manuale di siti e risorse web attraverso l'attribuzione di etichette relative al tipo di sito, al contenuto ecc.; le etichette ("categorie") sono organizzate in tassonomie, che possono essere navigate per livelli successivi di complessità. All'interno di DMOZ, quindi, le etichette si riferiscono all'intero sito, e determinano la collocazione del sito nelle liste di risorse recensite. Il requisito centrale delle tassonomie di DMOZ è la natura "user friendly".

Sulla base delle proprietà dei due strumenti sopra illustrati, per l'indicizzazione semantica delle risorse catalogate il sistema NS è stato scelto in virtù delle seguenti considerazioni:

[11] http://www.italica.rai.it/monografie/dante.
[12] http://www.italica.rai.it/monografie/grandi_narratori_900.
[13] http://www.italica.rai.it/monografie/storia_lingua_italiana.
[14] http://www.italica.rai.it/monografie/rinascimento.
[15] http://www.ilc.cnr.it/iwndb/iwndb_php.
[16] http://www.dmoz.org/World/Italiano.

1. il thesaurus NS è un vocabolario controllato, che ha una vasta copertura disciplinare e una stretta connessione con fonti bibliografiche e lessicografiche autorevoli;

2. è uno strumento recente (la prima versione è stata rilasciata nel 2006), in continuo aggiornamento;

3. esiste un vivo dibattito terminologico tra i gruppi che ne curano l'implementazione, secondo un percorso strutturato di proposta - discussione - approvazione - validazione di nuovi termini;

4. nel corso del progetto, si è presentata l'opportunità di inserirsi nel dibattito proponendo nuovi termini (v. infra, §§ 6 e 7);

5. la combinazione dei termini nelle stringhe di soggetto, secondo la sintassi pre-coordinata, consente di descrivere una vasta gamma di contenuti con un numero limitato di termini, e quindi di produrre, nel catalogo, raggruppamenti omogenei di siti;

6. l'indicizzazione semantica attraverso le stringhe di soggetto consente all'utente di fare ricerche sia per termini singoli (analogamente alle categorie di DMOZ) sia per combinazione di termini, e quindi di consultare il catalogo delle schede Servizio e Collezione per livelli successivi di specificità e raggruppando i siti per omogeneità di contenuto con gradi diversi di granularità.

4 Problemi di scrittura delle stringhe a copertura di oggetti eterogenei e livello di granularità dell'indicizzazione semantica

Così come il modello di dati, è stato necessario adattare alle caratteristiche degli oggetti catalogati (come definiti nel § 2) anche il Sistema NS: siamo stati costretti a reinterpretare e ricontestualizzare l'obiettivo delle stringhe "coestese con il contenuto di soggetto che debbono rappresentare" (Biblioteca nazionale centrale di Firenze 101-105). Malgrado l'adozione dell'approccio di descrizione di un sito su più livelli (descritto nel § 2), non è stato infatti sempre possibile stabilire una relazione biunivoca tra stringa di soggetto e contenuti oggetto di catalogazione. Nella costruzione delle nostre stringhe è stato quindi necessario tener conto della natura miscellanea e spesso non uniforme delle risorse da catalogare. In questi casi si è ritenuto di dover utilizzare più stringhe, assimilando di fatto i siti alla tipologia (prevista nel Sistema NS) degli studi miscellanei o "Scritti in onore", in modo da rendere quanto più possibile ragione dell'articolazione dei contenuti (Biblioteca nazionale centrale di Firenze). Un caso esemplare a questo proposito è la sezione dedicata al Rinascimento[17] del sito Italica.[18] Malgrado si fosse optato per una catalogazione a più livelli, infatti, indicizzare tale sezione con il solo termine "Rinascimento" sarebbe stato insufficiente, e d'altra parte al suo interno si trovano materiali di natura eterogenea: brani di opere dell'epoca, studi di ambito rinascimentale (nelle due sottosezioni Saggi e Monografie) e altri documenti. Nel campo "Tema" della

[17] http://www.italica.rai.it/monografie/rinascimento.
[18] http://www.italica.rai.it.

scheda dedicata al sito sono state quindi inserite più stringhe, tra le quali Rinascimento – Studi e Rinascimento – Opere – Antologie.

La scelta del livello di granularità dell'indicizzazione semantica è stata comunque fatta in un'ottica contrastiva rispetto alla totalità delle risorse catalogate. La scheda del ricchissimo sito web del centro di studi dedicato a Primo Levi[19] ha nel campo "Tema" unicamente "Levi, Primo", e lo stesso accade per molti altri siti incentrati su autori riguardo ai quali si trovano risorse specifiche solo in un paio di domini web, indipendentemente dalla qualità e ricchezza (a volte notevole) dei materiali in essi presenti. Invece, nel caso delle decine di siti che si occupano di Dante Alighieri e della sua opera, molto spesso nel campo "tema" di una singola scheda sono state inserite più stringhe, per segnalare esattamente quali risorse venissero rese disponibili dal sito catalogato (testi di opere, traduzioni degli stessi, bibliografie, trascrizioni o riproduzioni di manoscritti, studi, riviste scientifiche dedicate...). La scelta di procedere in questo modo è stata compiuta pensando alle difficoltà a cui sarebbe andato incontro un utente che dopo una richiesta apparentemente specifica si sarebbe trovato di fronte a decine di risultati: in questo modo, gli sono state offerte le possibilità da un lato di sfruttare l'indicizzazione semantica per raffinare la ricerca e trovare più agevolmente i materiali a cui poteva essere interessato, e dall'altro di farsi un'idea dei contenuti dei siti web anche per mezzo delle stringhe di soggetto ad essi associate, oltre che della loro descrizione.

[19] http://www.primolevi.it.

5 Uso del ruolo "forma" (intellettuale/bibliografica)

Una delle peculiarità delle risorse web o comunque digitali rispetto a quelle librarie è la potenziale disponibilità di diverse modalità di fruizione di uno stesso contenuto. Un esempio chiaro può essere il testo di un'opera letteraria, o di più opere nel caso di banche dati o biblioteche digitali: esso può essere reso fruibile in una forma in cui può essere solamente letto (come le riproduzioni in formato immagine di testi a stampa), oppure in una modalità che permette di fare al suo interno delle ricerche. In quest'ultimo caso, si può trattare di semplici ricerche di stringhe all'interno di un unico documento, con o senza caratteri jolly, o di ricerche molto più complesse (su tutti i testi di un determinato archivio o su un determinato sottoinsieme, con ricerca delle co-occorrenze, etc.). A volte ai testi sono associati metadati anche molto raffinati, che aprono possibilità di interrogazione altrimenti impossibili: ad esempio, nel caso di DanteSearch,[20] i testi delle opere di Dante sono corredati di lemmatizzazione e marcatura grammaticale e sintattica. I testi o i risultati della ricerca possono inoltre essere scaricabili o non esserlo. Discorsi in parte analoghi e in parte diversi possono essere fatti per le bibliografie, e a contenuti diversi da questi possono essere applicate modalità di fruizione ancor differenti (ad esempio, il testo di un dizionario o di un'enciclopedia può essere offerto in modalità parzialmente o totalmente ipertestuale, nel caso in cui alcuni o tutti i termini presenti in una voce permettano di arrivare direttamente alla voce ad essi corrispondente).

È ovvio che, per un utente, le forme assunte dai diversi contenuti e le loro modalità di fruizione rivestono un interesse notevole. Di conseguenza, si è cercato di permettere la ricerca di determinate ri-

[20] http://dante.di.unipi.it:8080/DanteWeb.

sorse sulla base della loro forma e delle operazioni che vi si possono compiere, e ciò ha portato ad un notevole uso, nella catalogazione, di termini nel ruolo di forma intellettuale/bibliografica, a volte associati in serie (ad esempio Dizionari – Ipertesti o Commenti – Archivi di dati). Malgrado ciò, in diversi casi non è stato possibile fornire un'indicizzazione semantica che rendesse conto delle effettive particolarità di una risorsa. Una delle cause che concorrono a questa insufficienza è ovviamente il fatto che la terminologia del Thesaurus NS non è stata pensata allo scopo di descrivere risorse come quelle del web e in particolare del web 2.0, soprattutto quelle interattive (vedi oltre il caso di "Forum"). Va inoltre aggiunto il fatto che la riflessione e la discussione su questi aspetti sono ancora carenti, ed è ovviamente difficile risolvere questioni così complesse in poco tempo e affrontandole a partire da una prospettiva in fondo limitata. D'altra parte, la ristrettezza dell'ambito di applicazione della catalogazione unita all'indicazione della forma bibliografica/intellettuale hanno permesso di risolvere la difficile questione della catalogazione dei siti web che contengono biblioteche digitali onnicomprensive, come Googlebooks o Archive.org, e anche di quelli che mettono integralmente a disposizione le pubblicazioni scientifiche legate a una determinata università (repository di ricerca o siti di case editrici universitarie). In entrambi i casi, si è optato per indicare i soggetti di interesse per il progetto (arte, letteratura e lingua italiana) seguiti dalla forma "Biblioteche digitali".

6 Termini nuovi e combinazione di termini già esistenti

Per via delle esigenze esposte nel paragrafo precedente, nell'ambito della discussione con BNCF le proposte di nuovi termini[21] hanno riguardato soprattutto – oltre ovviamente a etichette specifiche dei domini disciplinari arte, letteratura e lingua o linguistica italiana – termini che potessero rendere ragione delle diverse forme e modalità in cui le risorse web possono essere messe a disposizione dell'utente.

Nel thesaurus si trovano alcuni termini utili per descrivere la "natura" dei siti, sia dal punto di vista tecnico che della struttura dei dati: "Weblog" per i "Blog" (forma, quest'ultima, indicata come non preferita); "Biblioteche digitali" per i siti che contengono banche dati di opere digitalizzate, "Archivi di dati" per i database (online e scaricabili). Alcuni di questi termini sono stati inseriti su proposta delle catalogatrici del progetto Panoramafirb: tra essi, "periodici elettronici" e "forum".

Per quel che riguarda "forum", è facile intuire che si tratta di un termine fondamentale per la catalogazione di molti siti, essendo attivi allo stato attuale molti forum con una tradizione ormai consolidata negli anni, che costituiscono un punto di riferimento per gli utenti della rete (per esempio, per quelli interessati alla lingua italiana il forum SoloItaliano di Wordreference[22]).

Quanto invece a "Periodici elettronici" (inserito come NT di "Pubblicazioni elettroniche"), all'interno del catalogo Panoramafirb sarebbe stato utile affiancargli "Periodici scientifici elettronici". L'u-

[21] Nel corso della collaborazione con il Sistema Bibliotecario di Ateneo dell'Università di Pisa e con BNCF, sono stati proposti e accettati in totale 9 termini "comuni" (Abstract, Collane editoriali, Edizioni elettroniche, Excerpta, Forum, Frontespizi, Italianistica, Periodici elettronici, Bollettini elettronici) e circa 20 termini disciplinari appartenenti a letteratura, lingua e linguistica e arte italiana.

[22] http://forum.wordreference.com/forumdisplay.php?f=51.

so dei due termini consentirebbe infatti di distinguere le riviste online in generale dalle riviste scientifiche, che seguono precisi percorsi e standard di revisione, indicizzazione e pubblicazione dei contributi. Tuttavia, nell'ottica di limitare il più possibile la proliferazione di termini nel Thesaurus a favore piuttosto dell'espressione di un determinato concetto tramite la combinazione di termini già presenti, BNCF ha optato per non accettare il termine.

Quello della proposta di "Periodici scientifici elettronici" è stato uno dei tanti casi in cui è sorta una questione di grande rilievo, che concerne i meccanismi di equilibrio tra due esigenze contrapposte ed egualmente importanti: quella di utilizzare un vocabolario controllato ed esprimere concetti differenti attraverso la combinazione dei termini nella sintassi delle stringhe di soggetto da una parte e quella di "anticipare" le query dell'utente con termini di uso comune dall'altra.

Così, un termine che abbiamo proposto è stato "Edizioni elettroniche", che abbiamo definito come "Edizioni pubblicate in formato elettronico, destinate alla lettura e a funzioni avanzate di ricerca e di elaborazione dei contenuti." Questo termine sarebbe molto utile per descrivere le edizioni (nel senso filologico del termine) create in formato digitale e raccogliere in un unico gruppo omogeneo i numerosi siti frutto di progetti che avevano come scopo la creazione di edizioni digitali di opere: ad esempio le opere di Dante lemmatizzate, o ancora le grammatiche digitalizzate (come immagini e come testo) della Biblioteca dell'Accademia della Crusca.

Nel dibattito che è stato avviato su questo termine, BNCF ha proposto di scomporre "Edizione elettronica" in "Edizioni, Pubblicazioni elettroniche", tanto è vero che, allo stato attuale, Edizione elettronica è indicato come termine non preferito, e rimanda all'uso dei due termini combinati.

La continua dialettica tra la necessità di avere un vocabolario

controllato e uniforme da una parte e di rappresentare in maniera unitaria la specificità (terminologica e concettuale) di un settore disciplinare, dall'altra, si fa ancora più stringente nel caso della descrizione di risorse web per un catalogo elettronico, in cui l'utente accede al catalogo attraverso query, di cui il catalogatore deve in qualche modo tenere conto: così, un termine centrale per le ricerche sul web, "E-learning", nel thesaurus BNCF è indicato come forma non preferita, da sostituire con la combinazione di tre termini, cioè "Educazione, Impiego, Internet": tale divergenza tra indice di frequenza (e verosimilmente, di familiarità) di un termine nelle ricerche sul web e controllo del vocabolario nel thesaurus pone un interessante spunto di riflessione (allo stato attuale, una questione aperta) sul rapporto tra gli standard di indicizzazione semantica e le query digitate dagli utenti.

7 Termini di dominio utilizzati e termini nuovi

7.1 Arte italiana

Per quel che riguarda i termini o le categorie di termini utilizzati nell'indicizzazione semantica di risorse online relative al dominio Arte Italiana è stato necessario ricorrere a elementi non presenti nel Thesaurus, come i nomi propri di artisti e di opere. Tra i termini inseriti su nostra proposta si può citare "Pittura ferrarese". Per quanto riguarda le forme bibliografiche/intellettuali, nelle schede di Arte ricorrono frequentemente i seguenti termini: Collezioni, Monumenti, Collezioni digitali.

7.2 Letteratura italiana

Per quel che riguarda i termini o le categorie di termini del Thesaurus del NS più adoperati nell'indicizzazione semantica di risorse online relative alla letteratura italiana, un ruolo molto importante è stato giocato – com'era d'altra parte prevedibile – da elementi nel Thesaurus non presenti, ovvero i nomi propri (in particolare di autori e opere). Il termine in assoluto usato con la massima frequenza – e anche questo era prevedibile – è "Letteratura italiana"; altri termini specifici (come "Letteratura drammatica italiana", "Letteratura dialettale sarda", "Poesia per musica", etc.) sono stati adoperati molto più di rado. La "categoria di appartenenza" dei termini che nel loro insieme sono adoperati più spesso è però quella delle diverse etichette atte a ricoprire il ruolo di forma intellettuale/bibliografica, come: Opere, Edizioni, Libretti, Manoscritti, Descrizioni, Riproduzioni, Studi, Testi, Ipertesti, Biografie, Traduzioni, Periodici, Indici, Autografi, Incunaboli eccetera. Ciò appare una conseguenza scontata di quanto detto nel § 5; altra meno scontata è il fatto che i nuovi termini proposti e accettati per l'inserimento nel Thesaurus sono in buona parte termini non specificamente riconducibili alla letteratura italiana ma necessari per indicizzare in maniera adeguata alcune risorse: Abstract, Collane editoriali, Edizioni interpretative, Forum, Frontespizi, Periodici elettronici, Periodici umbri. Nella maggior parte dei casi, le nuove proposte riconducibili alla letteratura italiana rientrano nella serie "Luoghi carducciani", "Luoghi folenghiani", "Luoghi leopardiani"... Il termine di maggior rilievo disciplinare tra quelli entrati nel Thesaurus su nostra proposta è "Italianistica", che però non è specifico dell'ambito letterario, dal momento che la disciplina copre anche gli studi rivolti alla lingua italiana.

7.3 Linguistica italiana

Per quanto riguarda il dominio di linguistica italiana, i termini del thesaurus NS (e le combinazioni di termini) maggiormente usati rappresentano le due principali aree tematiche in cui possono essere raggruppati i siti web dedicati alla lingua e linguistica italiana:

Lingua italiana - Insegnamento [agli] Stranieri
Lingua italiana – Grammatica

Per quanto riguarda le forme bibliografiche/intellettuali, nelle schede di linguistica ricorrono frequentemente i seguenti termini: Biblioteche digitali, Corpora, Forum (termine inserito da BNCF su nostra proposta) e Weblog (forma preferita di "Blog"). Tali termini rappresentano molto bene la grande bipartizione dei siti dedicati alla lingua e linguistica italiana: da una parte, infatti, abbiamo siti "divulgativi", creati da e per utenti non specialisti e che hanno un interesse generico per la lingua italiana (forum, blog); dall'altra, una parte consistente di siti catalogati appartiene all'ambito scientifico/accademico, ed è costituita da biblioteche digitali e corpora testuali destinati alla comunità scientifica. Una questione aperta, che riteniamo opportuno attualizzare in questa sede, è rappresentata dalla catalogazione di materiali didattici di varia natura e pubblicati in vari tipi di siti: si tratta di documenti scaricabili (per esempio in formato .doc o .pdf), di pagine html o intere sezioni di siti contenenti esercizi, progettazione di percorsi didattici, spunti per attività di vario genere, indicazioni bibliografiche destinate ai docenti di lingua italiana sia come lingua materna che come lingua seconda.Ora, a prescindere al problema dell'autorevolezza e persistenza di questi documenti, abbiamo comunque ritenuto opportuno catalogare e indicizzare queste risorse con stringhe di soggetto dedicate, utilizzando gli unici termini disponibili nel thesaurus: Schede didattiche

ed Esercizi. Sarebbe senz'altro proficuo, non solamente per il dominio disciplinare della lingua e linguistica italiana, articolare meglio la tassonomia dell'etichetta di nodo [Parti e proprietà di documenti], introducendo termini specifici per documenti e siti del web progettati per scopi didattici (sia per docenti che per studenti).

8 Conclusioni

L'esperienza di costruzione del catalogo Panoramafirb dei siti web ha permesso di approfondire la questione della definizione di "sito web", "risorsa elettronica" e in generale della descrizione e catalogazione dei contenuti web. Inoltre, l'indicizzazione semantica delle risorse catalogate ha evidenziato la necessità di adattare gli strumenti terminologici esistenti (nel nostro caso, il NS) attraverso due percorsi: in primo luogo, la creazione di termini nuovi sia disciplinari (di lingua, linguistica e arte italiana) sia comuni (relativi alla natura dei contenuti descritti), in secondo luogo l'elaborazione di criteri specifici per la segmentazione dei contenuti e la descrizione attraverso le stringhe di soggetto. Il risultato è un catalogo uniforme dal punto di vista della soggettazione, in cui a stringhe di soggetto simili corrispondono risorse affini, e in cui quindi è possibile ravvisare una corrispondenza biunivoca tra contenuto e termini/stringhe di soggetto. Certamente, la costruzione del catalogo ha posto alcune questioni che rimangono tuttora aperte e che richiederebbero di essere sviluppate nell'ambito di ricerche ulteriori: in particolare, riteniamo che le due linee di sviluppo più promettenti siano da una parte i criteri di identificazione (e catalogazione) univoca delle risorse web selezionate, dall'altra lo sviluppo di funzionalità del database del catalogo che massimizzino il potenziale esplicativo e di raggruppamento di risorse omogenee delle stringhe di soggetto e dei termini del thesaurus NS.

Riferimenti bibliografici

Abiteboul, Serge, et al. «A First Experience in Archiving the French Web». *Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries*. ECDL '02. London: Springer-Verlag, 2002. 1–15. (Cit. a p. 5).

Biblioteca nazionale centrale di Firenze. *Nuovo soggettario. Guida al sistema italiano di indicizzazione per soggetto. Prototipo del Thesaurus*. Milano: Bibliografica, 2006. (Cit. a p. 9).

Brügger, Niels. «L'historiographie de sites Web: quelques enjeux fondamentaux». *Le Temps des Médias* 18.1 (2012): 159–169. (Cit. a p. 5).

Gambari, Stefano e Mauro Guerrini. *Definire e catalogare le risorse elettroniche*. Milano: Bibliografica, 2002. (Cit. a p. 2).

——.*Le risorse elettroniche. Definizione, selezione e catalogazione*. Milano: Bibliografica, 2002. (Cit. a p. 2).

ELISA BIANCHI, Consorzio ICoN.
e.bianchi@italicon.it

MARIA CLOTILDE CAMBONI, Università di Pisa.
m.c.camboni@humnet.unipi.it

ELENA LAZZARINI, Università di Pisa.
e.lazzarini@arte.unipi.it

Bianchi, E., M. C. Camboni. A., A. Lazzarini. "L'uso del sistema Nuovo Soggettario per l'indicizzazione semantica di risorse web: problemi e proposte". *JLIS.it*. Vol. 4, n. 2 (Luglio/July 2013): Art: #8828. DOI: 10.4403/jlis.it-8828. Web.

ABSTRACT: The essay deals with the creation of subject-headings for web resources related to Italian literature, art and linguistics with resort to the Nuovo soggettario system (NS). It describes the difficulties arisen and the results achieved in this regard during the development of a project aimed at creating web tools to facilitate the location of high quality web resources about Italian culture. One of these tools was a catalogue of web resources with subject headings, created using a modified version of the MICHAEL Data Model. The authors explain why they had to change the

model to meet the needs set by the peculiar items of the catalogue, why the choose the NS for the subject-headings, their choices about the granularity of the description, their particular use of the "Intellectual/Bibliographic form" roles of the NS to match the features of their items that could be relevant for a user, and their consequent proposals of new terms for the NS Thesaurus and the questions that arose from these proposals.

KEYWORDS: Digital cataloguing; Semantic indexing; Nuovo Soggettario

ACKNOWLEDGMENT: L'articolo è frutto di un lavoro condiviso; Elisa Bianchi ha scritto i §§ 1, 3, 6, 7.3 e 8; Maria Clotilde Camboni ha scritto i §§ 2, 4 5 e 7.2; Elena Lazzarini ha scritto il § 7.1.

Submission: 2013-03-14
Accettazione: 2013-04-08
Pubblicazione: 2013-07-01

JLIS.it

Le edizioni digitali come nuovo modello per dati di autorità concettuali

Francesca Tomasi

1 Introduzione

La progressiva estensione degli àmbiti di intervento computazionale agli oggetti del patrimonio culturale ha determinato un'attenzione maggiore al documento inteso come dato la cui capacità espressiva va oltre la sola descrizione metadatale a livello paratestuale. La trascrizione, per esempio, sta entrando nel circuito della rappresentazione del contenuto informativo di cui libri e documenti sono portatori. Sia in campo archivistico che librario l'attenzione verso il full-text ha obbligato a tradurre il sistema di metadatazione descrittivo, amministrativo-gestionale e strutturale, che si esprime comunemente al livello del paratesto, al livello del testo. E il metadato inizia così a configurarsi come un elemento di annotazione che può trasformare il testo, sia esso documento archivistico o fonte libraria, in edizione.

L'edizione digitale di un documento può essere intesa, attraverso l'annotazione, come un processo che porta alla progressiva stratificazione del sistema interpretativo dell'editore, in modo particolare nei sistemi di markup dichiarativo (Coombs, Renear e DeRose). I

JLIS.it. Vol. 4, n. 2 (Luglio/July 2013).
DOI: 10.4403/jlis.it-8808

diversi aspetti dell'analisi dei contenuti di un documento conducono alla creazione di una raccolta di informazioni multilivellari che nascono dal processo interpretativo. Tale processo è il modello del documento, inteso come oggetto informativo complesso, elaborato dall'editore critico. Tipicamente persone, luoghi, date, oggetti, eventi e parole chiave rappresentano istanze interpretative che si configurano come elementi dell'annotazione riferiti a valori che si presentano nella forma di stringhe di caratteri. Ogni stringa interpretata o annotata (composta da elemento descrittivo e valore associato) è potenzialmente un'informazione autonoma, legata al testo dell'edizione, necessaria a fornire i diversi punti di accesso al documento ovvero a determinare le possibili entries. Tale approccio è la base di partenza per creare liste controllate di valori di elementi, estraendo dal documento sia le forme attestate che le forme varianti di nomi di persona, di luoghi, date, titoli e soggetti, per associarle quindi alla forma controllata secondo lo standard adottato. Ma ogni stringa annotata (per esempio una stringa identificabile come un "nome di persona") richiama una serie di informazioni che vanno oltre la semplice annotazione e tali informazioni provengono sia dal contesto specifico di occorrenza della stringa che da fonti esterne (per esempio luogo e data di nascita, occupazione, relazioni con altre persone). E soprattutto gli elementi annotati non solo sono in relazione fra di loro, ma intrattengono anche relazioni con altre risorse distribuite. Si passa dall'edizione digitale alla raccolta di descrizioni di dati altamente strutturati che si possono caratterizzare come un nuovo modello di authority file, in cui il punto di accesso al documento è l'esito di una relazione fra elementi annotati in un determinato contesto testuale. L'authority si trasforma così da stringa a concetto e il processo di concettualizzazione è il risultato dell'accoppiata elemento-valore e della rete di collegamenti interni (fra elementi) ed esterni (far elementi e risorse distribuite).

In prima battuta diremo quindi che gli elementi annotati andranno posti in relazione attraverso adeguati predicati ontologici. Perché una stringa identificata come "data" e una identificata come "persona", o "luogo" o "evento" potrebbero avere una qualche connessione. Non è sufficiente un generico collegamento non tipizzato o sintattico, ma va specificata la ragione della relazione, individuando formalmente la tipologia di connessione fra gli elementi. La conoscenza che implicitamente nasce dalla lettura del documento viene così formalizzata attraverso relazioni semantiche esplicite: per esempio una data stabilisce il momento del trasferimento di una persona in un luogo; un luogo determina uno spazio in un cui un evento è stato organizzato da una persona; un soggetto identifica una feature di una persona. In secondo luogo ogni stringa annotata, oltre ad avere relazioni con altre stringhe interne al documento, ha relazioni con altri oggetti distribuiti che si riferiscono al medesimo contenuto informativo, sia a livello di singolo elemento (la stessa persona) che, soprattutto, a livello di concetto espresso in quel documento (una persona che intrattiene una relazione con un'altra persona in uno specifico contesto testuale).

Le persone, i luoghi, le date, i soggetti, gli eventi e gli oggetti vanno descritti secondo gli standard in uso, vanno messi in relazione fra di loro ad esprimere asserzioni, determinando concetti, e vanno relazionati con altre entità su WWW — che possono anche condividere lo stesso tipo di relazioni interne al documento — creando collegamenti incrociati.

Questo significa che le edizioni digitali devono confrontarsi con il mondo dei sistemi di metadazione in uso nel settore del cultural heritage, con i linguaggi formali del semantic web e con il crescente fenomeno linked data. Le edizioni digitali sono una base di conoscenza "naturaliter" linked. Le relazioni fra le stringhe annotate nascono cioè spontaneamente, all'atto della lettura del testo. Diremo

che il contesto in cui ogni stringa occorre rappresenta le ragioni del collegamento e stabilisce il dominio di riferimento. Contesto e dominio sono due concetti chiave nella trasformazione dell'annotazione in base di conoscenza perché identificano l'àmbito di modellazione dell'edizione.

Contesto in letteratura significa che le relazioni fra stringhe nascono dall'ambito semantico in cui tali stringhe compaiono (Lee). Le relazioni che possono essere formalizzate derivano quindi dalla specifica co-occorrenza di stringhe. Ma contesto è anche un concetto che richiama inevitabilmente lo standard ISAAR-CPF[1] e la sua formalizzazione EAC-CPF.[2] Il ruolo di ISAAR-CPF in particolare diventa importante nel processo di identificazione univoca di entità come persone e come relazioni fra persone, veicolando i concetti di soggetto produttore (sia esso persona, famiglia o ente), di relazione fra il soggetto produttore e gli oggetti prodotti (vale a dire le risorse di cui il soggetto assume una forma di paternità) e di collegamento fra soggetti produttori o in generale fra persone. Di EAC-CPF peraltro c'è l'ontologia recentemente proposta che formalizza classi e proprietà dello schema (Mazzini e Ricci).[3] Ce lo insegna l'archivistica "separating description of people from description of record" (Pitti) che in campo di edizione può essere tradotto nel separare la descrizione delle persone dal testo dell'edizione, ma mantenendo il collegamento fra la persona e il documento in cui quella persona occorre, che stabilisce il contesto. Affermazione, quella di Pitti, che può essere estesa dalle persone a ogni fenomeno dell'analisi. E

[1] International Standard Archival Authority Record for Corporate Bodies, Persons and Families. Second Edition, 2003. http://www.icacds.org.uk/eng/ISAAR(CPF)2ed.pdf.

[2] Encoded Archival Context – Corporate Bodies, Persons and Families. La specifica dello schema si può leggere all'indirizzo: http://eac.staatsbibliothek-berlin.de.

[3] EAC-CPF Vocabulary Specification 1.0 si può leggere all'indirizzo: http://archivi.ibc.regione.emilia-romagna.it/ontology/reference_document/referencedocument.html.

trasformare le entità e le loro relazioni in ontologie significa trasformare i testi in basi di conoscenza. Le edizioni digitali diventano allora sistemi su cui sviluppare forme di "knowldege representation" (Clement).

Ed entriamo così nel concetto di dominio come spazio di riferimento semantico. L'ontologia è per sua stessa natura una concettualizzazione di una realtà osservata rispetto ad un àmbito di riferimento. Allo stesso modo diverse edizioni di testi se avranno fra loro entità comuni (la stessa persona, lo stesso luogo, la stessa keyword) potranno avere relazioni diverse a seconda dell'àmbito in cui queste entità compaiono. Il concetto di ontologia di dominio deve fare quindi i conti sia con la realtà osservata rispetto allo specifico contesto, sia con il punto di vista assunto sull'oggetto dell'analisi. Scopo del presente contributo è quindi di: ragionare sulle entità, nella forma di stringhe estratte da un testo annotato (elemento-valore), come entries, e quindi come punti di accesso al documento, e ragionare su come queste ultime possono configurarsi come authority files; ragionare su come estendere il concetto di authority a quello di relazione in quanto ogni authority è legata ad un contesto e ad un dominio; ragionare sul concetto di relazione come collegamento fra le authorities così configurate nello spazio del WWW in un sistema di interlinking. Tentare quindi di "andare oltre le colonne d'Ercole" (Crupi) diventa lo scopo del processo che si intende qui descrivere.

Le edizioni digitali fanno parte del patrimonio culturale e vanno quindi valorizzate al pari delle raccolte librarie, archivistiche e museali, anche in considerazione della realizzazione di digital libraries nella forma di aggregatori di risorse come strumento di accesso integrato al patrimonio culturale (come è ad esempio Europeana,[4] v. Aloia, Concordia e Meghini).Il metadato aggregato non sarà quindi più solo un elemento estratto dalla descrizione della

[4]Il portale può essere consultato all'indirizzo: http://www.europeana.eu/portal.

risorsa, ma un elemento che proviene dal testo pieno dell'oggetto digitale. Le edizioni digitali in campo letterario possono fornire ai sistemi archivistici e librari un modello già testato e oggetto di studi e sperimentazioni che può favorire il processo di trascrizione integrale delle fonti documentali e librarie. Se il processo di dialogo avviato fra archivi, biblioteche e musei[5] si estendesse al settore delle digital humanities il patrimonio culturale amplierebbe le prospettive di interesse allargando la base di conoscenza a disposizione dell'utente finale. Le già esistenti authorities in settore archivistico e librario potranno poi essere arricchite di nuovi dati provenienti da nuove fonti ancora inesplorate.

2 Il panorama di riferimento

Nel campo delle edizioni digitali di testi si registra un numero crescente di sperimentazioni (Sahle). Solo per fare qualche esempio si può esplorare la classificazione delle edizioni del XIX secolo inglesi ed americane fatta da Nines.[6] o si possono consultare i numerosi progetti editoriali del DDH (Department of Digital Humanities) del King's College di Londra;[7] si possono anche vedere i lavori del CDS (Center for Digital Scholarship) della Brown University, come lo storico Women's Writers Project,[8] o accedere ai progetti dei vari centri che si occupano di digital humanities[9] o ancora consultare l'elenco

[5]Come dimostra l'interessante progetto italiano MAB (Musei, Archivi e Biblioteche): http://www.mab-italia.org.

[6]Networked Infrastructure for Nineteenth-Century Electronic Scholarship: http://www.nines.org. Si tratta di un aggregatore di metadati provenienti da "peer-reviewed digital objects'.

[7]http://www.kcl.ac.uk/artshums/depts/ddh/research/index.aspx.

[8]http://www.wwp.brown.edu.

[9]Una classificazione si può leggere sul sito CenterNet: http://digitalhumanities.org/centernet.

di edizioni, e di progetti di digital libraries o collezioni digitali in generale, che si basano su XML/TEI[10] sullo stesso sito dedicato allo schema.[11]

Anche le istituzioni archivistiche hanno avviato procedure di trascrizione integrale delle fonti,[12] arrivando al livello dell'item come nel progetto Datini, datato 2002, condotto sulla porzione dell'omonimo fondo delle lettere di Margherita Datini a Francesco di Marco[13] o come nel lavoro sul Codice diplomatico della Lombardia medievale.[14] O ancora non si può non menzionare, in campo di trascrizione di manoscritti, l'egregio lavoro di UCL (University College London) su Jeremy Bentham[15] come esempio di progetto collaborativo in un'ottica di "social edition" (Siemens et al.).

Anche il rapporto fra le edizioni e il ruolo delle tecnologie legate al semantic web ha portato alla realizzazione di prodotti digitali di eccellenza, come, per fare un esempio, il Discovery Project (D'Iorio e Barbera)[16] relativo alla filosofia. Senza dimenticare che digital libraries di testi, come la raccolta di classici della letteratura prodotti in seno al progetto Gutenberg, sono già esposti come linked data

[10]Si tratta del principale schema in uso in campo di markup di testi letterari e umanistici in senso ampio: http://www.tei-c.org.

[11]Projects using TEI: http://www.tei-c.org/Activities/Projects.

[12]Anche se non si può non notare che il neonato SAN (Sistema Archivistico Nazionale): http://san.beniculturali.it che vuole aggregare progetti digitali in ambito archivistico, riserva il concetto di digitalizzazione alla conversione di oggetti analogici, anche in termini di documenti di testo, nel solo formato immagine, riservando al metadato il solo ruolo descrittivo. La ragione evidentemente è che il numero di progetti di trascrizione annotata di documenti in campo archivistico è ancora limitata

[13]Progetto dell'Archivio di Stato di Prato: http://datini.archiviodistato.prato.it/margherita/index.htm.

[14]Progetto del Centro Scrineum dell'Università di Pavia: http://cdlm.unipv.it.

[15]Transcribe Bentham Transcription Desk: http://blogs.ucl.ac.uk/transcribe-bentham.

[16]http://www.discovery-project.eu/home.html.

sets[17] e già collegati ad altri data sets come DBpedia.[18]

Non è un caso poi se molte edizioni di testi si configurino come "archivi": Walt Whitman Archive,[19] Willa Cather Archive,[20] William Blake Archive,[21] Dante Gabriel Rossetti Archive,[22] Emily Dickinson's Archive;[23] si tratta di un processo che intende tradurre il concetto di edizione in quello di raccolta di documenti necessari alla classificazione del lavoro di un autore (Price). E l'edizione come archivio allarga il concetto di edizione a quello di base di conoscenza.

Un serbatoio quindi di informazione annotata che può essere arricchita e trasformata, diventare oggetto di riflessione alla ricerca di relazioni interne fra gli elementi e posta in collegamento con altre risorse per diventare una fonte di conoscenza. Se il processo di annotazione delle risorse a testo pieno, che ad oggi avviene nella maggior parte dei casi in forma manuale, potesse poi avvalersi di strumenti di riconoscimento automatico delle stringhe (information extraction, IE), e conseguente etichettatura, elaborati nel settore del natural language processing, come la named entity recognition, il sistema di costruzione di punti di accesso semantici ne trarrebbe ampio giovamento (per una visione d'insieme dei sistemi di IE si veda Chang et al.).

Ai vocabolari di annotazione in uso nel settore dell'edizione di testi, primo fra tutti lo schema Text Encoding Iniziative (TEI) basato

[17] Project Gutenberg Catalog: http://wifo5-03.informatik.uni-mannheim.de/gutendata.

[18] "DBpedia is a crowd-sourced community effort to extract structured information from Wikipedia and make this information available on the Web." http://dbpedia.org.

[19] http://www.whitmanarchive.org.

[20] http://cather.unl.edu.

[21] http://www.blakearchive.org/blake.

[22] http://www.rossettiarchive.org.

[23] http://www.emilydickinson.org.

sull'embedded markup XML, si aggiungono gli standard, vale a dire sets di metadati e relativi valori o ontologie, che identificano il sistema descrittivo delle risorse digitali in uso negli ambienti di gestione e trattamento del patrimonio culturale. A livello di metadati/ontologie, ovvero di element sets, il mondo degli archivi ha gli schemi EAD[24] e il già citato EAC-CPF, i musei hanno il CIDOC-CRM,[25] il Web, e i sistemi di esposizione di metadati, investono su DC[26] come strumento per la disseminazione. SKOS[27] è un modello in uso nel settore della costruzione di reti lessicali. FRBR[28] è un altro modello, standard dell'IFLA, che dalle biblioteche si sta estendendo ai diversi àmbiti della metadatazione di risorse in cui il processo di stratificazione, o il punto di vista multilivello, svolge un ruolo fondamentale nella descrizione dell'oggetto dell'analisi. E poi c'è Europeana che ha elaborato un data model finalizzato a raggruppare e mappare vari modelli concettuali e ontologie.[29] Al set di descrittori, elementi o classi, si aggiunge la questione dei valori. Altrettanto numerosi i vocabolari in uso nella forma della tassonomia o del thesaurus: p.e. AAT (Art and Architecture Thesaurus) del Getty, lo storico DDC (Dewey Decimal Classification), IconClass, GeoNames, Wordnet.[30] E poi esistono le authorities della Library of Congress[31] e il progetto

[24]Encoded Archival Description: http://www.loc.gov/ead.

[25]CIDOC - Conceptual Reference Model: http://www.cidoc-crm.org.

[26]Dublin Core: http://dublincore.org.

[27]Simplified Knowledge Organisation Systemhttp: http://www.w3.org/2004/02/skos.

[28]Functional Requirements for Bibliographic Records: http://www.ifla.org/publications/functional-requirements-for-bibliographic-records.

[29]Europeana Data Model (EDM) Documentation: http://pro.europeana.eu/edm-documentation.

[30]Un elenco completo dei value vocabularies si può leggere nel report del W3C Incubator Group del 25 ottobre 2011, *Library Linked Data Incubator Group: Datasets, Value Vocabularies, and Metadata Element Sets*: http://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset-20111025.

[31]Library of Congress authority: http://authorities.loc.gov; Library of Congress

VIAF[32] che vogliono proporsi come descrittori univoci, anche in un'ottica linked data. E numerosi sono anche gli aggregatori di vocabolari, ontologie e linked data sets: dal Metadata Registry[33] al LOV,[34] da LOD Cloud[35] ai Semantic Web Search Engines[36] finalizzati al recupero di informazione semanticamente consistente. I principi del semantic web, e di linked data in particolare, si stanno imponendo come modello teorico e tecnologico di riferimento nel settore delle humanities e in particolare delle biblioteche, degli archivi e dei musei allo scopo di allargare le prospettive di interlinking fra risorse prodotte dagli istituti di conservazione (Guerrini e Possemato).[37]

Ovviamente l'esigenza nella rappresentazione di un dominio è usare standard condivisi sulla base delle regole condivise e rendere le descrizioni compatibili con altri domini e quindi altri standard. Grande lavoro sul cross-mapping e su problemi di allineamento si sta facendo (Haslhofer e Klas) e fin dal 1996 la molteplicità di standard di metadati è sentito come un problema (Day). Ma molte questioni sono ancora da risolvere.

Se dal punto di vista di metadati/ontologie e vocabolari il panorama è estremamente eterogeneo, dal punto di vista delle tecnologie, intese come linguaggi formali per la descrizione delle risorse, uno sforzo comune si sta invece registrando. XML, RDF, URI e OWL sono ormai termini comunemente in uso nel settore del digital cultural

Linked Data Service http://id.loc.gov.

[32]Virtual International Authority File: http://viaf.org.

[33]http://metadataregistry.org.

[34]Linked Open Vocabularies: http://lov.okfn.org/dataset/lov.

[35]Linking Open Data Cloud di Ckan: http://datahub.io/group/lodcloud.

[36]Un elenco si può consultare sul wiki del W3C sul semantic web, nell'àmbito delle attività della Task Force su linking open data: http://www.w3.org/wiki/TaskForces/CommunityProjects/LinkingOpenData/SemanticWebSearchEngines.

[37]Come dimostra la bella raccolta di contributi del convegno *Global Interoperability and Linked Data in Libraries* tenutosi a Firenze il 18 e 19 giugno 2012 e i cui atti sono pubblicati da JLIS.it: http://leo.cilea.it/index.php/jlis/issue/view/536.

heritage. Che si produca un'annotazione embedded (p.e. XML/-TEI) o una annotazione stand-off ogni elemento interpretativo, che può diventare un authority record, deve essere identificato univocamente. Le tecnologie del semantic web aiutano a far fronte al problema dell'identificazione univoca e della sua modalità di espressione attraverso il meccanismo degli URIs. A livello URI è possibile attribuire ad ogni entità una serie di informazioni, mettendo in relazione tale entità con altri URIs attraverso asserzioni RDF, che possono anche prevedere l'utilizzo di predicati ontologici esistenti. Sempre attraverso le stesso meccanismo se si dispone degli URIs di altre risorse Web, magari esposti come data sets, è possibile creare relazioni fra gli elementi annotati e le altre risorse che condividono con le prime determinate features. Tale annotazione, che può venire quindi trasformata in questo modo in data set, può essere esposta su Web attraverso grafi RDF e di conseguenza essere visibile ad altri utenti. Anche il testo dell'edizione, esposto come grafo RDF, può essere mostrato, e volendo anche popolato, da altri ricercatori.[38] In questo contesto un ruolo importante ricopre il framework OAC (Open Annotation Collaboration)[39] come strategia per la gestione delle relazioni fra documento e annotazione e per l'interoperabilità fra annotazioni in prospettiva RDF (Barbera et al.). Il processo di estrazione di triple RDF da file, che utilizzano per esempio il vocabolario TEI, attraverso il modello OAC risulta peraltro un ambito di riflessione critica estremamente interessante nell'ambito delle digital humanities (Jordanous, Stanley e Tupman).

[38]Sul ruolo delle tecnologie nell'ambito linked data si vedano guide e tutoriali sul sito: http://linkeddata.org.

[39]Si veda il recente Open Annotation Data Model: http://www.openannotation.org/spec/core.

3 Le fasi del processo

La costruzione di authority files come raccolta di dati controllati che vengono estratti delle edizioni di documenti si scontra con l'importanza del contesto in cui ogni authority compare e quindi con il dominio di riferimento in cui quell'authority può essere ricompreso. Il problema si articola su tre livelli: come descrivere gli elementi dell'annotazione, che possono essere le entries di un authority record; come creare le relazioni fra tali elementi, che può diventare un sistema di approfondimento del concetto di authority come raccolta di dati contestuali; come far dialogare tali elementi e quindi l'edizione, con il WWW attraverso linked data. E quindi come trasformare autority files, che nascono da un contesto testuale e sono relativi ad un dominio, in linked data sets autoesplicativi, coerenti e appropriati e in grado quindi di dialogare con altre risorse correlate. L'informazione che proviene dai testi delle edizioni può fornire importanti concetti che possono essere formalizzati per la costruzione di basi di conoscenza.

Partiamo da un caso di studio per esemplificare il procedimento: un'edizione digitale di una raccolta di lettere manoscritte, conservate in istituzioni archivistiche e in biblioteche nazionali, ricevute e inviate, nel corso del XV secolo, dal/al copista e libraio fiorentino Vespasiano da Bisticci (Tomasi, «L'edizione digitale e la rappresentazione della conoscenza. Un esempio: Vespasiano da Bisticci e le sue lettere»; «Digital editions between embedded markup and external representation. A case study: Vespasiano da Bisticci's Letters»).

3.1 Elementi e valori

Il primo problema in un approccio finalizzato a stabilire descrittori e relativi valori per la creazione di authority files è la selezione dei metadati quindi la scelta di un vocabolario controllato sia a livello

di elementi che di valori. Due sono quindi i piani su cui ragionare: quali element sets è più opportuno utilizzare per esprimere il punto di vista dell'interprete sulla fonte, che rappresenta il modello; quali value vocabularies sono i più appropriati per esprimere il valore di un elemento. Supponiamo di voler esprimere il seguente concetto, o la seguente asserzione, come lo possiamo dedurre da una lettera di Vespasiano da Bisticci a Piero de' Medici:[40]

> Vespasiano da Bisticci ha copiato le Vite di Plutarco per Piero de' Medici a Firenze nel 1441

Possiamo formalizzare il concetto iniziando a scomporne gli elementi costitutivi secondo il modello "who, where, when, what" e impiegando i nomi di elementi, o le denominazioni delle classi, come stabiliti dai più comuni modelli di metadati o ontologie (per esempio i già citati TEI, CIDOC-CRM, DC, EAC-CPF, EDM).[41]

In un approccio finalizzato a ridefinire il ruolo e la funzione di un authority come stringa estratta da un contesto specifico d'uso e relativa ad un altrettanto specifico dominio i problemi riguardano sia la definizione dei nomi delle etichette descrittive che i valori associati.

Da questo esempio è facile dedurre che denominazioni di elementi diversi esprimono in realtà lo stesso concetto (es. "placename"

[40] Supponendo di voler tradurre la forma attestata di nomi di persona, date, luoghi ed eventi in un documento nella corrispettiva forma controllata come stabilita da una authority condivisa.

[41] Senza ambire ovviamente ad una mappatura dei modelli o all'esaustività della rappresentazione. Alcuni valori potrebbero essere suscettibili di ulteriori scomposizioni (e.g. manuscript-of-Plutarchus-Vitae). Peraltro TEI sta lavorando al mapping, come si può leggere sul wiki dedicato all'attività dello Special Interest Group (SIG) sulle ontologie: http://wiki.tei-c.org/index.php/SIG:Ontologies, in particolare TEI su CIDOC-CRM (Eide e Ore). Grande lavoro sul mapping ha poi fatto Europeana per il suo data model, fornendo peraltro linee guida specifiche. Le *Mapping Guidelines* v1.0.1 (del 24.02.2012) si possono consultare all'indirizzo: http://pro.europeana.eu/documents/900548/ea68f42d-32f6-4900-91e9-ef18006d652e.

Elemento/Classe	**Valore**
persname/creator/actor/agent/person	Bisticci_Vespasiano_da
persname/person	Medici_Piero_de
placename/place_appellation/place	Firenze – Florence
date	1441
event	copy-of-codex
object/physical_thing	manuscript-of-Plutarchus-Vitae

e "place_appellation") e che i valori associati non sempre sono formulabili secondo i precetti di un vocabolario controllato (es. un evento). Diciamo che a livello di mapping molte ambiguità terminologiche sono risolvibili, anche se non sempre lo stesso elemento è interpretato esattamente con lo stesso significato dai modelli in uso (e questo deriva principalmente dalle circostanze di implementazione del modello e dal contesto d'impiego, e.g. "actor" in TEI è utilizzato in modo diverso rispetto al CIDOC-CRM).

Per quanto riguarda i valori esistono, come noto, numerosi vocabolari controllati (già menzionati in precedenza: per le persone, ma anche per i titoli e le keywords, ci sono per esempio le authorities della Library of Congress, per i luoghi il database GeoNames, per i soggetti in Italia c'è il nuovo soggettario della BNCF, ma esiste anche Wordnet a livello internazionale, per le date lo standard ISO 8601). Ma non è detto che tali vocabolari siano sufficienti ad esprimere ogni valore associato all'elemento oltre a soddisfare le esigenze di comunità diverse (sul vocabulary alignment, in modo particolare per i soggetti, si veda, storicamente, Doerr).

Come in un sistema di authority attenzione speciale la si qui vuole dedicare al concetto di persona. Si tratta di un elemento su cui numerosi modelli di metadazione hanno riflettuto. Il primo problema nella definizione di un authority file per le persone è la definizione della forma accettata del nome. E su questo problema una volta che ogni progetto dichiari a quale istituzione deputata

a stabilire il controllo d'autorità si rivolge (es. l'Istituto Centrale per il Catalogo Unico per l'Italia o VIAF a livello internazionale) è possibile rivolvere la questione, anche se ci dovrebbe essere condivisione a livello internazionale circa chi debba ricoprire questo ruolo di garante del controllo di autorità. A cui possiamo aggiungere che le forme attestate nei documenti possono fornire utili forme varianti che possono arricchire authorities esistenti.[42]

Ma particolarmente importante è la connotazione del concetto "persona" nei diversi modelli di metadazione. Diciamo che è evidente che un'etichetta come "EDM:agent" o "CIDOC-CRM_E39:actor" determina un'azione della persona ed è quindi altro rispetto a "person". Allo stesso modo "DC:creator" determina una funzione o meglio un ruolo. Particolare attenzione andrà allora prestata alla descrizione del concetto di persona in quanto il ruolo, la funzione e l'azione sono caratteristiche che possono cambiare a seconda del contesto testuale in cui l'entità compare. Ecco quindi che, astraendo, l'authority, come stringa estratta da un concetto espresso dal documento, inizia a configurarsi: la persona identificata ha un ruolo e ha svolto una specifica funzione che ha portato alla realizzazione di qualcosa a favore di un'altra persona in un certo luogo e in un certa data come attestato dalla fonte in cui l'entità compare.

3.2 Relazioni fra elementi o classi

Passiamo quindi dalla riflessione in termini di accoppiate elemento-valore a quella di asserzione in termini soggetto/predicato/oggetto,

[42] Per esempio VIAF attesta diverse forme di Vespasiano da Bisticci (il cui VIAF ID è 76466245 e il permalink http://viaf.org/viaf/76466245): Vespasiano, da Bisticci, 1421-1498; Vespasiano da Bisticci, Fiorentino, 1421-1498; Vespasiano, da Bisticci, ca. 1421-1498; Vespasiano Da Bisticci, Fiorentino; Bisticci, Vespasiano Da. Dalla collezione di lettere in questione desumiamo invece che Vespasiano si firma sempre come "Vespasiano di Filippo".

secondo i precetti di RDF. Ovviamente nel momento in cui si ontologizza la conoscenza alcune classi diventano proprietà e i valori, intesi come risorse, diventano istanze potenzialmente dotate di URIs e quindi univocamente identificabili.

Le relazioni, o meglio la definizione delle proprietà, diventa un modo per esplicitare formalmente le interpretazioni dell'editore critico. La lettura del testo da parte dell'editore comporta quindi la determinazione del sistema di collegamenti. Il contesto in cui una persona, un luogo o una data sono inseriti fa di quell'istanza una fonte di informazioni proprio in quanto contestualizzata rispetto a quella specifica situazione testuale. La stessa istanza potrebbe assumere un valore diverso quando calata in un differente contesto.

A questo problema si aggiunge la modalità della dichiarazione delle proprietà, vale a dire la definizione dei criteri con cui esprimere le relazioni fra gli elementi annotati, ovvero la scelta dei predicati ontologici e la verifica degli esistenti, allo scopo di comprendere se altre ontologie soddisfino i bisogni interpretativi dell'editore critico. Prendiamo un caso semplice. La relazione fra una persona, identificata da un elemento "persname", e associata ad un letterale in vocabolario controllato, e il luogo in cui quella persona è nata, utilizzando gli elementi TEI e la proprietà "birth":

TEI:persname#Bisticci_Vespasiano_da
birth
TEI:placename#Florence

Caso già più particolare potrebbe essere il seguente concetto:

TEI:persname#Bisticci_Vespasiano_da
copyied-where
TEI:placename#Florence

In questo ultimo caso si sta esprimendo una proprietà che collega le stesse istanze precedenti (Vespasiano e Firenze), ad identificare la

relazione fra una persona e un luogo come desunta da uno specifico contesto testuale in cui la proprietà (luogo in cui è avvenuta la copia di un codice) è specifica per l'occorrenza che si vuole documentare. Ma potremmo anche dire (in una linearizzazione non standardizzata):

> actor/person#Bisticci_Vespasiano_da
> copyied-for
> addressee/person#Medici_Piero_de

A specificare anche i ruoli ("actor" e "addressee") che diverse persone hanno in uno specifico contesto in cui accade un determinato evento (una copia effettuata da un individuo per un altro individuo) in un dato momento.

Ovviamente il problema della compatibilità e dell'interscambio fra i modelli concettuali se deve avvenire in termini di classi e sottoclassi deve avvenire anche a livello di predicati. Sarà dunque necessario mappare i predicati utilizzati in uno specifico contesto con i predicati affini utilizzati in altri modelli affinché la collezione sia davvero interoperabile a livello semantico. Il data model proposto da Europeana, il già citato EDM (Doerr et al.), può essere un riferimento, anche perché per sua stessa natura deve confrontarsi con standard di metadati diversi e renderli compatibili attraverso la definizione di uno schema unico condiviso (Peroni, Tomasi e Vitali).

Per quanto riguarda le relazioni fra persone ISAAR-CPF è un buon modello di riferimento. In ISAAR-CPF il concetto di relazione lega fra loro i soggetti produttori (in senso estensivo le persone) ma anche i soggetti produttori con le risorse prodotte. Ogni relazione fra soggetti può essere classificata (es. gerarchica, cronologica, familiare, associativa), descritta (volendo utilizzando anche un vocabolario controllato) e datata (impiegando p.e. una convenzione come ISO 8601). Allo stesso modo le relazioni fra un soggetto e una risorsa possono essere tipizzate, può essere descritta la natura della relazione e

fornita una datazione. EAC-CPF acquisisce le specifiche ISAAR-CPF e propone un "eac:relations" che si basa sul principio degli "agents" come soggetti produttori e dei collegamenti fra soggetti intesi come unità complesse ("entities"), fornendo poi gli strumenti per specificare la funzione della relazione ("functionRelation"), e per determinare e rappresentare relazioni fra soggetti e risorse correlate ("resourceRelation").

Per ragionare in termini di authority records oltre ad EAC-CPF dovrebbero essere seguite le indicazioni di MADS[43] che, nel definire un modello di authority record, insiste sul problema delle relazioni fra persone e RDA[44] che, fra le altre cose, e sulla scorta di FRBR, ragiona sul concetto di persona, sia a livello di attributes, che di relationships.[45] L'authority estratta da un documento diventa quindi un'entità più strutturata che prevede, oltre a forme controllate delle entries, anche la serie delle relazioni necessarie a documentare un contesto. Si inizia così a semantizzare con collegamenti tipizzati che determinano una nuova authority come punto di accesso ai concetti intesi come relazioni fra istanze contestuali, in cui la fonte svolge un ruolo fondamentale nella definizione del concetto.

3.2.1 Relazioni con linked data sets

Affinché authority records così configurati possano essere interoperabili anche a livello semantico è necessario porli in dialogo con la realtà del WWW. Questo significa trasformare le authorities in data sets e rendere questi ultimi pubblicamente disponibili; ma significa anche conoscere ed utilizzare data sets esistenti qualora ci siano

[43] Metadata Authority Description Schema: http://www.loc.gov/standards/mads.

[44] Resource Description & Access: http: http://www.rda-jsc.org/rda.html.

[45] Un bel progetto denominato SNAC (Larson e Janakiraman) è un esempio prototipale di riflessione sul concetto di persona e sulle associazioni: http://socialarchive.iath.virginia.edu. L'accesso al prototipo all'indirizzo: http://socialarchive.iath.virginia.edu/xtf/search.

possibili collegamenti, per aprire il concetto di relazione a quello di contesto esteso, determinato dal collegamento. Ovviamente con RDF e URIs dereferenziabili creare un data sets non è operazione complessa. E la scelta degli URIs può essere fatta consapevolmente impiegando data sets già esistenti e certificati (es. i già citati VIAF per le forme controllate dei nomi, il progetto Gutenberg per autori e testi, LC Linked Data Service per gli authority records, o ancora DBpedia per i nomi e Wordnet per le keywords).[46] Più complesso concettualmente riconoscere che il data set documenta occorrenze relative ad uno specifico dominio e relative ad un determinato contesto testuale in cui un'entità occorre. La complessità deriva dal fatto che se la proprietà "owl:same-as", utilizzata comunemente per definire forme di corrispondenza fra entità, aiuta a documentare l'esistenza di URIs affini, bisogna ricordare che la stessa entità, se calata in un diverso contesto testuale, potrebbe veicolare un diverso concetto.

Certamente non bisogna dimenticare che la vera interoperabilità è determinata dall'impiego di risorse già formalizzate e che la moltiplicazione di URIs relativi alla stessa istanza inficia il processo di dialogo. Quindi certamente creare collegamenti fra una risorsa e la sua forma standardizzata, o acquisirne l'URI (authority control via permalink), è importante, anche se è necessario sia esito di un ragionamento che tiene conto della specificità in cui la risorsa è calata. Ne deriva che il data set prodotto da ogni edizione produce documentazione relativa a istanze contestuali e che quindi la relazione fra data sets è determinata dalla condivisione di un concetto non di

[46]Un elenco completo dei data sets ad oggi disponibili, e dei relativi URL di progetto e URIs di risorse, si può leggere nella sezione della Task Force del W3C SWEO Community Project: *Linking Open Data on the Semantic Web*, http://www.w3.org/wiki/TaskForces/CommunityProjects/LinkingOpenData/DataSets oppure sul già citato Ckan, "a registry of open data and content packages provided by the Open Knowledge Foundation": http://datahub.io.

semplici stringhe.

Certamente tanto più aumenteranno i data sets esistenti e verranno rese disponibili le triple su WWW, aprendo le risorse al dialogo e non mantenendole "siloed", tanto più la rete della conoscenza diventerà efficace. Non bisogna poi dimenticare che se linked data è una modalità di rappresentazione dell'informazione che ambisce alla costruzione di relazioni la comunità del semantic web, e dell'intelligenza artificiale in particolare, coglie ancora dei limiti derivati dall'assenza di una "upper level ontology" che davvero agevoli forme di ragionamento automatico (Jain et al.).

4 Conclusioni

Scopo del presente lavoro è quindi di aprire una strada verso l'edizione digitale come raccolta testuale da cui acquisire dati che possano essere rappresentati come un nuovo modello di authority record, in cui cioè le stringhe annotate ed estratte dai testi pieni delle edizioni diventino punti di accesso al bagaglio informativo trasmesso dai documenti e in cui la fonte dove l'entità appare è determinante a stabilire il significato. In prima battuta le informazioni già etichettate possono essere estratte da testi marcati, che già presentano un primo livello di descrizione e forniscono le entries. Queste ultime diventano un'authority, arricchita con altre entità correlate a diversi livelli, e la relazione rappresenta una nuova authority. Esporre questi dati sotto forma di open data sets garantisce una ricchezza di risorse aggiuntive per l'interscambio; utilizzare data sets certificati per costruire relazioni e collegamenti deve fare i conti con le diverse situazioni in cui le entità occorrono. Il principio del contesto testuale in questa argomentazione, anche secondo le modalità con cui tale espressione viene utilizzata in campo archivistico, è fondamentale per la costruzione di nuove authorities, che documentano il domi-

nio in cui le entità occorrono. E l'interscambio è determinato dalla condivisione di concetti. Il concetto diventa un nuovo strumento per esplorare i contenuti espressi dai documenti, trasformando le authorities in punti di accesso semantici. Questo processo, oltre a valorizzare i documenti digitali, fornisce nuove fonti utili per l'arricchimento di liste di autorità e fornisce una nuova metodologia di esplorazione del full-text dei documenti; l'authority si viene a configurare come un record complesso in cui contesto e dominio determinano nuovi concetti.

Riferimenti bibliografici

Aloia, Nicola, Cesare Concordia e Carlo Meghini. «Europeana v1.0». *Digital Libraries and Archives*. A cura di Maristella Agosti, et al. Vol. 249. Communications in Computer and Information Science. Berlin, Heidelberg: Springer, 2011. 127–129. <http://link.springer.com/content/pdf/10.1007\%2F978-3-642-27302-5_16.pdf#page-1>. (Cit. a p. 25).

Barbera, Michele, et al. «Annotating Digital Libraries and Electronic Editions in a Collaborative and Semantic Perspective». *Digital Libraries and Archives*. Berlin, Heidelberg: Springer, 2012. 45–56. <http://link.springer.com/chapter/10.1007/978-3-642-35834-0_7>. (Cit. a p. 31).

Chang, Chia-Hui, et al. «A Survey of Web Information Extraction Systems». *IEEE Transactions on Knowledge and Data Engineering* 18.10. (2006): 1411–1428. (Cit. a p. 28).

Clement, Tanya. «Knowledge Representation and Digital Scholarly Editions in Theory and Practice». *Journal of the Text Encoding Initiative* 1. DOI: 10.4000/jtei.203. (2011). (Cit. a p. 25).

Coombs, James H., Allen H. Renear e Steven J. DeRose. «Markup systems and the future of scholarly text processing». *Communications of the ACM* 30.11. DOI: 10.1145/32206.32209. (1987): 933–947. (Cit. a p. 21).

Crupi, Gianfranco. «Beyond the Pillars of Hercules: Linked data and cultural heritage». *JLIS.it* 4.1. DOI: 10.4403/jlis.it-8587. (2013). (Cit. a p. 25).

Day, Michael. *Mapping between metadata formats*. 1996. <http://www.ukoln.ac.uk/metadata/interoperability>. (Cit. a p. 30).

D'Iorio, Paolo e Michele Barbera. «Scholarsource: A Digital Infrastructure for the Humanities». *Switching Codes. Thinking through New Technology in the Humanities and the Arts*. A cura di Roderick Coover e Thomas Bartscherer. Chicago: University of Chicago Press, 2011. 61–87. (Cit. a p. 27).

Doerr, Martin. «Semantic Problems of Thesaurus Mapping». *Journal of Digital Information* 1.8. (2001). <http://journals.tdl.org/jodi/index.php/jodi/article/viewArticle/31/32>.

Doerr, Martin, et al. «The Europeana Data Model (EDM)». *Proceedings of 76th IFLA General conference and Assembly*. 2010. <http://conference.ifla.org/past/ifla76/149-doerr-en.pdf>. (Cit. a p. 37).

Eide, Øyvind e Christian-Emil Ore. «TEI and cultural heritage ontologies: Exchange of information?» *Literary and Linguist Computing* 24.2. (2009): 161–172. (Cit. a p. 33).

Guerrini, Mauro. «Global Interoperability and Linked Data in Libraries: Special issue.» *JLIS.it* 4.1. (2013). <http://leo.cilea.it/index.php/jlis/issue/view/536>.

Guerrini, Mauro e Tiziana Possemato. «Linked data: a new alphabet for the semantic web». *JLIS.it* 4.1. DOI: 10.4403/jlis.it-6305. (2013). (Cit. a p. 30).

Haslhofer, Bernhard e Antoine Isaac. «data.europeana.eu - The Europeana Linked Open Data Pilot». *DCMI International Conference on Dublin Core and Metadata Applications*. The Hague, The Netherlands. 2011.

Haslhofer, Bernhard e Wolfgang Klas. «A survey of techniques for achieving metadata interoperability». *ACM Computing Surveys* 42.2. DOI: 10.1145/1667062.1667064. (2010): 7:1–7:37. (Cit. a p. 30).

Jain, Prateek, et al. «Linked Data Is Merely More Data». *Linked Data Meets Artificial Intelligence*. 2010. <http://www.aaai.org/ocs/index.php/SSS/SSS10/paper/view/1130>. (Cit. a p. 40).

Jordanous, Anna, Alan Stanley e Charlotte Tupman. «Contemporary transformation of ancient documents for recording and retrieving maximum information: when one form of markup is not enough». *Balisage: The Markup Conference*. 2012. (Cit. a p. 31).

Larson, Ray R. e Krishna Janakiraman. «Connecting Archival Collections: The Social Networks and Archival Context Project». *Research and Advanced Technology for Digital Libraries*. A cura di Stefan Gradmann, et al. Berlin, Heidelberg: Springer, 2011. 3–14. <http://link.springer.com/chapter/10.1007/978-3-642-24469-8_3>. (Cit. a p. 38).

Lee, Christopher A. «A framework for contextual information in digital collections». *Journal of Documentation* 67.1. DOI: 10.1108/00220411111105470. (2011): 95–143. (Cit. a p. 24).

Mazzini, Silvia e Francesca Ricci. «EAC-CPF Ontology and Linked Archival Data». *Proceedings of the 1st International Workshop on Semantic Digital Archives*. Berlin: CEUR, 2011. (Cit. a p. 24).

Peroni, Silvio, Francesca Tomasi e Fabio Vitali. «Reflecting on the Europeana Data Model». *Digital Libraries and Archives*. A cura di Maristella Agosti, et al. Communications in Computer and Information Science 354. Berlin, Heidelberg: Springer, 2012. 228–240. <http://link.springer.com/chapter/10.1007/978-3-642-35834-0_23>. (Cit. a p. 37).

Pitti, Daniel. «Creator Description: Encoded Archival Context». *Authority control in organizing and accessing information: definition and international experience*. A cura di Arlene G. Taylor, et al. Binghamton N.Y.: Haworth Information Press, 2004. 201–226. (Cit. a p. 24).

Price, Kenneth M. «Edition, Project, Database, Archive, Thematic Research Collection: What's in a Name?» *DHQ* 3.3. (2009). <http://www.digitalhumanities.org/dhq/vol/3/3/000053/000053.html>. (Cit. a p. 28).

Sahle, Patrick. *A catalog of Digital Scholarly Editions*. 2013. (Cit. a p. 26).

Siemens, Ray, et al. «Toward modeling the social edition: An approach to understanding the electronic scholarly edition in the context of new and emerging social media». 27.4. DOI: 10.1093/llc/fqs013. (2012): 445–461. (Cit. a p. 27).

Tomasi, Francesca. «Digital editions between embedded markup and external representation. A case study: Vespasiano da Bisticci's Letters». *Quaderni digilab* 2.1. (2012): 201 –218. <http://digilab-epub.uniroma1.it/index.php/Quaderni_DigiLab/article/view/24>. (Cit. a p. 32).

——. «L'edizione digitale e la rappresentazione della conoscenza. Un esempio: Vespasiano da Bisticci e le sue lettere». *Ecdotica* 9. (2012). in print. (Cit. a p. 32).

FRANCESCA TOMASI, Università di Bologna.
francesca.tomasi@unibo.it

Tomasi, F. "Le edizioni digitali come nuovo modello per dati di autorità concettuali". *JLIS.it*. Vol. 4, n. 2 (Luglio/July 2013): Art: #8808. DOI: 10.4403/jlis.it-8808. Web.

ABSTRACT: Projects related to cultural heritage enhancement are facing a gradual transition from the description of the sources, at the level of metadata, to their digitization. When this heritage is textual a special attention is recognized to digitization as annotated or "marked-up" transcription, having the aim of textual or documentary edition. Each feature of a document that can be element of annotation - and is therefore subject of interpretation - takes the form of an authority data to be analyzed under the different aspects that attest the specific instance of the element in context. Tools of description of resources, as product of context and domain, contribute to transform the edition of a document in a knowledge base. Semantic Web and Linked Data provides the theoretical and technological tools to convert siloed authority files, which represent the conceptual or semantic access points to digital editions, in interoperable resources.

KEYWORDS: Authority control; Digital editions; Linked data; Semantic indexing; Semantic web.

Submission: 2013-03-02
Accettazione: 2013-05-16
Pubblicazione: 2013-07-01

JLIS.it

Il diritto d'autore nell'era digitale: uno studio pilota su comportamenti, percezione sociale e livello di consapevolezza

Simone Aliprandi

1 Introduzione alla ricerca empirica condotta

1.1 Obiettivi della ricerca

La ricerca presentata in queste pagine[1] ha lo scopo generale di fornire lo spunto scientifico per un'analisi della proprietà intellettuale che tenga conto debitamente della visuale delle "persone comuni" e non solo degli operatori del settore (cosiddetti "stakeholders") o

[1] La ricerca deriva da un progetto di dottorato di ricerca da me condotto tra il 2010 e il 2011 per il dottorato in Società dell'informazione dell'Università degli Studi di Milano-Bicocca. I risultati completi e dettagliati, nonché altri articoli connessi allo stesso progetto sono disponibili in modalità open access (e con licenza Creative Commons) al sito web http://www.aliprandi.org/copyrightsurvey. In questo articolo si cercherà di fornire una panoramica introduttiva sugli scopi della ricerca, un inquadramento generale a livello metodologico e un resoconto commentato dei risultati più interessanti emersi.

JLIS.it. Vol. 4, n. 2 (Luglio/July 2013).
DOI: 10.4403/jlis.it-8802

degli esperti. Infatti, sono davvero poche le ricerche attualmente disponibili che si siano poste in quest'ottica. Gran parte di esse, specie quelle commissionate dalle grandi multinazionali della produzione di contenuti creativi o di software, sono state condotte con un chiaro intento di sondare le tendenze del mercato, indagando le opinioni delle persone in quanto consumatori e potenziali acquirenti, e non come individui in senso più neutrale. Da ciò non può che derivare una distorsione nel tipo di dati raccolti, o quantomeno una portata limitata della loro significatività descrittiva (Una panoramica commentata delle varie ricerche pregresse è disponibile in Aliprandi, «Misurare la cosiddetta "pirateria": una rassegna commentata delle principali ricerche empiriche»).

Dunque questa ricerca ha voluto mettersi in un'ottica differente e ha spostato il focus su tre grandi aree di indagine:

- i comportamenti più comuni fra gli utenti della rete, ovvero come solitamente gli utenti si comportano quando devono acquisire o diffondere materiali coperti da diritto d'autore;

- la percezione che gli utenti della rete effettivamente hanno del problema "diritto d'autore", cioè se lo sentono come un problema importante o secondario, come un elemento utile o solamente fastidioso, etc.;

- il livello di consapevolezza degli utenti della rete sui meccanismi e principi che stanno alla base del diritto d'autore, così da capire quanto effettivamente essi siano informati sull'argomento.
  Inoltre la ricerca si è preoccupata di declinare queste tre grandi aree anche sulla base del tipo di attività svolta dagli utenti della rete in materia di contenuti creativi. In altre parole, si è cercato di indagare quali siano le differenze in fatto di comportamenti, percezione e consapevolezza a seconda che le risposte

provenissero da utenti meramente passivi dei contenuti creativi (cioè che frequentano la rete solo per fruire di contenuti creativi), da utenti creativi (cioè che a loro volta immettono in rete contenuti creativi) o da utenti creativi professionali (cioè che immettono in rete contenuti da essi stessi prodotti).

1.2 La struttura del questionario

Lo strumento di rilevazione scelto per lo svolgimento di questa indagine è stato quello del questionario somministrato online con metodo CAWI (Computer Assisted Web Interviewing). Nel questionario ho cercato di condensare gran parte dei nodi tematici emersi durante le mie attività di divulgazione e formazione (ovvero la pregressa fase di osservazione partecipante che ha portato all'ideazione della ricerca), cercando il più possibile di utilizzare un linguaggio allo stesso tempo rigoroso e semplice. È stata questa una delle difficoltà maggiori poiché i più discussi temi in materia di diritto d'autore nell'era digitale vengono spesso trattati dalle persone non addette ai lavori (ovvero di formazione non strettamente giuridica) attraverso l'uso di semplificazioni, così forti che rischiano di snaturare i concetti tecnico-giuridici in esse insiti. Come si può notare anche da una lettura di massima,[2] si tratta di un questionario sostanzioso, composto da un totale di:

- 9 informazioni di carattere demografico (Sezione 1);
- 35 item che tutti i rispondenti sono chiamati a riempire affinché il questionario possa ritenersi completo;

[2] Il questionario integrale e una sua descrizione più dettagliata sono disponibili nel sito ufficiale della ricerca all'indirizzo http://copyrightsurvey.blogspot.it/2012/08/struttura-ricerca.html (dove, tra l'altro, è disponibile anche la sua versione inglese).

- altri 10 item da riempire solo eventualmente, qualora il rispondente ricada in specifiche categorie di utente della rete (grazie ad alcune domande filtro).

Dalla lettura del testo completo del questionario emerge che esso è stato strutturato in modo da rispecchiare le tre aree di indagine evidenziate nel paragrafo precedente. Dopo una Sezione 1 inerente alle informazioni demografiche sul rispondente, troviamo quindi una Sezione 2 dedicata ai comportamenti, una Sezione 3 dedicata ad opinioni e percezione e una Sezione 4 dedicata al livello di consapevolezza. Come anticipato, nella Sezione 5, dedicata ad approfondire le tre aree in relazione con il tipo di attività svolta in rete, le domande sono state impostate in modo da filtrare in più *step* i rispondenti, così da poterli raggruppare in sottocategorie, secondo il diagramma rappresentato in figura 1 a fronte.

Ne consegue che le quattro sottocategorie generate dal sistema dei filtri si configurano in questo modo:

categoria 0: utenti generici (passivi), cioè coloro che non pubblicano in rete contenuti creativi e sono quindi semplici fruitori;

categoria 1: utenti attivi, cioè che oltre a scaricare contenuti dalla rete li immettono, indipendentemente che si tratti di contenuti originali (creati da loro) o semplicemente creati da altri e semplicemente ripubblicati;

categoria 2: utenti creativi semplici , cioè coloro che immettono contenuti effettivamente creati da loro;

categoria 3: utenti creativi professionali, cioè coloro che creano e immettono contenuti in rete in un'ottica professionale (nel senso che la pubblicazione e condivisione online è effettivamente parte della loro attività lavorativa).

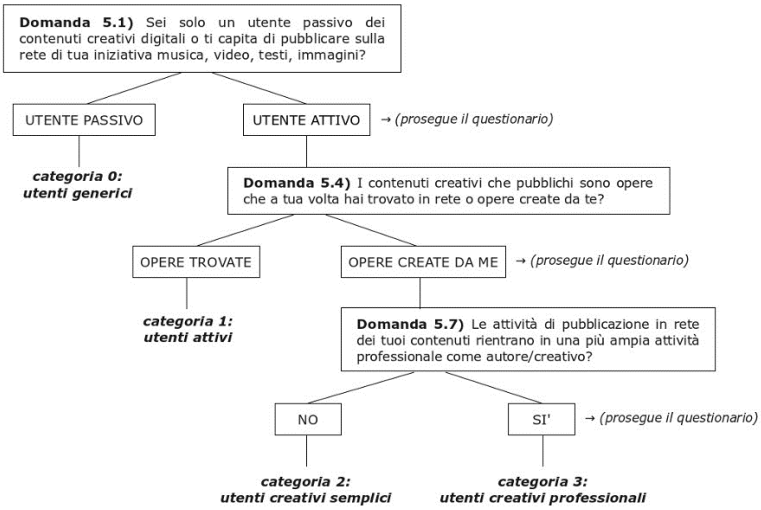


Figura 1: Diagramma esplicativo del sistema di domande-filtro per la profilazione dei rispondenti.

2 Informazioni sui rispondenti

2.1 Risposte totali

Le risposte utili ammontano complessivamente a 1735, di cui 1289 relative allo Studio 1 (Italia) e 446 relative allo Studio 2 (estero). In questo articolo si tratteranno unicamente i dati raccolti per lo Studio 1 (Italia), i quali per vari motivi (principalmente la dimensione del gruppo dei rispondenti) denotano maggior rappresentatività e significatività statistica. Da punto di vista della suddivisione per genere, lo Studio 1 vede una percentuale di 57.3% (739) di maschi e di 42.7% (550) di femmine.

2.2 Area geografica

Utilizzando le informazioni raccolte relativamente alla regione di residenza, ho proceduto alla divisione delle risposte nelle tre classiche aree Nord, Centro e Sud. Il sottoinsieme Nord conta un totale di 933 risposte utili (72,4%) con una buona concentrazione delle risposte provenienti dalla Lombardia centrale. Il sottoinsieme Centro conta un totale di 213 risposte utili (16,5%) principalmente provenienti dalle province di Roma e Firenze. Il sottoinsieme Sud (comprendente anche le due isole maggiori e l'Abruzzo) conta un totale di 143 risposte utili (11,1%) e non presenta aree di particolare concentrazione.

2.3 Fascia di età

1. sotto i 18 anni: 52 (4.0%)
2. tra i 18 e i 24 anni: 396 (30.7%)
3. tra i 25 e i 34 anni: 363 (28.2%)

4. tra i 35 e i 44 anni: 252 (19.6%)
5. tra i 45 e i 54 anni: 156 (12.1%)
6. tra i 55 e i 64 anni: 57 (4.4%)
7. oltre i 64: 13 (1.0%)

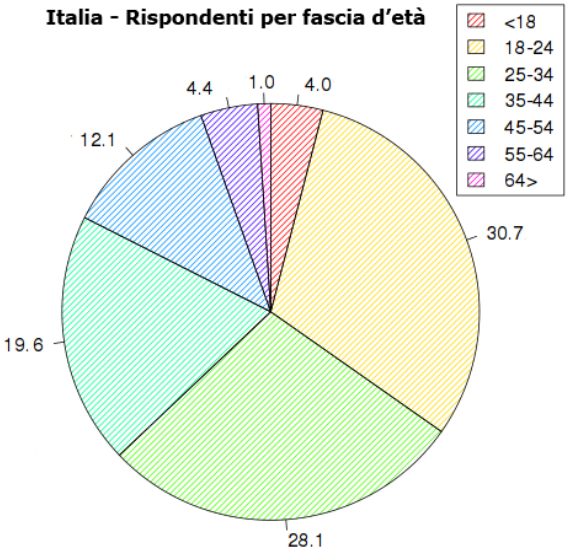


Figura 2: Rispondenti per fascia di età.

2.4 Titolo di studio

1. scuola primaria: 0 (0.0%)
2. scuola media: 119 (9.2%)

3. scuola superiore: 550 (42.7%)
4. università — primo livello: 150 (11.6%)
5. università — secondo livello: 325 (25.2%)
6. post lauream (master/dottorato): 145 (11.3%)

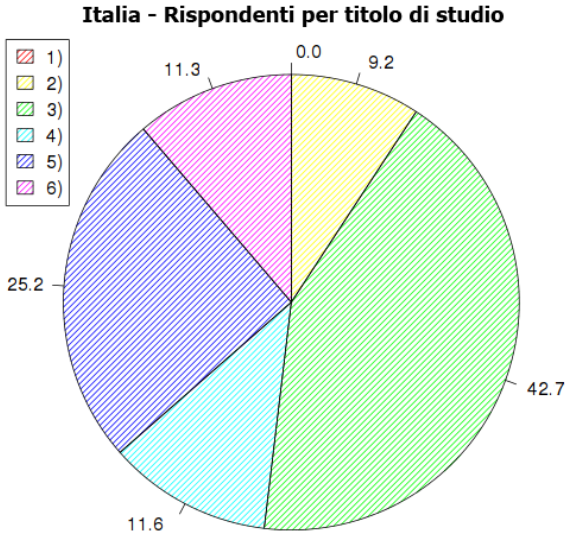


Figura 3: Rispondenti per titolo di studio.

2.5 Occupazione

1. studente: 517 (40.1%)
2. operaio: 30 (2.3%)

3. impiegato (settore pubblico): 179 (13.9%)
4. impiegato (settore privato): 196 (15.2%)
5. dirigente/manager: 33 (2.6%)
6. libero professionista: 207 (16.1%)
7. imprenditore: 33 (2.6%)
8. casalingo: 1 (0.1%)
9. attualmente non occupato: 68 (5.3%)
10. pensionato: 25 (1.9%)

2.6 Tipologia di utenti

Infine, dopo queste suddivisioni basate sulle risposte fornite nella prima sezione del questionario (Informazioni demografiche), se ne può fornire una ulteriore creata a posteriori grazie al sistema di domande filtro della Sezione 5 del questionario, dove — come già spiegato nei paragrafi precedenti — una serie di domande filtro permettevano di individuare sottoinsiemi di intervistati accomunati da specifici comportamenti e atteggiamenti. Lo Studio 1 vede una distribuzione delle risposte utili per tipologia di utenti di questo tipo:

1. utenti generici: 589 (45,7%)
2. utenti attivi: 238 (18,5%)
3. utenti creativi: 294 (22,8%)
4. utenti creativi professionali: 168 (13,0%)

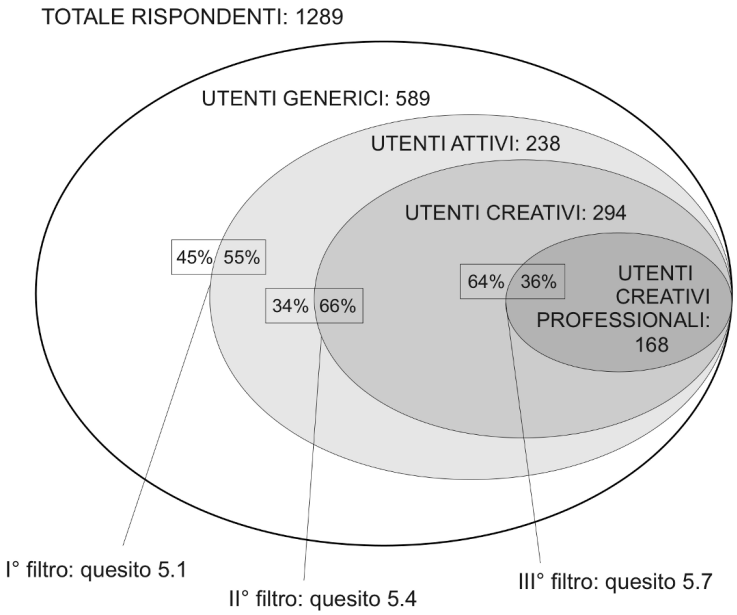


Figura 4: Il gruppo di rispondenti italiano, rappresentato per sottoinsiemi secondo le tipologie di utenti. A cavallo di ciascun sottoinsieme sono indicate le percentuali parziali relative alle risposte fornite alle domande filtro.

3 Risultati più interessanti della ricerca

3.1 Comportamenti (sezione 2 del questionario

La Sezione 2 del questionario ha appunto lo scopo di monitorare i comportamenti più frequenti messi in atto dagli utenti. Si tratta in realtà della parte meno originale della ricerca, dato che moltissime sono le ricerche che si sono occupate di questi aspetti. Tuttavia l'utilità dei dati relativi a questa sezione è legata al fatto che i quesiti, benché riferiti a comportamenti già sufficientemente indagati e monitorati, sono stati posti il più possibile in un'ottica nuova e approcciando gli utenti intervistati in maniera neutrale, diversamente da quanto accade in altre ricerche (in particolare quelle condotte da aziende multinazionali o da enti anti-pirateria).

3.1.1 L'uso della rete a fini di fruizione di contenuti creativi

Quello della disponibilità di banda e del tipo di attività che più frequentemente gli utenti svolgono in internet è proprio uno degli aspetti già sufficientemente indagati, costantemente monitorati e su cui disponiamo di dati quantitativi consistenti. Tuttavia mi sembrava opportuno cogliere l'occasione del questionario per isolare e indagare specificamente gli usi che toccano la questione del diritto d'autore in rete. Infatti, se quasi tutti i rispondenti dichiarano di essere ormai connessi stabilmente con banda larga (benché quello di "broadband" sia un concetto abbastanza relativo e passibile di diverse interpretazioni a seconda delle aree geografiche), ciò che mi interessava indagare con il quesito 1.9 era la frequenza e l'intensità dell'uso di internet specificamente per attività di fruizione (download o streaming) di contenuti creativi digitali. Il risultato è stato quello di una acquisizione di contenuti creativi tramite internet molto frequente: giornaliera per il 41% dei rispondenti e addirittura

costante per circa il 15% ("il mio computer è sempre connesso e scarica automaticamente"); quasi il 20% dichiara di farlo qualche volta alla settimana, il 18,5% di farlo qualche volta al mese, infine solo il 5.5% dichiara di non farlo mai.

3.1.2 Modalità di fruizione e acquisizione di contenuti creativi digitali

Il macro-quesito 2.1 (composto da quattro quesiti) ha l'obbiettivo di approfondire le modalità di fruizione considerando quattro ipotesi: il download attraverso sistemi di *file-sharing* e *peer-to-peer*, la fruizione in streaming, l'acquisizione da appositi negozi online, l'acquisizione da amici e conoscenti. La scala a 5 gradi "mai-spesso" consente di fornire indicazioni sulla frequenza di questi comportamenti. Ho scelto di estrarre le risposte totali a ciascuno dei quattro quesiti e rappresentarle in un unico grafico per poterne apprezzare le correlazioni e le proporzioni interne.

Ad un primo sguardo, ciò che balza immediatamente all'occhio è una netta avversione verso la modalità per così dire più "ufficiale", ovvero l'acquisizione attraverso negozi online come ad esempio iTunes, Amazon e simili, che registra una decisa risposta "mai" da parte del 55% dei rispondenti. Invertendo l'ottica, non vi è un'altra modalità di fruizione che sia vistosamente predominante sulle altre. Se sommiamo le risposte "più di qualche volta" e "spesso" arriviamo a circa un 42% per l'acquisizione tramite *file-sharing* e a quasi un 25% per la fruizione tramite streaming. Su risposte come queste è però importante tenere in considerazione un importante filtro di desiderabilità sociale, che può in qualche modo aver affievolito le risposte a favore di pratiche che, benché molto diffuse e comunemente accettate, sono formalmente considerate illecite. Infine è curioso il dato che si raccoglie dall'ultimo dei quattro quesiti da cui pare potersi dedurre che il "prestito" tra amici e conoscenti

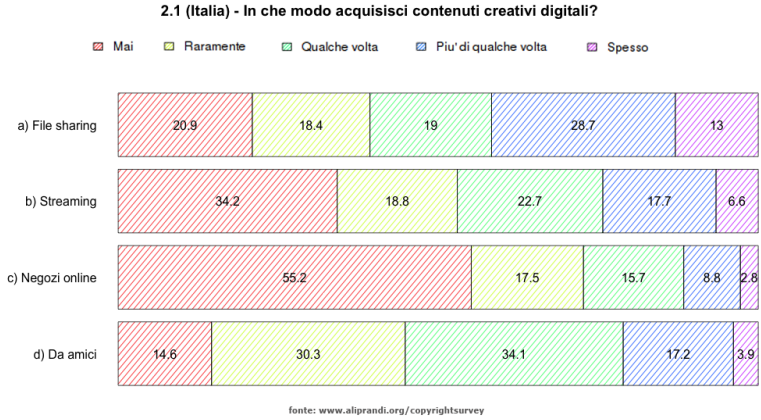


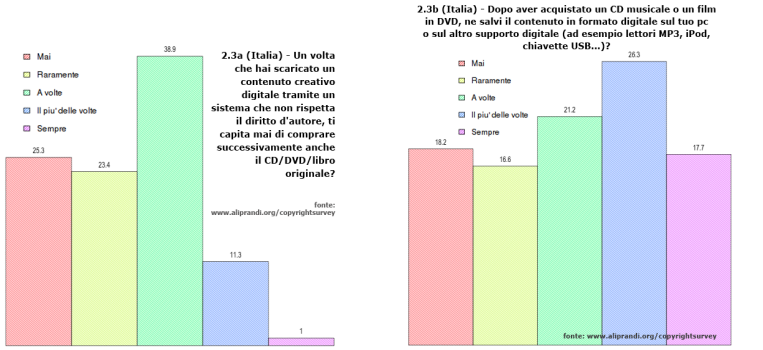
Figura 5: Grafico delle risposte al macro-quesito 2.1.

(attraverso email, supporti ottici, chiavette USB...) è una pratica che sopravvive, nonostante internet per sua natura tenda a metterci in contatto con un numero indefinito di sconosciuti disponibili a condividere contenuti creativi.

3.1.3 Supporto fisico vs file digitale

I quesiti 2.3a e 2.3b sono stati concepiti per indagare due comportamenti che possono essere considerati speculari tra loro. Con il primo si chiede infatti all'utente se, dopo aver acquisito un contenuto digitale con metodi che non rispettano pienamente il copyright, gli capita di comprare anche il supporto fisico originale; con il secondo si chiede invece se, una volta acquistato il supporto fisico originale, vi è l'abitudine di digitalizzare (*ripping*) il contenuto del supporto per "consumarlo" più agevolmente su altri dispositivi. Si noti tuttavia la sostanziale differenza semantica tra i due quesiti.

Il primo infatti presuppone esplicitamente che il comportamento pregresso sia da intendere in qualche modo illecito; il secondo è invece più neutrale e parte da un presupposto lecito per poi muoversi idealmente verso un comportamento solo potenzialmente illecito (la copia privata ad uso personale è infatti un diritto garantito all'utente da quasi tutti i sistemi giuridici). In questa immagine si presentano con due grafici affiancati gli esiti delle risposte totali ai due quesiti.



(a) **(b)**

Figura 6: Grafici delle risposte ai quesiti 2.3a e 2.3b.

Il quesito 2.3a rivela una forte concentrazione delle risposte sull'opzione più neutra "a volte" che tuttavia fornisce già un'indicazione positiva sulla disponibilità degli utenti a comportarsi in quel modo. L'opzione "sempre" rimane invece ad una percentuale quasi nulla. Diversa è invece la situazione per il quesito 2.3b in cui le preferenze sono molto più sbilanciate in direzione positiva (con un 26.3% di "il più delle volte" e un 17.7% di "sempre") ad indicare che quella del *ripping* è una pratica molto frequente, probabilmente proprio perché consente di ampliare le modalità di fruizione dei contenuti.

3.1.4 Modalità di ricerca e fruizione di brani musicali

Con il quesito 2.4 ho cercato di indagare le modalità di ricerca e fruizione di brani musicali. Si noti che la domanda era espressamente impostata in modo da riferirsi ad una fruizione veloce, istantanea, per così dire "usa e getta": nel preambolo si leggeva infatti "ti viene in mente un brano musicale che vorresti ascoltare velocemente. Qual è il primo posto in cui lo cerchi?". Tra le quattro opzioni considerate ve n'è una — la prima — di particolare impatto perché implica che l'utente medio abbia già un hard-disk colmo di brani musicali e che quindi disponga già di tutta la musica di cui potrebbe aver bisogno. Le rimanenti opzioni riguardano invece il download da una rete *peer-to-peer*, la ricerca del brano sotto forma di video musicale su Youtube, il download da canali più "ufficiali" come iTunes o simili. Anche in questo caso le risposte totali sono state disposte in un unico grafico comparativo.

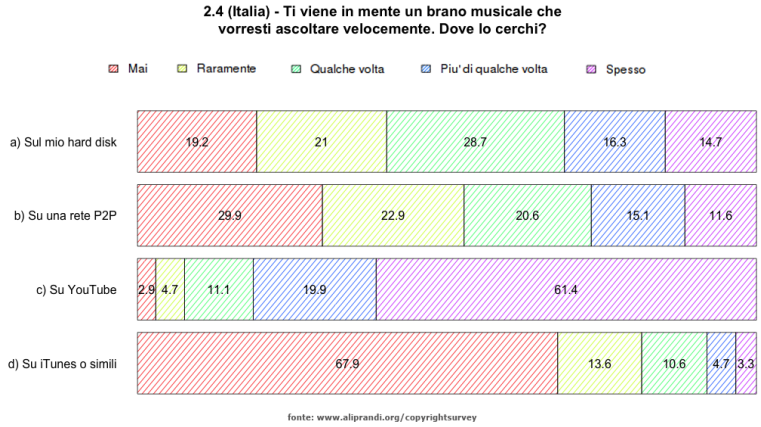


Figura 7: Grafico delle risposte al macro-quesito 2.4.

È lampante a chiunque la specularità delle ultime due barre del grafico. L'opzione YouTube registra più del 61% di risposte "spesso" e meno del 3% di risposte "mai", confermando che YouTube è ormai molto di più di un semplice servizio di hosting in cui gli utenti possono caricare i loro filmati amatoriali. Al contrario l'opzione "iTunes o simili" registra quasi un 68% di risposte "mai" e un 3.3% di risposte "spesso", confermando invece una poca predisposizione da parte dei rispondenti italiani a rifornirsi attraverso canali commerciali (*Discovering behaviors and attitudes related to pirating content*; Aliprandi, «Misurare la cosiddetta "pirateria": una rassegna commentata delle principali ricerche empiriche»). Decisamente più equilibrata la distribuzione delle risposte per le opzioni "sul mio hard-disk" e "in una rete p2p". La prima, con un totale di risposte positive del 31% circa e con un 28.7% di "qualche volta" conferma che in effetti buona parte degli utenti dispone già di una cospicua mole di contenuti creativi digitali nei propri archivi senza quindi avere bisogno di approvvigionarsi ulteriormente dalla rete. La seconda, con un 30% di "mai" e un 23% di "raramente" conferma che le risposte a favore del p2p sono inferiori rispetto a quanto ci si potrebbe aspettare; anche in questo caso bisogna tenere in considerazione eventuali ritrosie nel fornire risposte veritiere in merito ad un comportamento considerato come illecito dall'ordinamento giuridico.

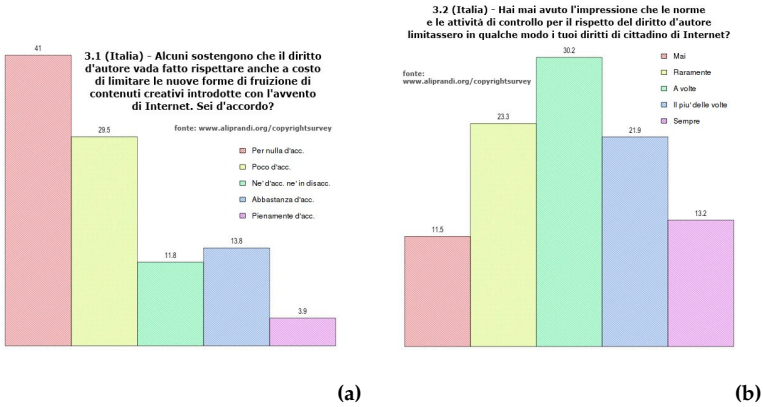
3.2 Opinioni e percezioni (sezione 3 del questionario)

La sezione 3 del questionario rappresenta il fulcro della ricerca condotta poiché si pone come un vero elemento di novità. Sono infatti davvero pochi i casi di ricerche che si siano preoccupate di indagare le opinioni e la percezione degli utenti della rete intesi in senso neutro, quindi senza trattarli da "consumatori" e da potenziali

acquirenti di beni commerciali. Come vedremo nel corso dell'analisi dettagliata, alcuni dei temi trattati da questa sezione sono soggetti ad una distorsione nelle risposte dovuta alla cosiddetta desiderabilità sociale, cioè a quella tendenza a fornire risposte non sempre veritiere bensì idealizzate in quanto più vicine a quelle ritenute socialmente più accettabili e condivise.

3.2.1 Diritto d'autore vs libertà digitali

I quesiti 3.1 e 3.2 si riferiscono rispettivamente al rapporto esistente tra *enforcement* del diritto d'autore e nuove modalità di fruizione del contenuti creativi derivanti da internet, e al rapporto esistente tra enforcement del diritto d'autore e rispetto dei diritti di cittadinanza digitale (sulla cui definizione si sta discutendo proprio negli ultimi anni).



(a) **(b)**

Figura 8: Grafici delle risposte ai quesiti 3.1 e 3.2.

Il primo dei due quesiti mostra una distribuzione piuttosto sbilanciata verso il disaccordo, con un 41% in totale disaccordo e quasi un 30% di parziale disaccordo, a significare che più del 70% dei rispondenti non fa sua l'idea che il diritto d'autore debba essere fatto rispettare anche a costo di limitare le nuove forme di fruizione dei contenuti. Un risultato davvero significativo, che difficilmente si può trovare in altre ricerche empiriche sul diritto d'autore promosse e finanziate da portatori di interessi economici. Il secondo quesito invece mostra risposte più neutre, con una distribuzione a campana in cui il valore centrale (in questo caso "a volte") prende il sopravvento con un 30% lasciando le opzioni estreme al di sotto del 15%.

3.3 La percezione dei comportamenti contrari al diritto d'autore

Il macro-quesito 3.5 con i suoi otto quesiti rappresenta una delle parti centrali di tutta la ricerca empirica e sicuramente la parte più caratterizzante della sezione dedicata ad opinioni e percezione. In tali quesiti si è chiesto agli utenti di esprimere in massima sincerità e libertà la loro opinione in merito ad alcune affermazioni relative alle problematiche più dibattute sul diritto d'autore in ambito digitale. Anche in questo caso la rappresentazione sinottica di tutti gli item aiuta a compiere considerazioni comparative.

I primi tre quesiti ponevano alcuni interrogativi in merito a quale sia l'effettivo danno arrecato dalla diffusa pratica della fruizione di contenuti senza il rispetto del diritto d'autore. Il primo (3.5a), sintetizzabile nell'assunto "questo fenomeno danneggia principalmente gli autori", vede una distribuzione delle risposte praticamente simmetrica, da cui non si può dedurre una reale presa di posizione del totale dei rispondenti che si mostrano divisi a metà tra l'accordo e il

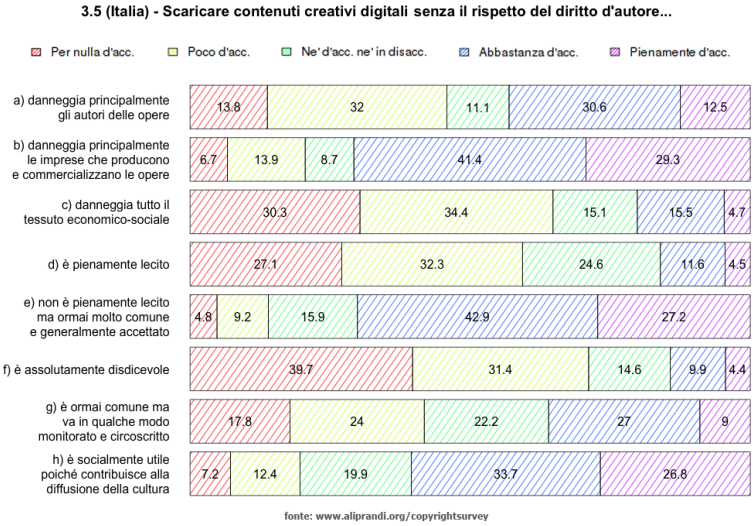


Figura 9: Grafico delle risposte al macro-quesito 3.5.

disaccordo. Al contrario il secondo (3.5b) mostra un netto sbilanciamento delle risposte con più del 70% dei rispondenti che si professano d'accordo con l'idea che tale fenomeno danneggi principalmente le aziende produttrici. Quando invece con l'item 3.5c si pone l'assunto "questo fenomeno danneggia tutto il tessuto economico-sociale" le risposte si distribuiscono in maniera quasi speculare rispetto al precedente quesito, con quasi un 65% di risposte in disaccordo. L'andamento speculare di questi due item è una delle informazioni più interessanti di tutta la ricerca: in sostanza i rispondenti da un lato si dichiarano consapevoli del fatto che la fruizione di contenuti creativi senza il rispetto del diritto d'autore danneggi le aziende produttrici più che i singoli autori, dall'altro rifiutano l'assunto secondo cui tale fenomeno sia un danno per l'intero sistema economico (assunto che per altro fa da *leitmotiv* a quasi tutte le campagne di sensibilizzazione sulla cosiddetta pirateria).

I successivi quesiti (dal 3.5d al 3.5h) indagavano nello specifico la percezione sociale del fenomeno della fruizione di contenuti creativi senza il rispetto del diritto d'autore fornendone alcune qualificazioni e misurandone sempre il grado di accordo. Se l'assunto secondo cui si tratta di "un comportamento pienamente lecito" lascia abbastanza perplessi i rispondenti (che esprimono circa un 60% di disaccordo contro circa un 15% di accordo, dimostrando anche un buon livello di consapevolezza a riguardo), l'assunto seguente secondo cui si tratta di "un comportamento non pienamente lecito ma comunque molto comune e generalmente accettato" raccoglie un amplissimo consenso, con circa il 70% di risposte in accordo e solo il 14% di risposte in disaccordo. Coerentemente, la situazione è invertita sulla qualificazione di tale fenomeno come "assolutamente disdicevole" dove le risposte in disaccordo superano il 71% mentre quelle in accordo raggiungono a stento il 15%. L'assunto 3.5g secondo cui "questo fenomeno, pur essendo abbastanza comune, dovrebbe essere

monitorato e circoscritto" raccoglie anch'esso risposte equamente distribuite tra l'accordo e il disaccordo con una buona percentuale (22%) di indifferenti. Infine vi è il quesito 3.5h con il provocatorio assunto secondo cui "questo comportamento è socialmente utile poiché contribuisce alla diffusione della cultura" e che raccoglie più del 60% di risposte in accordo, segnalando (in maniera nemmeno così implicita) una percezione sociale generalmente favorevole e accondiscendente verso il fenomeno in generale.

3.4 Livello di consapevolezza (sezione 4 del questionario)

Se la sezione del questionario dedicata ad opinioni e percezione sociale subisce l'influenza del fattore "desiderabilità sociale", quella dedicata al livello di consapevolezza subisce l'influenza di quello che molti chiamano "effetto esame", ovvero di quel disagio ed irrigidimento che il rispondente può provare quando i quesiti tendono a mettere in evidenza la sua impreparazione su certi temi; effetto che appunto può portare a distorsioni nella risposta. Come già scritto nelle considerazioni metodologiche, ho cercato di mettere in atto tutte le cautele del caso durante la predisposizione dei quesiti; tuttavia è importante tenere presente questo aspetto nell'analisi e nel commento dei dati.

3.4.1 Livello di curiosità sul tema "diritto d'autore"

Per limitare il rischio della distorsione tipica da "effetto esame", i quesiti 4.1 e 4.3 sono stati modellati su due situazioni tipo che possono fungere da indicatore per misurare il livello di curiosità e interesse verso il tema del copyright. Il quesito 4.1 fa infatti riferimento al caso in cui l'utente installa un software sul PC e vede comparire il testo della licenza d'uso. L'esito è stato che quasi il

50% dei rispondenti dichiara di evitarla completamente scorrendo direttamente alla fine e il 25% dichiara di farla scorrere velocemente leggendo solo i punti che ritiene più importanti; solo l'1.8% dichiara invece di leggerla approfonditamente.

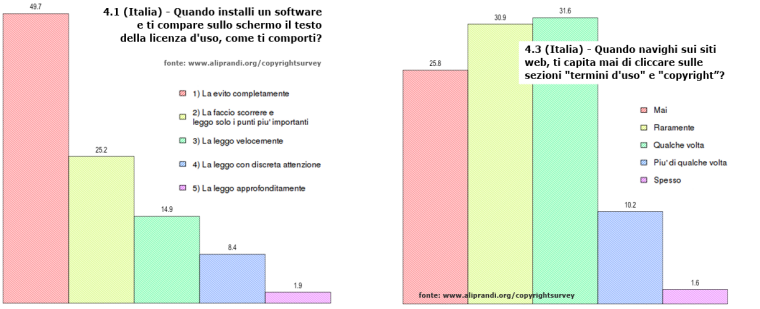


Figura 10: Grafici delle risposte ai quesiti 4.1 e 4.3.

Meno sbilanciate sono le risposte del quesito 4.3 dedicato alla situazione in cui l'utente si trova a navigare nel web e alla frequenza con cui egli si preoccupa di cliccare sulle sezioni dedicate ai termini d'uso e al copyright. Il 25.7% dichiara di non farlo mai e il 30.8% dichiara di farlo solo raramente; tuttavia la maggior parte delle risposte (31.6%) si attesta sull'opzione intermedia "qualche volta". Le risposte raccolte per questi quesiti, specialmente quelle relative al primo dei due, sono sintomo di un basso livello di curiosità verso ciò che attiene al copyright.

3.5 Informazione in materia di diritto d'autore

Il macro-quesito 4.2 (composto di due quesiti) si poneva l'obbiettivo di indagare la propensione dei rispondenti ad informarsi in materia di diritto d'autore. E' ovvio che non tutti gli utenti della rete sono tenuti a sostenere un esame universitario di diritto d'autore, ma bisogna anche considerare che, soprattutto negli ultimi anni, le iniziative di formazione e i canali di informazione su questi argomenti si sono moltiplicate. Dunque mi sembrava interessante misurare quest'aspetto e l'ho fatto chiedendo da un lato quanto spesso è capitato di leggere materiale informativo e dall'altro quanto spesso è capitato di partecipare ad occasioni di formazione sul tema (seminari, lezioni, conferenze). Il risultato è stato molto più favorevole all'ipotesi "materiale informativo" che raccoglie un totale del 40% di risposte positive ("più di qualche volta" e "spesso") e quasi un 28% di risposte intermedie, rispetto a quella di "seminari e lezioni" che invece raccoglie un drastico 54% di risposte "mai" e solo un 6% scarso di risposte "spesso".

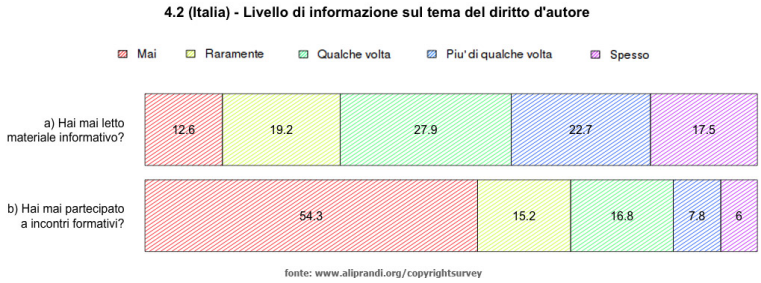


Figura 11: Grafico delle risposte al macro-quesito 4.2.

3.5.1 Effettiva conoscenza dei principi base del diritto d'autore

Gli ultimi due quesiti di questa sezione (4.4 e 4.5) si preoccupavano di misurare l'effettiva conoscenza dei principi base del diritto d'autore. Ovviamente si tratta di un aspetto molto difficile da misurare in un'indagine come questa; soprattutto se si intende farlo nel limite di due semplici quesiti. Inoltre, in questo specifico caso, il rischio di cadere nel già citato "effetto esame" è altissimo. Ciò che ho scelto di fare è stato soffermarmi sui due aspetti che sembrano interessare maggiormente gli utenti di internet e sui quali nello stesso tempo circolano varie "leggende metropolitane": come comportarsi se si intende riutilizzare un contenuto creato da altri (item 4.4) e come si acquisiscono i diritti d'autore su qualcosa che abbiamo creato (4.5). Per ovvie ragioni, le ipotesi di risposta a questi due quesiti non erano rappresentate in una scala di frequenza o di accordo ma indicavano le cinque opzioni che, secondo la mia esperienza di consulente e formatore in materia, vengono maggiormente prese in considerazione.

Il primo quesito è stato in qualche modo "mascherato" come se fosse un quesito sui comportamenti più frequenti (assimilabile quindi a quelli della sezione "comportamenti"), così da poter minimizzare l'influenza dell'effetto esame. Ciononostante, esso celava comunque in sé risposte giuste e risposte sbagliate dal punto di vista giuridico. Infatti, le ipotesi giuridicamente più corrette sarebbero — quantomeno in linea di principio — la n. 4 ("verifico i termini d'uso del sito in cui l'ho trovato e lo uso solo se essi mi autorizzano a farlo") e la n. 5 ("contatto il titolare del sito o direttamente l'autore per chiedergli un'espressa autorizzazione"). Tuttavia quella che raccoglie il maggior consenso (quasi il 49%) è l'ipotesi n. 3 ("lo prendo ma mi preoccupo di citare correttamente la fonte").
Anche in questo caso non si può non rilevare la forte discrasia tra il precetto giuridico e la prassi sociale ritenuta più corretta. Quasi a

significare che il digitale e la rete diventano idealmente luoghi in cui regna una sorta di *fair use* generalizzato, grazie al quale tutto si può utilizzare alla semplice condizione di rispettare il cosiddetto diritto morale al riconoscimento della paternità dell'opera.

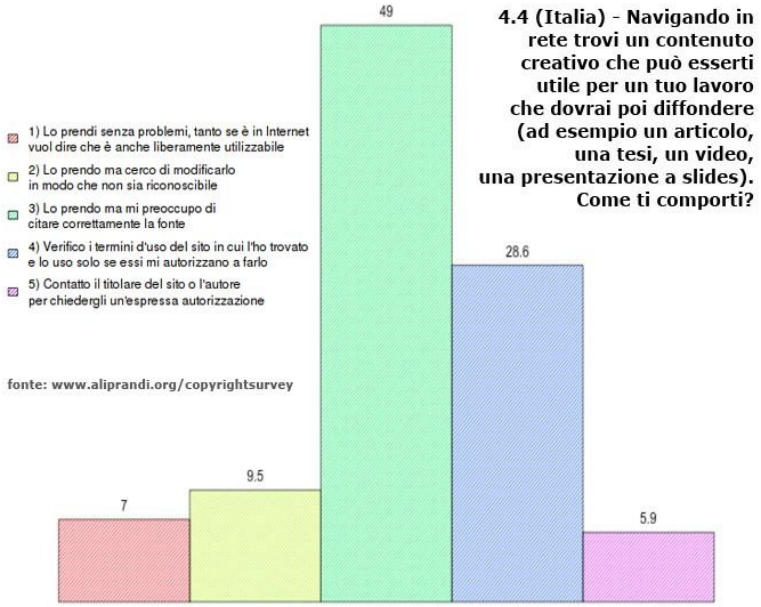


Figura 12: Grafico delle risposte al macro-quesito 4.4.

Passando al secondo dei due quesiti, c'è da osservare che — per esperienza personale — è proprio sulla modalità di acquisizione dei diritti d'autore che circola la più pesante disinformazione. Ed è forse uno degli aspetti più problematici della corretta comprensione delle dinamiche di funzionamento del diritto d'autore. Infatti, se si chiede ai rispondenti dello Studio 1 in quale modo si acquisisca il diritto d'autore sulle proprie creazioni, solo il 18.7% sceglie l'opzione

corretta (ovvero, "in modo automatico, senza fare nulla"), e ben il 22.6% dichiara direttamente di non saper fornire una risposta. Ad ottenere il maggior numero di adesioni (entrambe attorno al 23%) sono proprio le due opzioni che incarnano le più frequenti "leggende metropolitane" in materia di diritto d'autore: cioè quella per cui il diritto d'autore si acquisisca mediante deposito dell'opera presso un apposito ufficio e quelle per cui lo si acquisisca applicando una licenza all'opera. Infine la terza e altrettanto diffusa "leggenda" (quella per cui il copyright si ottenga con l'iscrizione dell'autore ad una *collecting society*) raccoglie il risultato minore con il 12.6%.

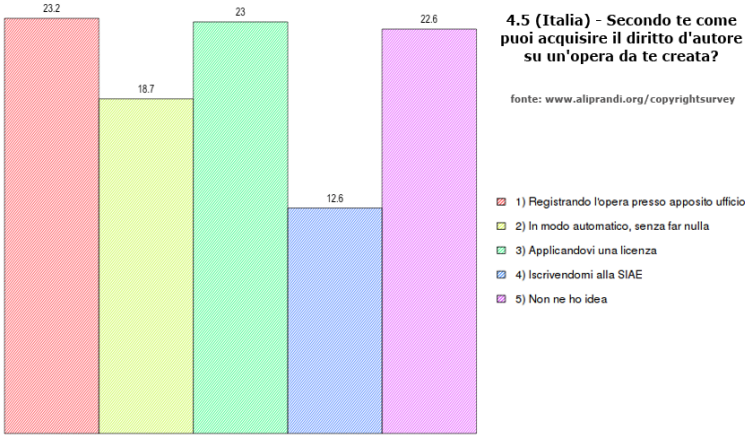


Figura 13: Grafico delle risposte al macro-quesito 4.5.

3.6 Tipologie di utenti: alcune considerazioni specifiche

Come è già stato parzialmente spiegato, l'ultima parte del questionario è stata concepita per creare dei percorsi specifici per le diverse

tipologie di utenti di internet. Questa sezione contiene infatti tre domande filtro, il cui unico scopo è quello di collocare il rispondente in una delle quattro categorie di utenti previste: generici, attivi, creativi e creativi professionali. Oltre alle domande filtro, si trovano in questa sezione alcuni item mirati ad indagare specificamente i comportamenti e gli atteggiamenti del rispondente a seconda della tipologia di appartenenza. E' importante però tenere conto di una questione non irrilevante dal punto di vista statistico: il sistema dei filtri utilizzato, se da un lato permette di creare utili classificazioni tra gruppi di utenti, dall'altro fa sì che il numero di rispondenti diminuisca man mano che il questionario prosegue. Di conseguenza, più ci si avvicina alla fine del questionario e più diminuisce la rappresentatività delle risposte a causa dell'assottigliarsi del numero di rispondenti.

3.6.1 Utenti generici

Analizzando per variabili demografiche le risposte al quesito 5.1, è possibile fornire alcune utili informazioni su com'è composto il gruppo degli utenti generici rispetto a quello degli utenti attivi. Si nota che gli utenti generici sono principalmente femmine (57.8% contro un 35.6% per i maschi) e che l'essere non attivi in rete è una caratteristica propria più delle fasce di età più basse (under 25) e dei soggetti con titolo di studio inferiore. Di riflesso gli utenti attivi sono principalmente maschi e tendenzialmente più adulti e più istruiti.

3.6.2 Utenti attivi

Ai rispondenti che hanno proseguito il questionario è stato chiesto con il quesito 5.2 se sia mai capitato loro che un contenuto caricato in rete fosse rimosso o segnalato poiché violava il diritto d'autore. L'esito è stato circa un 13% di "sì" e un 87% di "no". Successiva-

mente è stato chiesto (quesito 5.3) se il rispondente sfrutta uno o più social network per l'attività di immissione e diffusione in rete dei contenuti creativi. L'esito — abbastanza indicativo del ruolo centrale che il *social networking* svolge nella pubblicazione e diffusione dei contenuti — è stato del 73.4% di risposte positive e del 17.9% di risposte negative; mentre solo l'8.7% ha dichiarato di non fare uso di alcun social network.

3.6.3 Utenti creativi

Il quesito 5.4 è il secondo filtro di questa sezione e aveva lo scopo di dividere gli utenti che effettivamente producono contenuti (utenti creativi) da quelli che semplicemente immettono e diffondono contenuti creati da altri. Per questi ultimi, che sono risultati il 34% degli utenti generalmente attivi, il questionario prevedeva due quesiti mirati ad indagare alcuni sentimenti e preoccupazioni tipici dell'autore della società dell'informazione. Proseguendo nell'analisi, il quesito 5.5 chiedeva di esprimere in una scala di intensità da "per nulla" a "molto" quanto il rispondente si interessi e si preoccupi dell'aspetto della tutela e gestione dei suoi diritti d'autore. Il grafico delle risposte totali fornisce un quadro abbastanza equilibrato tra tutti e cinque gli step della scala, con un lieve sbilanciamento verso l'opzione "poco" (25.8%). Ma ancor più interessante — visto il tipo di domanda — è la visualizzazione grafica per tipologie di utenti. Non a caso, l'andamento delle risposte appare quasi speculare tra "utenti creativi" e "utenti creativi professionali".

Interessante è anche l'esito del quesito 5.6 con il quale si indagava la vera motivazione che spinge gli utenti creativi a svolgere tale attività. Tra le quattro opzioni previste, un netto 52.8% ha dichiarato di farlo per il semplice piacere di potersi esprimere e confrontare con altri creativi; il 24.9% ha dichiarato di essere interessato più che altro alla notorietà e riconoscimento intellettuale come autore; il

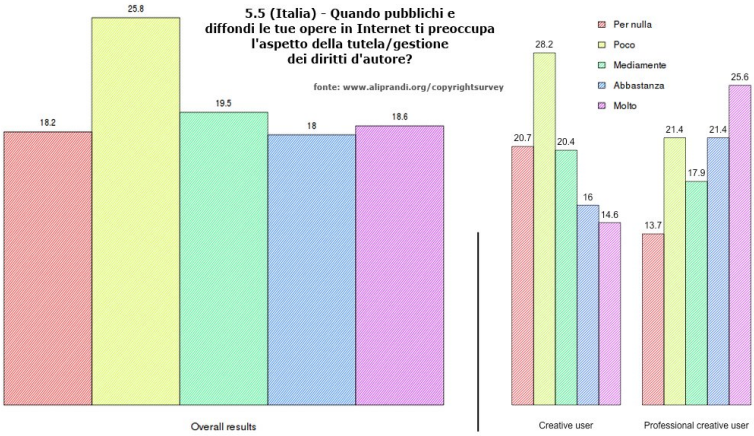


Figura 14: Grafico delle risposte al quesito 5.5. A sinistra si trovano le risposte complessive; a destra invece le risposte sono rappresentate per categorie di utenti (nota: solo gli utenti creativi e creativi professionali sono giunti fino a questo punto del questionario).

20.1% ha invece dichiarato filantropicamente di farlo per arricchire il patrimonio culturale del mondo; e infine solo un irrilevante 2.2% ha dichiarato di mirare ad un ritorno economico diretto o indiretto. Insomma, abbiamo a che fare con una forma di creatività digitale quasi totalmente scevra da logiche di profitto e arricchimento. Questo quesito si presta ad interessanti valutazioni specialmente se analizzato in rapporto alla variabile "tipologia di utente".

Visualizzando i risultati per le due tipologie di utenti che sono arrivate a compilare questa parte della survey (creativi e creativi professionali), si nota infatti un certo sbilanciamento tra le opzioni 1 e 4. Gli utenti creativi hanno scelto per il 15.3% l'opzione 1 ("notorietà e riconoscimento intellettuale come autore") e per il 66.3% l'opzione 4 ("il piacere di potermi esprimere e confrontare con altri creativi

come me"); al contrario gli utenti creativi professionali hanno scelto l'opzione 1 per il 41.7% e l'opzione 4 per 30.9%. Tale esito è in effetti coerente con il profilo delle categorie di utenti; il creativo professionale è più interessato del creativo amatoriale all'ottenimento di un riconoscimento per la sua attività, anche se meramente morale e non monetario.

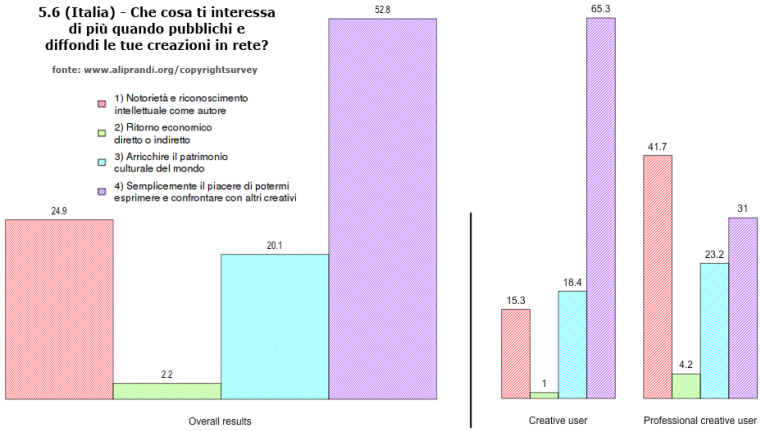


Figura 15: Grafico delle risposte al quesito 5.6. A sinistra si trovano le risposte complessive; a destra invece le risposte sono rappresentate per categorie di utenti (nota: solo gli utenti creativi e creativi professionali sono giunti fino a questo punto del questionario).

3.6.4 Utenti creativi professionali

Un ultimo filtro (posizionato al quesito 5.7) ha separato gli utenti creativi in senso generico da quelli che svolgono tali attività di produzione di contenuti in un'ottica professionale e quindi non solo per hobby personale. Gli utenti che si sono fermati a questo filtro sono circa il 64%; al restante 36% sono stati sottoposti gli ultimi item

del questionario, mirati ad indagare alcuni comportamenti e atteggiamenti strettamente connessi ad un'attività creativa. Essi sono principalmente femmine (42% contro il 34% dei maschi), concentrati nelle fasce d'età adulte e in possesso di un'istruzione tendenzialmente elevata. Il quesito 5.8 (figura 16 nella pagina successiva) chiedeva se per la diffusione dei contenuti prodotti i rispondenti si rivolgano o meno ad un consulente in diritto della proprietà intellettuale per chiarirsi le idee. Ben il 51% ha risposto di non farlo mai e di preferire "arrangiarsi da soli"; e il 22% ha dichiarato di farlo solo raramente. Questo esito è abbastanza indicativo di come, quantomeno in Italia, l'attività di produzione di contenuti, anche se fatta a titolo professionale, si affidi per tre quarti ad una sorta di "fai-da-te" giuridico che di certo non eleva la professionalità del lavoro; e d'altro canto quest'esito ben si sposa con gli esiti dei quesiti relativi al livello di interesse in materia di diritto d'autore.

Passando al quesito 5.9, si chiedeva se l'utente avesse mai subito una violazione dei suoi diritti d'autore e le opzioni previste erano tre: "sì, diverse volte", "sì, ma solo qualche volta" e "no, mai". In generale quasi il 60% dei rispondenti giunti a questa fase del questionario hanno dichiarato di non aver mai subito una violazione; mentre l'opzione "sì, ma solo qualche volta" ha raccolto il 35.7% delle preferenze e la restante opzione "sì, diverse volte" ha raccolto solo il 4.8%.

L'ultimo quesito del questionario (5.10, figura 17 a pagina 78) aveva un carattere più che altro esplorativo e intendeva indagare quale sia la reazione più frequente dell'utente creativo professionale che scopre una violazione dei suoi diritti d'autore. Quasi la metà dei rispondenti ha risposto che solitamente "lascia correre, perché in fondo la cosa non lo mi danneggia più di tanto"; e il 45% che solitamente "cerca di rintracciare il responsabile e di contattarlo per sistemare bonariamente la questione". Solo il 5.5% dichiara di

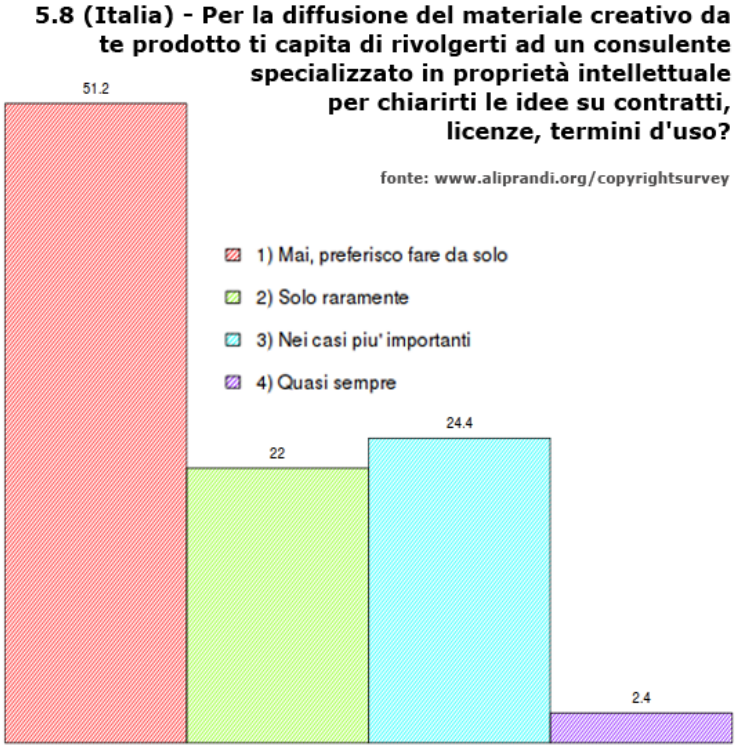


Figura 16: Grafico delle risposte al quesito 5.8.

rivolgersi ad un avvocato per avviare un'azione legale, mentre le altre due opzioni (ovvero "rintraccio il responsabile e lo contatto privatamente richiedendogli subito un risarcimento economico" e "mi rivolgo direttamente alle forze dell'ordine per fare denuncia") hanno raccolto un secco 0%. Emerge dunque un altro dato che di certo non farà piacere agli avvocati: i creativi professionali nel mondo di internet hanno un bassissimo grado di litigiosità e sono piuttosto propensi a risolvere bonariamente e privatamente eventuali problemi di violazione dei loro diritti d'autore.

4 Conclusioni

I risultati emersi dalla ricerca si inseriscono pienamente nel solco dello scenario teorico tracciato dalla ormai fitta letteratura dedicata agli impatti della rivoluzione digitale sul modello di copyright classico. A detta di moltissimi autori,[3] infatti, la cultura della condivisione dei contenuti creativi è senza alcun dubbio parte integrante della società dell'informazione ed è entrata ormai nel DNA degli utenti della rete.

Inoltre i dati confermano tutte le perplessità emerse in varie sedi in merito all'approccio utilizzato dalle numerose ricerche empiriche prese in considerazione. Poca incisività possono avere ricerche condotte con il mero scopo di cogliere gli orientamenti di consumo, dato che ci troviamo in contesti in cui non si può più parlare propriamente e strettamente di "consumo". Ancora minore incisività possono avere ricerche condotte con il sotterraneo scopo di diffondere informazioni distorte in merito al fenomeno del condivisione delle opere protette da diritto d'autore, o ancor peggio a criminalizzare i comportamenti degli utenti.

[3] Si pensi principalmente alle teorie di Lawrence Lessig, di Yochai Benkler e di Philippe Aigrain

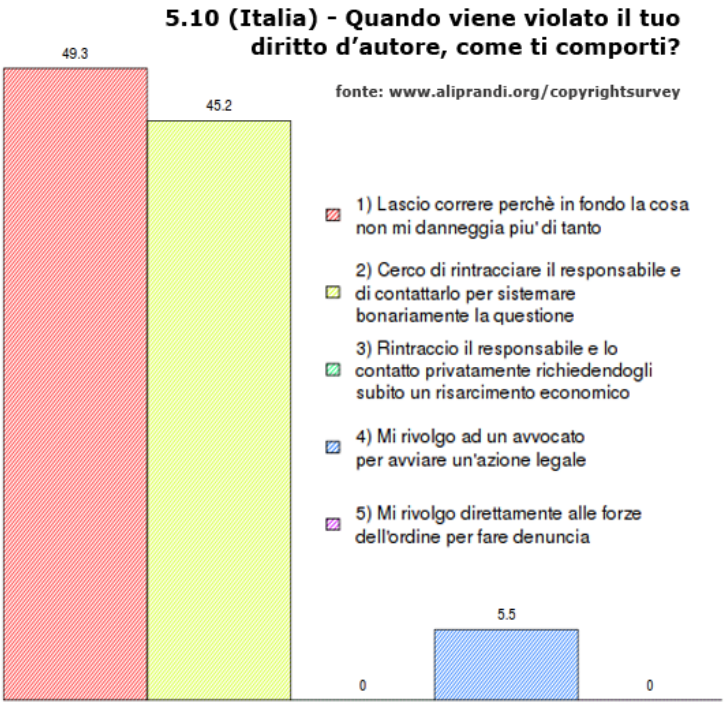


Figura 17: Grafico delle risposte al quesito 5.10.

Dai dati empirici si deduce chiaramente che i comportamenti vanno nella direzione di uno sharing digitale sempre più intenso, costante e — per così dire — "usa e getta". La diffusione di comportamenti divergenti rispetto al modello tradizionale di copyright è direttamente proporzionale all'aumento della disponibilità di tecnologie che appunto consentono tali attività. Nonostante gli interventi normativi degli ultimi 20 anni abbiano cercato di arginare un sostanziale superamento del copyright inteso in senso classico, gli utenti della rete e delle nuove tecnologie digitali trovano ogni anno sempre nuove opportunità per acquisire, diffondere e fruire opere dell'ingegno. Ciò rende quindi anacronistico ogni principio giuridico che non tenga conto di questa continua evoluzione.

Dal punto di vista della percezione, i dati mostrano che i comuni utenti della rete hanno un'opinione non sempre edificante del sistema "diritto d'autore" (inteso nel suo complesso), che viene spesso percepito come un ostacolo alle possibilità di espressione e informazione nella rete, senza significative differenze per fasce di età e titolo di studio. Un simile aspetto può difficilmente emergere in ricerche condotte da enti con forti interessi e legami con l'industria del copyright; salvo qualche raro esempio, i casi da me analizzati in altra sede (si veda l'articolo "Misurare la cosiddetta pirateria" citato in bibliografia) hanno tralasciato questo aspetto, o lo hanno indagato con un approccio di per sé distorsivo, ponendo le domande in modo da far presupporre al rispondente l'indiscussa illiceità di alcuni comportamenti in fatto di sharing di opere. Infine, sul piano del livello di consapevolezza, benché il grado di informazione tenda ad aumentare con l'aumento del titolo di studio in generale e dell'alfabetizzazione tecnologica, gli utenti mostrano di avere ancora le idee abbastanza confuse sui meccanismi che stanno alla base del diritto d'autore. E' anche questo un aspetto connaturato ai nuovi paradigmi di comunicazione portati dalla rete; essa infatti porta anche

i comuni utenti ad essere toccati da aspetti che, fino a pochi anni fa, erano percepiti come strettamente riservati agli addetti ai lavori (tra cui appunto il diritto d'autore, la privacy, la sicurezza informatica). Tuttavia non tutti hanno il tempo, la voglia e la necessaria preparazione per approfondirli, specie in un sistema di comunicazione, qual è quello di internet, dove la tendenza è quella di impigrirsi intellettualmente e di preferire approvvigionarsi di informazioni in modo veloce e superficiale. Opinione di chi scrive è che lo studio del diritto della proprietà intellettuale nonché ogni iniziativa legislativa di sua modifica non può fare a meno di tenere in considerazione opinioni, percezioni e comportamenti più diffusi degli utenti; pena una perpetuazione di quello scollamento tra diritto positivo e norma sociale e una generale nevrotizzazione del sistema giuridico. Un costante monitoraggio dei tre macro-temi trattati da questa ricerca, condotto ovviamente con una approccio il più laico e neutrale possibile, è di certo un buon punto di partenza, dato che permette di capire meglio i punti deboli dell'attuale sistema e cogliere quali possono essere le strade da prendere per proposte di modifica.

Riferimenti bibliografici

Aliprandi, Simone. *Alcuni dati dalla survey sul copyright nell'era digitale*. 2012. http://www.dirittodautore.it/page.asp?mode=News&IDNews=5878.

——.*Diritto d'autore nell'era digitale. Gli aspetti sociologici in un questionario da compilare*. Aduc.it, 2012. http://www.aduc.it/articolo/diritto+autore+nell+era+digitale+aspetti_18733.php.

——.«Il diritto d'autore tra criminalizzazione ed effettività delle norme». *Ciberspazio e diritto* 13.45 (2012). http://www.aliprandi.org/pub/aliprandi_cibersp&dir_2-2012.pdf.

——.«Misurare la cosiddetta "pirateria": una rassegna commentata delle principali ricerche empiriche». *SCIRES-IT* 2.1. DOI: 10.2423/i22394303v2n1p59 (2012). (Cit. alle pp. 46, 60).

Autorità per le Garanzie nelle Comunicazioni. *Il diritto d'autore sulle reti di comunicazione elettronica. Indagine conoscitiva*. 2010. http://www.agcom.it/default.aspx?DocID=3790http://www.agcom.it/default.aspx?DocID=3790.

Bennato, Davide. «L'utente di file-sharing oltre il senso comune». *Sociologia della Comunicazione* (2009).

CIG Customer Insight Group. *The Psychology of Sharing: why do people share online?* New York: The New York TImes, 2011. http://nytmarketing.whsites.net/mediakit/pos.

Cittadini e nuove tecnologie. Roma: ISTAT, 2010. http://www.istat.it/salastampa/comunicati/in_calendario/nuovetec/20101223_00/testointegrale20101223.pdf.

Dei, F. «Tra dono e furto: la condivisione della musica in rete». *Cultura in Italia. Nuovi media, vecchi media*. A cura di M. Santoro. Bologna: Il Mulino, 2008. 49–74.

Discovering behaviors and attitudes related to pirating content. 2011. Online discussion held on october 2011,http://download.pwc.com/ie/pubs/2011_discovering_behaviors_attitudes_related_to_pirating_content.pdf. (Cit. a p. 60).

Eighth annual BSA global software piracy study. Washington, DC: Business Software Alliance, 2010. http://portal.bsa.org/globalpiracy2010.

Gross, Michael. *Online software piracy poll*. Paris: IPSOS, 2004. disponibileonlinesuwww.ipsos-na.com/news-polls/pressrelease.aspx?id=2452.

I comportamenti di consumo di contenuti digitali in Italia. Il caso del file sharing. Roma: Fondazione Luigi Einaudi.

La cultura dell'innovazione in Italia. Rapporto 2009. Milano: Condè Nast, 2009. http://www.cotec.it/it/wp-content/uploads/2009/06/cultura_innovazione_italia_rapporto2009.pdf.

L'import-export dei diritti d'autore per libri, in Italia. Indagine condotta dall'Istituto DOXA per l'Istituto Commercio Estero (I.C.E.) Milano: Doxa, 2004. http://www.aie.it/

Portals / _default / Skede / Allegati / Skeda105-1644-2007.5.2 / 04.03.25%20-%20indagine%20Doxa%20-%20relazione.pdf?IDUNI=41.

P2P Survey 2006. Leipzig: IPOQUE, 2006. http://www.ipoque.com/sites/default/files/mediafiles/documents/p2p-survey-2006.pdf.

SIMONE ALIPRANDI, Progetto Copyleft-Italia.it.
simone.aliprandi@gmail.com
http://www.aliprandi.org

Aliprandi, S. "Il diritto d'autore nell'era digitale: uno studio pilota su comportamenti, percezione sociale e livello di consapevolezza". *JLIS.it*. Vol. 4, n. 2 (Luglio/July 2013): Art: #8802. DOI: 10.4403/jlis.it-8802. Web.

ABSTRACT: This article shows an empiric research within the context of a graduate studies thesis and provides a summary report of obtained results. The research has been carried on through a web-survey which was available online between February and June 2011. The article first explains the research goals and introduces the web-survey. Then main methodological nodes met in this research and methods used to collect and analyze data will be discussed. Then is given the results of the total number of useful responses and their distribution based on demographic variables (geographic area, age, degree, occupation and user typology). Therefore it brings into focus on the results about the Studio 1 – Italia, which are the most representative answers. The results (in few cases showed with bar charts) will be discussed concerning the three macro-themes of the research: behaviours (that is how net users usually acquire creative contents and software); opinions and perceptions (that is how net users relate themselves to the copyright problem, which are their opinions and perceptions about the hottest themes on the subject); consciousness level (in other words what is the real information level and the copyright working principles consciousness of net users and not of experts).

KEYWORDS: Copyright; Surveys; Creative works; Sharing; Social perceptions.

Submission: 2013-03-14
Accettazione: 2013-04-08
Pubblicazione: 2013-07-01

JLIS.it

Verso un nuovo modello di OPAC. Dal recupero dell'informazione alla creazione di conoscenza

Antonella Iacono

Lo scenario nel quale si svolge l'attuale evoluzione del catalogo elettronico è quello della "società dell'informazione", digitale, convergente e pervasiva in cui le condizioni del benessere e del progresso scientifico sono intrinsecamente legate alla gestione consapevole ed efficace del ciclo di vita dell'informazione (Floridi; Hess e Ostrom).[1] Nell'odierno contesto informativo, caratterizzato dall'enorme crescita della quantità delle conoscenze prodotte, assume un'importanza strategica la creazione di strumenti di recupero dell'informazione che siano in grado di creare le migliori condizioni affinché l'informazione possa trasformarsi in conoscenza.[2] Il catalogo, dunque, quale

[1] Il ciclo di vita dell'informazione comprende la creazione, la trasmissione, il processo, la gestione e l'uso dell'informazione stessa attraverso la condivisione, la modificazione, l'indicizzazione ai fini del suo recupero, la conservazione e l'immagazzinamento, ma anche e soprattutto l'apprendimento e l'istruzione (Floridi).

[2] Quest'affermazione ci induce a chiarire la differenza tra informazione e conoscenza, ove per conoscenza si intende il processo di modifica della propria struttura cognitiva, un atto dunque "soggettivo", legato alle esperienze e all'interpretazione da parte di ogni singolo individuo. La conoscenza è ottenibile solo attraverso un processo di rielaborazione dell'informazione ricevuta per inserirla nel proprio bagaglio conoscitivo e, dunque, a differenza dell'informazione che può essere comunicata, trasferita, manipolata, consumata, commercializzata, non è mai oggettiva o misurabile

JLIS.it. Vol. 4, n. 2 (Luglio/July 2013).
DOI: 10.4403/jlis.it-8903

apparato strumentale interpretativo e di mediazione tra l'universo documentario e l'utente, è costretto a misurarsi costantemente con il mutare delle condizioni in cui si realizzano l'accesso e l'uso dell'informazione fortemente influenzate dalle tecnologie digitali. Oggi un nuovo importante cambiamento aspira a rivoluzionare le modalità di creazione, condivisione e utilizzo dell'informazione stessa, riconducendola alla struttura logica primaria di collegamento tra i dati e creando le premesse per un'integrazione delle conoscenze presenti in rete. La costituzione di un nuovo "Web dei dati", di cui le biblioteche cominciano ad intravedere le potenzialità applicative, è in grado di creare effetti di enorme portata sulla struttura del catalogo elettronico, di ampliare le potenzialità della ricerca e favorire la costruzione di nuovi servizi all'utenza basati sui dati. Le relazioni tra dati bibliografici e di altra natura, rese possibili dall'adozione dei linked data, nuova tecnologia del Web semantico,[3] consentiranno alle biblioteche di produrre cataloghi profondamente integrati con il resto del Web, imprimendo una svolta decisiva alla struttura dei record catalografici, alle modalità di accesso al catalogo e alle funzioni che esso potrà svolgere nel più ampio spazio globale dell'informazione. Di fronte a un panorama di forte cambiamento, questo contributo offre una riflessione sull'evoluzione dei cataloghi elettronici interrogandosi sulla loro capacità di rispondere ai bisogni informativi degli utenti. La trattazione è articolata in due parti distinte; in questa prima parte si indagheranno i recenti sviluppi che

(Serrai; Floridi; Case; Salarelli; «Information»; Capurro e Hjørland). Nelle pagine che seguono ci riferiremo al concetto di informazione in un senso ampio intendendo per "informazione" quel concetto, già individuato da Bateson, di "differenza percepita" o di riconoscimento di schemi nel mondo che ci circonda, che si può trasmettere, organizzare, distinguendola dalla "conoscenza" che invece in quanto profondamente soggettiva non si è in grado di gestire o misurare, ma si può solo facilitare tramite un'organizzazione appropriata dell'informazione.

[3] http://linkeddata.org/. Per approfondire si veda Heat e Bizer («Synthesis Lectures on the Semantic Web: Theory and Technology»).

hanno condotto ai *next generation catalogs* e ai *discovery tools*. L'analisi si concentrerà sul processo di ricerca dell'informazione che l'utente svolge nel catalogo, individuando le attuali criticità. Accogliendo gli spunti provenienti dagli studi cognitivi del recupero dell'informazione si proporrà un modello teorico di sviluppo fondato sui modelli comportamentali della ricerca informativa e incentrato su un'analisi approfondita del processo di ricerca che si svolge nell'interazione tra utente e OPAC. Nella seconda parte, di prossima pubblicazione, si esplorerà la possibilità che i *linked data* possano essere la tecnologia più appropriata per la costruzione di nuovi OPAC basati sulla creazione di conoscenza all'interno del processo informativo. Verrà, dunque, valutato il loro impatto in relazione alle principali attività che l'utente svolge nel processo di ricerca dell'informazione nell'OPAC al fine di progettare nuove funzionalità che consentano di migliorare la ricerca.

1 L'OPAC e il modello di sviluppo attuale: *next generation catalogs* e *discovery tools*: dalla ricerca locale alla scoperta globale.

Non è superflua una riflessione che si interroghi sui modelli che attualmente ispirano lo sviluppo del catalogo elettronico, sulle funzionalità e le qualità peculiari che lo rendono riconoscibile e lo distinguono dagli altri strumenti di recupero dell'informazione disponibili nel web. Come è noto, l'OPAC sta attraversando una fase di grande evoluzione che ne ha ridefinito profondamente la fisionomia: il suo sviluppo è fortemente legato al passaggio dalla gestione fisica dei documenti a quella digitale; ciò ha comportato il declino della concezione del catalogo come principale strumento informativo della biblioteca, poiché di fronte alla crescita delle risorse informative

il segmento di informazione, tradizionalmente rappresentato nel catalogo elettronico e limitato alle sole risorse di un'istituzione, perde inevitabilmente di interesse. Se, dunque, in passato il catalogo elettronico rappresentava lo strumento che descriveva ciò che era disponibile localmente, oggi si assiste al contrario ad una predisposizione di un ambiente di ricerca nel quale il catalogo è solo una delle componenti (Dempsey). Il risultato è lo scivolamento verso una rischiosa "perdita d'identità" del catalogo in favore di strumenti che vengono approntati all'interno delle biblioteche (sistemi per la scoperta o *web scale discovery services*) e al di fuori (altri strumenti di ricerca presenti nel Web). Da qualche anno *next generation catalogs* e *discovery tools* sostituiscono i cataloghi tradizionali delle biblioteche.[4] Questi prodotti rappresentano il punto d'arrivo di un lungo percorso di rinnovamento dei cataloghi elettronici che si è svolto a partire dalla metà del decennio scorso, quando si è avviato un corposo dibattito sulle funzionalità degli OPAC che ha visto autorevoli studiosi confrontarsi sulla possibilità di rinnovare i cataloghi arricchendoli e dotandoli di funzionalità più avanzate. Il ripensamento del modello di sviluppo dei cataloghi, oltre ad essere al centro di un dibattito internazionale molto intenso, è stato anche oggetto delle politiche bibliotecarie di importanti istituzioni che hanno ridefinito obiettivi e pratiche della catalogazione. Il quadro è ulteriormente arricchito dagli sviluppi originati dal ripensamento complessivo di principi e funzioni del catalogo avviate dallo studio Functional Requirements for Bibliographic Records (FRBR) che rappresenta ad oggi la riflessione più matura sulla natura del catalogo e le sue funzioni e un

[4]Per un inquadramento generale sulle tematiche dell'evoluzione dei cataloghi elettronici in OPAC di nuova generazione e strumenti per la scoperta si vedano Breeding e Vaughan (*Next-Gen Library Catalogs*; *Web Scale Discovery Services*); per una breve introduzione alla tematica si veda il recente volume di Marchitelli e Frigimelica (*OPAC*); la vastissima letteratura sull'argomento è stata di recente condensata in Ceroti («Rassegna critica della letteratura scientifica italiana sugli OPAC»).

solido modello di rappresentazione delle entità che compongono l'universo bibliografico.[5]

Il modello di sviluppo in uso in questi nuovi strumenti risponde a due principali necessità: quella di offrire all'utente una ricerca semplificata adottando paradigmi più vicini alla ricerca nei motori del web e fornendo servizi aggiuntivi all'utente e quella di gestire le risorse elettroniche e digitali in crescente aumento. Possiamo semplificare affermando che la prima necessità ha condotto allo sviluppo di quelli che oggi sono noti come *next generation catalogs*, mentre la seconda ha generato un'ulteriore evoluzione di questi strumenti nella direzione dei cosiddetti *discovery tools*.[6] In questa rincorsa verso l'accesso globale all'informazione, il catalogo tradizionale è stato sostituito con nuovi strumenti più in linea con le caratteristiche della rete (pervasività, integrazione e semplicità d'uso); tuttavia il modello offerto dai nuovi OPAC presenta ancora alcune significative criticità:

1. i cataloghi elettronici sono ancora lontani da un'adeguata rappresentazione delle entità che compongono l'universo bibliografico, poiché non si interviene sulla struttura dei record catalografici che vengono codificati in un formato non più adatto a realizzare la ricchezza e l'espressività offerta dal modello FRBR;

[5]Modello che tuttavia le biblioteche non hanno applicato pienamente alla struttura dei propri cataloghi elettronici in quanto, com'è noto, per la sua implementazione non si interviene sulla struttura del record bibliografico, ma ci si serve di algoritmi per lo schiacciamento di record e la loro visualizzazione per lo più in base all'entità opera (Zhang e Salaba).

[6]La maggior parte dei software sviluppati attualmente, o in via di sviluppo, ricade in questa categoria; tra questi, alcuni software proprietari come WorldCat Local, Primo, Summon, Ebsco discovery and delivery, sviluppati a partire dal 2009. La loro particolarità è la presenza di grandi indici centralizzati e prepopolati di risorse gratuite e a pagamento che hanno come tratto distintivo quello della ricerca globale e onnicomprensiva di tutte le risorse cui la biblioteca ha accesso.

2. i cataloghi si avvalgono di tecnologie non adatte all'apertura dei dati e all'interoperabilità. I dati sono costretti entro griglie di rappresentazione che favoriscono lo scambio esclusivamente all'interno della comunità bibliotecaria e non si integrano con il più ampio spazio delle risorse presenti sul web;

3. la progettazione dei cataloghi ha imboccato la strada dell'imitazione del processo di ricerca tipico della ricerca nel web senza un'adeguata progettazione del processo informativo che l'utente compie quando consulta un catalogo.

Mentre il web si avvia ad un mutamento che rivoluzionerà l'accesso ai dati, lo sviluppo dei cataloghi elettronici ha seguito vie diverse che hanno isolato le biblioteche e i dati prodotti e immagazzinati nei cataloghi elettronici dal resto del web. Un'opportunità di cambiamento viene oggi offerta dal Web Semantico e dalla tecnologia dei *linked data* che consente la produzione di dati aperti, interoperabili e riutilizzabili nel web. I *linked open data* offrono un enorme potenziale per le istituzioni del patrimonio culturale come biblioteche, archivi, musei; i dati bibliografici possono avere un uso più ampio ed essere collegati a dati prodotti da altre istituzioni accrescendo il valore dei dati stessi, formando la creazione di un grafo globale in grado di collegare le risorse culturali tra loro e alle altre risorse del web. Nel quadro così delineato i dati bibliografici non rimarranno "chiusi" negli OPAC, ma saranno disponibili per l'utilizzo in altre applicazioni del Web Semantico. Appoggiare questa nuova modalità di produzione dei dati bibliografici può offrire grandi vantaggi alle biblioteche e ai loro utenti: le biblioteche potranno collegare i loro dati fra di loro, condividere strumenti bibliografici, schemi, ontologie e sistemi di organizzazione delle conoscenze, migliorare il controllo bibliografico e, infine, costruire sui dati nuovi

servizi per i propri utenti.[7] Queste brevi riflessioni ci inducono a concludere che oggi il processo di rinnovamento degli OPAC debba passare per altre strade e non imboccare quelle "scorciatoie" poco praticabili rappresentate dall'approccio algoritmico, dalla chiusura dei dati, dalla riproduzione automatizzata dei processi della conoscenza. La possibilità di preservare la qualità e l'autorevolezza del catalogo elettronico e di garantirne un rinnovato ruolo nel più ampio spazio globale dell'informazione dipende da un'adeguata riprogettazione del catalogo che preveda un ripensamento dei modelli funzionali attuali. Il passaggio fondamentale che distinguerà gli odierni cataloghi da quelli del futuro sarà infatti la possibilità di ottenere tramite questi strumenti un'esperienza di ricerca che dal recupero globale dell'informazione si orienti verso la creazione di conoscenza ottenibile attraverso una struttura rinnovata del record bibliografico e un'adeguata progettazione del processo informativo. E' necessario, dunque, prima di tutto, ripensare i modi in cui avviene la ricerca nel catalogo; ciò comporta la necessità di una progettazione dell'OPAC che si basi su modelli nuovi, verso il superamento della logica del motore di ricerca, della rilevanza algoritmica e che possa fondarsi invece sull'utente, sui suoi bisogni informativi, i suoi comportamenti e sull'analisi delle componenti che entrano in gioco nel processo di ricerca dell'informazione. In questa prospettiva, "cercare' l'informazione bibliografica vuol dire prima di tutto "costruire' un percorso di ricerca e "comprendere" l'informazione recuperata.

[7] A tal fine nell'ottobre 2011 in seno al consorzio W3C si è creato un gruppo dedicato ai dati di natura bibliografica: il W3C Library Linked Data Incubator Group (LLD XG). Il gruppo ha avuto come obiettivo lo studio di fattibilità dei Library linked data (LLD), cioè ha stabilito quali requisiti debbano possedere i dati bibliografici per poter essere interconnessi e pienamente utilizzabili nel web Semantico. I risultati sono stati esposti nel Library Linked Data Incubator Group Final Report all'indirizzo http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/.

2 La ricerca nell'OPAC. Verso un nuovo modello di sviluppo basato sull'analisi del processo informativo per l'apprendimento e la conoscenza.

La ricerca per soggetto, l'esplorazione del catalogo e l'esposizione significativa dei risultati di una ricerca sono a tutt'oggi le principali aree di criticità nell'uso dei moderni "OPAC di nuova generazione' e negli strumenti di *discovery*.[8] Tali carenze sono il risultato di una visione "debole' di sviluppo degli OPAC che ha preso a riferimento modelli inidonei a favorire l'apprendimento e la conoscenza e che si manifesta in alcuni elementi di criticità riguardanti la progettazione dell'ambiente di ricerca:[9] 1) l'arricchimento del record bibliografico, oggi molto comune, da solo è insufficiente a comprendere e valutare l'informazione nel contesto e 2) l'approccio "esplorativo" fornito dalle "faccette" risulta inefficace in quanto è fondato sull'estrazione automatica di dati dai record bibliografici e non, come si dovrebbe, sull'esposizione della struttura delle relazioni bibliografiche tra le entità che compongono l'universo bibliografico rappresentato nel catalogo. Infine, la pericolosa deriva verso lo strumento di *discovery* comporta l'annegamento dei dati dei cataloghi in questi nuovi stru-

[8]Dato che la prerogativa dei *discovery tool* è l'indicizzazione preventiva nell'indice di risorse non omogenee per tipologia e trattamento catalografico provenienti da varie fonti esterne vengono penalizzati proprio gli accessi semantici, sempre meno presenti tra i filtri utili per la navigazione o per il *browsing* di soggetti o classi.

[9]Tale progettazione è legata alle teorie dell'IR classico e imperniata sulla coppia ordinamento per rilevanza / navigazione a faccette; le due funzionalità sono complementari poiché l'ordinamento per rilevanza produce sovente un lungo elenco di risultati che vanno necessariamente filtrati ed è conseguenza di un meccanismo di recupero dell'informazione "ingenuo' basato sul *matching*, cioè sulla corrispondenza algoritmica tra l'interrogazione e le parole presenti nel documento o nei suoi metadati.

menti unificati basati su una presentazione scarsamente strutturata dell'informazione contenuta e, così, incapaci di supportare l'apprendimento dell'informazione nel contesto o fornire mappe semantiche appropriate per la navigazione. Progettare OPAC che siano davvero "di nuova generazione" vuol dire, dunque, prima di tutto, rovesciare questa visione di sviluppo per creare nuovi cataloghi basati sulla facilitazione del processo di conoscenza la quale – riprendendo il pensiero di Svenonius - non deriva dalla quantità di informazione recuperabile, ma dall'intelligenza adoperata nell'organizzare l'informazione. Il primo e più importante rilievo riguarda il modo in cui negli attuali OPAC vengono presentati i risultati delle interrogazioni; nel *relevance ranking* il concetto di rilevanza è modellato sul concetto di circalità (rilevanza topica o *aboutness*) ottenuta mediante procedimenti di tipo algoritmico e basata sull'analisi dei testi: una presunta rilevanza che è molto lontana dal considerare tutti gli altri aspetti che rientrano in tale concetto, che è invece un fenomeno complesso in cui sono coinvolti processi, strutture, sistemi, fenomeni e concetti. Il concetto di rilevanza si è modificato nel tempo grazie all'evoluzione dei modelli di *information retrieval* verso paradigmi più interattivi capaci di rappresentare le diverse variabili legate non solamente al giudizio di rilevanza, ma all'intero processo di recupero dell'informazione.[10] L'analisi più approfondita della nozione di rilevanza dovrebbe, dunque, orientare la progettazione degli OPAC anche verso una rilevanza soggettiva e dinamica che includa anche

[10]Gli studi sulla rilevanza introducono nel concetto di *information retrieval* la componente dinamica, la componente cognitiva e situazionale della ricerca dell'informazione e il concetto del "contesto", che è centrale nell'approccio qualitativo alla ricerca dell'informazione (Saracevic, «Relevance: a review of and a framework for the thinking on the notion in information science»; «Relevance: a review of the literature and a framework for thinking on the notion in information science. Part II: nature and manifestations of relevance»; «Relevance: a review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of relevance»).

la dimensione della *pertinenza*, cioè la corrispondenza del risultato alle necessità informative "soggettive" dell'utente (Biagetti). Affinché ciò avvenga non è necessario progettare algoritmi più sofisticati, ma offrire all'utente le funzionalità più appropriate al compito da svolgere che derivano da un'adeguata analisi del processo di ricerca. L'adozione di questa nuova prospettiva teorica consente di intendere la ricerca di informazione nell'OPAC come un'attività che va ben oltre il recupero dell'informazione (*information retrieval*) e rientra nel più generale "comportamento informativo' che comprende il bisogno informativo, la ricerca, lo scambio e l'uso dell'informazione. Lo studio dei comportamenti di ricerca è da decenni il campo d'indagine delle teorie dell'*information behaviour*, ossia di quel filone di studi che nell'ambito della Scienza dell'informazione propone un approccio qualitativo del recupero dell'informazione, focalizzando l'attenzione sul comportamento dell'utente e mettendo in evidenza il contesto sociale nel quale si compie il processo di ricerca (Bates 2010). Nell'ambito di questo importante campo di studi, che ha impresso una svolta qualitativa nella disciplina, si sono originati vari approcci teorici (o "metateorie"), numerose teorie e un gran numero di modelli di *information seeking behaviour* che consentono di esaminare più a fondo i diversi aspetti del recupero dell'informazione mettendo l'accento su come le persone cercano e usano l'informazione. Tali modelli possono costituire la base per progettare i sistemi non solo più intelligenti, ma anche più "realistici".[11]

[11]Le origini di questo filone di studio risalgono agli anni 70' come reazione alla influente quanto restrittiva della teoria matematica della comunicazione (MTC). Tra i primi e più grandi sostenitori della "svolta cognitiva" vi furono gli studiosi Jesse Shera, Birger Hjørland e Brian Vickery («Information Science»). Il consolidamento di queste teorie nel campo della Scienza dell'Informazione risale alla metà degli anni '80 quando si assiste a un nuovo orientamento degli studi sull'*information retrieval* verso una dimensione sociale e cognitiva che considera nel recupero dell'informazione anche il contesto nel quale avviene, i processi mentali e i comportamenti messi in atto dagli utenti nella ricerca dell'informazione che sono al centro del confronto

Accogliere queste teorie come base per la progettazione degli OPAC consente, infatti, di superare l'attuale visione squisitamente algoritmica della ricerca basata sui sistemi (e sull'analisi dei testi) e di concentrarsi sulla ricerca come processo, poiché l'utente dell'OPAC è coinvolto prima di tutto in un processo informativo che lo conduce dalla definizione del suo bisogno d'informazione fino alla scoperta e al recupero dell'informazione *pertinente*:

> [la ricerca dell'informazione è] una sorta di processo graduale che evolve per fasi successive attraverso l'interazione tra l'utente e il sistema informativo; è un vero e proprio processo di apprendimento: mano a mano che la conoscenza aumenta, cresce anche lo spettro di informazioni che giudichiamo pertinenti e utili. (Vickery 10)[12]

---

tra studiosi di tutto il mondo. Meritano di essere citate per l'ampiezza degli studi prodotti le conferenze Information Seeking in Context (ISIC) che si tengono ogni due anni a partire dal 1996. Dato che il campo della ricerca informativa è multidisciplinare, le conferenze vedono confrontarsi ricercatori di diverse discipline quali la scienza dell'informazione, la gestione dell'informazione, la psicologia, la psicologia sociale, l'informatica, e altre discipline. Nonostante la una vasta mole di scritti e di contributi, nella teoria biblioteconomica queste teorie non trovano a oggi lo spazio che meriterebbero.

[12]La traduzione è a cura dell'autrice. Un riconoscimento implicito di questa funzione di 'apprendimento' nell'uso del catalogo si ritrova a mio parere nell'obiettivo di navigare il catalogo introdotto tra le funzioni del catalogo nei nuovi Principi internazionali di catalogazione, che consente, appunto, attraverso l'implementazione del modello nelle interfacce, di offrire una visualizzazione delle entità oggetto di interesse del lettore e di poter navigare nella rete delle entità correlate, offrendo un'informazione contestualizzata e dunque capace di creare processi di conoscenza. Tale funzione non presente originariamente in FRBR, venne introdotta nei nuovi ICP in seguito all'importante studio di Svenonius (*The intellectual foundation of information organization. Digital libraries and electronic publishing*).

Le "metateorie del comportamento informativo"[13] e alcuni noti modelli di ricerca formulati entro quei paradigmi (Ellis; Kuhlthau; Bates; Godbold; Ingwersen; Saracevic, «Relevance: a review of and a framework for the thinking on the notion in information science») costituiscono un utile orizzonte teorico nel quale si può proporre un modello alternativo di sviluppo per l'OPAC[14] basato su due principali approcci: 1) la teoria cognitiva dell'informazione, che considera nel processo dell'informazione il ruolo chiave del soggetto interpretante e mette in primo piano la struttura cognitiva dell'utente[15] e 2) la teoria costruttivista dell'informazione, che considera il processo di ricerca come un processo di costruzione e i cui ambiti di indagine sono i concetti di bisogno informativo, di strutture della conoscenza, di uso dei sistemi informativi e di recupero dell'informazione per migliorarne l'usabilità e le funzioni.[16] La ricerca nell'OPAC è dunque

[13]Il concetto di metateoria ha affinità e sovrapposizioni con quello di paradigma individuato da Thomas Kuhn, che tuttavia rappresenta un concetto ancora più ampio in quanto racchiude, oltre alle metateorie, anche teorie e la metodologia di un campo disciplinare. Le principali teorie del comportamento informativo sono ripercorse nel saggio introduttivo di Marcia Bates al volume curato da Karen E. Fisher (10-14) e possono essere ricondotte ad alcuni principali approcci: storico, costruttivista, costruzionista o discorsivo analitico, filosofico-analitico, della teoria critica, etnografico, socio-cognitivo, cognitivo, bibliometrico, fisico, ingegneristico, della progettazione centrata sull'utente, evoluzionista.

[14]Per mancanza di spazio in questa sede non è possibile entrare nel merito dei vari aspetti del modello che sono stati oggetto di approfondimento nel mio lavoro di ricerca; verranno presi in considerazione soltanto alcuni aspetti generali che si reputano maggiormente funzionali alla trattazione rimandando per un approfondimento alla lettura dell'elaborato.

[15]Secondo questa metateoria l'informazione può essere considerata come una frattura che produce una differenza nella struttura cognitiva dell'utente ovvero come una «differenza che crea una differenza». Il concetto fu teorizzato da Bateson ed espresso in termini matematici da Bertram C. Brookes. La comprensione dell'informazione avviene quando creatore e interprete utilizzano e comprendono lo stesso sistema di segni, concetto da cui discende la teoria "socio-cognitiva" (Bateson 470).

[16]Secondo tale teoria, che trova le origini nelle discipline della pedagogia e della filosofia, ogni individuo crea la propria realtà in base a modelli mentali che non sono

da intendersi come un'interazione complessa nella quale entrano in gioco non solo l'utente e il sistema, ma tutte le componenti che interagiscono per il tramite dell'interfaccia: 1) gli oggetti informativi (es. testi e la loro rappresentazione nei record bibliografici e le entità bibliografiche rappresentate in FRBR, Functional Requirements for Subject Authority Data (FRSAD), Functional Requirements for Autority Data (FRAD) che sono le principali funzioni-utente delineate in FRBR e negli Statement of International Cataloguing Principles (ICP); 2) lo spazio cognitivo dell'utente formato dal suo bisogno informativo, dall'incertezza, dalle particolarità del compito da svolgere, dal suo interesse personale, dalle esperienze pregresse, dalle sue abitudini di ricerca; 3) le caratteristiche del sistema ossia le funzionalità di ricerca offerte, le particolari caratteristiche del software in uso, la struttura utilizzata per codificare dei dati; 4) gli intermediari coinvolti nel processo di recupero dell'informazione, ad esempio i bibliotecari, e, nei nuovi OPAC sociali, le communities di utenti che si creano attorno al catalogo. Nel processo di ricerca vanno adeguatamente considerati alcuni concetti chiave quali ad esempio: il contesto sociale e culturale in cui il bisogno informativo nasce e si sviluppa (Wilson; Dervin) gli obiettivi e i compiti concreti che svolgono gli utenti (Xie), la struttura cognitiva dell'utente e le conoscenze pregresse (Ingwersen), le strategie di ricerca dell'informazione che gli utenti mettono in campo (Bates; Ellis), gli oggetti informativi/ i testi/ i documenti rappresentati nei sistemi, i requisiti hardware e software, le interfacce dei sistemi di recupero dell'informazione, gli altri intermediari coinvolti nel processo di recupero dell'informazione (Ingwersen). L'utente si rivolge al catalogo attivando la ricerca dell'informazione per colmare un bisogno, un gap

innati, ma derivano dall'esperienza e si trasformano in base ad essa. A questo filone si ascrivono i principali studi sul miglioramento dei sistemi di recupero dell'informazione con un'attenzione alle caratteristiche di usabilità e di personalizzazione delle funzionalità di ricerca (Salarelli).

(Dervin), per risolvere un'incertezza cognitiva o uno "stato anomalo" nella propria conoscenza (Belkin) e compie nelle sue esplorazioni dell'OPAC un percorso di ricerca non lineare, ma dinamico e iterativo; si può supporre che in vari momenti della ricerca egli analizzi il proprio bisogno informativo e decida se continuare la ricerca, cambiare strategia o percorso, reiterare i passi precedentemente fatti, o abbandonare la ricerca (Bates; Godbold). Intendendo dunque la ricerca come un processo di 'costruzione' della conoscenza, il catalogo non deve solo fornire in risposta un set di risultati, ma aiutare l'utente a costruire il proprio personale percorso di ricerca, mettendo a disposizione tutte quelle funzionalità che gli consentano di compiere il proprio processo di "informarsi", cioè di raggiungere, comprendere e usare l'informazione bibliografica (Kuhlthau).

Un OPAC centrato sull'utente dovrebbe quindi offrire le funzionalità che aiutino il lettore a soddisfare le esigenze che manifesta nel corso di una ricerca, cioè supportare le attività compiute da chi si trova allo stadio iniziale, dall'utente che ha già definito l'oggetto della sua ricerca, o da chi invece si trova alla conclusione dell'attività, attivando le funzionalità più appropriate. Vi sono, inoltre, alcuni elementi di rilievo che caratterizzano ulteriormente il processo di ricerca (raffigurato in fig. 1 a pagina 100):

1. il processo di ricerca viene considerato all'interno del "contesto", rappresentato dalle componenti che interagiscono nell'OPAC: l'utente condizionato dal suo "stato" (cognitivo, affettivo e situazionale), il sistema, i documenti, gli intermediari;

2. il processo informativo non consiste in un iter lineare che si compie all'interno di un catalogo elettronico, ma in una ricerca dinamica e iterativa influenzata da più fattori, variabili e barriere che si manifestano nel corso della ricerca e che derivano dall'interazione delle componenti coinvolte;[17]

[17]Vari modelli e teorie che si iscrivono nel filone di studi dell'information be-

3. l'individuazione delle attività legate alle diverse fasi del processo informativo si traduce nel riconoscimento di funzioni-utente più ampie rispetto a quelle già previste da FRBR e che dipendono dal processo e dall'interazione delle componenti in esso coinvolte;

4. tali funzioni - utente (fig. 2 a pagina 102) a loro volta sono alla base dell'analisi delle funzionalità che gli OPAC dovrebbero offrire per il soddisfacimento delle funzioni stesse e dunque per un più efficace recupero dell'informazione.

Allo scopo di analizzare nel dettaglio la ricerca che si svolge in un OPAC è stata presa a riferimento la nota schematizzazione del processo di ricerca (ISP) elaborata da Kuhlthau (*Seeking meaning: a process approach to library and information services*) che offre un modello in grado di illustrare nel dettaglio le fasi in cui si articola la ricerca "attiva" dell'informazione e che, dunque, può essere utilizzato come schema dei tre principali stadi in cui l'utente si trova quando utilizza l'OPAC:[18]

haviour hanno investigato nel dettaglio particolari aspetti del processo di ricerca dell'informazione come l'attivazione del processo (Dervin) e le principali barriere che intervengono all'attivazione e lungo il processo (Wilson; Godbold). Centrale è il concetto di "gap", quale situazione problematica o carenza cognitiva che l'utente vuole colmare attivando e proseguendo nelle fasi del processo informativo (Dervin). L'attivazione del processo di ricerca non è scontata in quanto intervengono una serie di barriere che derivano dallo stato in cui l'utente si trova e nel quale sorge e si sviluppa il suo bisogno d'informazione (Wilson; Godbold). Tali barriere condizionano non solo i meccanismi di attivazione, ma sono presenti durante tutto il processo di ricerca, poiché l'utente è iterativamente coinvolto nella comprensione e nel perfezionamento della sua ricerca; egli dunque modifica continuamente il suo bisogno man mano che procedendo nella ricerca costruisce e comprende il suo focus e modifica lo stato cognitivo, affettivo e situazionale di partenza (Kuhlthau).

[18] La struttura delle fasi del modello di Kuhlthau (*Seeking meaning: a process approach to library and information services*) vengono qui proposte nella versione semplificata elaborata da Vakkari («A theory of the task-based information retrieval process: A summary and generalization of a longitudinal study»).

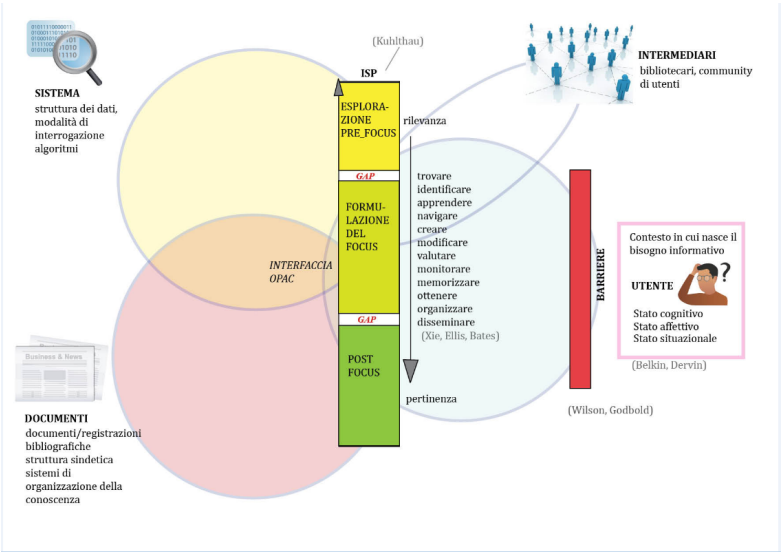


Figura 1: Una rappresentazione delle componenti che entrano in gioco nell'interazione in un OPAC: utente, sistema, documenti, intermediari. Al centro del modello è rappresentato il processo informativo e le principali funzioni che l'utente compie durante il processo. Sono anche presenti le principali barriere che intervengono all'attivazione e nel corso del processo di ricerca.

.

1. la fase precedente alla formulazione del focus, nella quale l'utente è impegnato nell'attività di trovare un argomento specifico per la sua ricerca ed è predisposto verso un'esplorazione del catalogo;

2. la fase della formulazione e del perfezionamento del focus in cui l'utente è in grado di manifestare modalità di ricerca "attive" e porre domande al sistema;

3. la fase successiva alla formulazione del focus ove l'utente compie l'attività di raccolta e organizzazione dell'informazione ricevuta per avviarsi alla conclusione della ricerca (Kuhlthau).

Le fasi enumerate condizionano i modi in cui l'utente cerca e usa l'informazione e le strategie di ricerca adottate, che non sono prevedibili a priori, ma sono ulteriormente condizionate dalla tipologia e complessità del compito da svolgere, dal tempo a disposizione, dell'interesse personale, e delle conoscenze di cui dispone. Nel processo informativo è tuttavia possibile ravvisare un certo numero di attività cognitive "generiche" che vengono svolte dall'utente (Xie; Fattahi):

- riconoscere un bisogno informativo ovvero un gap nella conoscenza;
- focalizzare un argomento di ricerca e identificare l'informazione necessaria;
- esplorare il catalogo / formulare una richiesta al sistema;
- analizzare e valutare la rilevanza/la pertinenza dei risultati ottenuti;
- aggiungere nuova conoscenza alla propria "struttura cognitiva";
- organizzare e riutilizzare l'informazione ottenuta.

La visione costruttivista e cognitiva ci consente, infine, di intendere il processo di ricerca come "creazione di senso" (Dervin), ovvero di comprensione dell'informazione nel suo contesto avvicinandola a quella recentemente prefigurata da Fattahi laddove lo

Attività	Pre focus	Focus	Post focus
***Trovare**	■	■	□
***Identificare**	■	■	□
Apprendere/correlare	■	□	□
***Navigare**	■	□	□
Creare	■	■	□
Modificare	□	■	□
Valutare / (Selezionare*)			
-la *rilevanza*	■	□	○
-la *pertinenza*	○	■	□
Monitorare			
il processo di ricerca	■	■	□
Memorizzare	○	□	■
***Accedere /Ottenere**	■	■	■
Organizzare	□	■	■
Disseminare	○	○	■

■ valore alto □ valore medio ○ valore scarso *Funzioni-utente in FRBR

Figura 2: Le attività dell'utente in relazione alle fasi del processo di ricerca dell'informazione nell'OPAC.

studioso auspica il passaggio dei cataloghi "da sistemi orientati al documento ai sistemi orientati alla conoscenza":[19]

> «il catalogo di biblioteca non è un sistema con il quale l'utente deve acquisire conoscenza a forza di sperimentare, per tentativi ed errori: dovrebbe essere un sistema cognitivo integrato che si autodefinisce e in grado di armonizzare l'ambiente e il comportamento di ricerca informativa dell'utente, al di là e oltre la ricerca d'informazione. Dovrebbe mostrare come è rappresentata, strutturata e visualizzata la conoscenza nel catalogo».

[19]Lo studioso iraniano affida al "super-record" ossia al record della "super-opera" quella funzione organizzativa necessaria alla costruzione di una tale visione di sviluppo dell'OPAC (Fattahi 42).

3 Conclusioni

Le tecnologie digitali hanno cambiato significativamente i modi con cui oggi le persone si approcciano all'informazione e la utilizzano. Tuttavia, ancora oggi i principali miglioramenti apportati agli OPAC sono incentrati sul recupero dell'informazione orientato ai sistemi e basato sull'analisi dei testi, sugli algoritmi, sul *ranking* e sull'ampliamento a livello globale del recupero dell'informazione. Le considerazioni svolte in queste pagine rappresentano il tentativo di analizzare più in profondità il processo di ricerca che si compie nel catalogo e di accogliere in quest'analisi le teorie dell'*information retrieval* cognitivo e orientato all'utente, capace di associare i fattori sociali che influenzano il recupero dell'informazione alle strategie di ricerca e ai comportamenti che gli utenti manifestano dinamicamente nel corso dell'interazione con il catalogo. Un'evoluzione dell'OPAC in questa direzione porta, quindi, a concepire la ricerca non più in termini di corrispondenza tra una *query* e il risultato, ma ad ampliare la visuale all'intero processo informativo in cui i sistemi devono supportare diverse strategie comportamentali dell'utente. Il modello presentato brevemente in questo contributo si offre come strumento di analisi delle componenti e delle variabili coinvolte nel recupero dell'informazione così considerato in una prospettiva olistica, immaginando che nell'interazione che si svolge nell'OPAC, l'utente compia al contempo un processo di ricerca e di apprendimento. Si auspica, quindi, che una maggiore apertura verso questi aspetti del recupero dell'informazione, ancora poco presenti nella progettazione delle interfacce dei cataloghi, possa condurre allo sviluppo di funzionalità in grado di migliorare l'evoluzione futura di questi autorevoli strumenti.

Riferimenti bibliografici

Bates, Marcia J. «The Design of Browsing and Berrypicking Techniques for the Online Search Interfaces». *Online Information Review* 13. DOI: 10.1108/eb024320 (1989): 407–424. (Cit. alle pp. 96–98).

Bates, Marcia J. e Mary Niles Maack, cur. «Information». *Encyclopedia of Library and Information sciences*. (Cit. a p. 86).

——, cur. «Information Science». *Encyclopedia of Library and Information sciences*. (Cit. a p. 94).

Bateson, Gregory. *Verso un'ecologia della mente*. Milano: Adelphi, 2004. (Cit. a p. 96).

Biagetti, Maria Teresa. «Nuove funzionalità degli OPAC e relevance ranking». *Bollettino AIB* 50.4 (2010): 339–356. <http://bollettino.aib.it/article/view/5340>. (Cit. a p. 94).

Breeding, Marshall. *Next-Gen Library Catalogs*. New York: Neal-Schuman Publishers, 2010. (Cit. a p. 88).

Capurro, Rafael e Birger Hjørland. «The concept of information». *Annual Review of Information Science and Technology* 37. DOI: 10.1002/aris.1440370109 (1 2003): 343–411. (Cit. a p. 86).

Case, Donald Owen. *Looking for information: a survey of research on information seeking, needs, and behaviour*. Bingley: Emerald, 2010. (Cit. a p. 86).

Ceroti, Mario. «Rassegna critica della letteratura scientifica italiana sugli OPAC». *Biblioteche Oggi* 30.9 (2012): 15–27. <http://www.biblio.liuc.it/scripts/bibloginj/ricerche.asp?tipo=articolo\&art=4344>. (Cit. a p. 88).

Dempsey, Lorcan. «Thirteen Ways of Looking at Libraries, Discovery, and the Catalog: Scale, Workflow, Attention». *EDUCAUSE Review Online* (2012). http://www.educause.edu/ero/article/thirteen-ways-looking-libraries-discovery-and-catalog-scale-workflow-attention. (Cit. a p. 88).

Dervin, Brenda. «From the mind's eye of the user: the sense-making qualitative-quantitative methodology». *Qualitative research in information management*. A cura di Jack D. Glazier e Ronald R. Powell. (Cit. alle pp. 97–99, 101).

Ellis, David. «A behavioural approach to information retrieval system design». *Journal of Documentation* 45. DOI: 10.1108/eb026843 (1989): 171–212. (Cit. alle pp. 96, 97).

Fattahi, Rahmatollah. *From Information to Knowledge: SuperWorks and the Challenges in the Organization and Representation of the Bibliographic Universe =Dall'informazione alla conoscenza: le super-opere e le sfide dell'organizzazione e rappresentazione dell'universo bibliografico : Lectio magistralis in Biblioteconomia, Università degli Studi di Firenze*. Firenze: Casalini Libri, 2010. (Cit. alle pp. 101, 102).

Fisher, Karen E., Sanda Erdelez e Lynne Mckechnie, cur. *Theories of information behaviour*. Medford: Information Today, 2005. (Cit. a p. 96).

Floridi, Luciano. *La rivoluzione dell'informazione*. Bingley: Emerald, 2010. (Cit. alle pp. 85, 86).

Godbold, Natalia. «Beyond information seeking: towards a general model of information behaviour». *Information Research* (2006). <http://informationr.net/ir/11-4/paper269.html>. (Cit. alle pp. 96, 98, 99).

Heath, Tom e Christian Bizer. «Synthesis Lectures on the Semantic Web: Theory and Technology». *Linked Data: Evolving the Web into a Global Data Space*. A cura di James Hendler e Frank Van Harmelen. (Cit. a p. 86).

Hess, Charlotte e Elinor Ostrom. *La conoscenza come bene comune: dalla teoria alla pratica*. Milano: Mondadori, 2009. (Cit. a p. 85).

Ingwersen, Peter. «Cognitive perspectives of information retrieval interaction». *Journal of Documentation*. DOI: 10.1108/eb026960 (1996): 3–50. (Cit. alle pp. 96, 97).

Kuhlthau, Carol Collier. *Seeking meaning: a process approach to library and information services*. Westport: Libraries Unlimited, 2004. (Cit. alle pp. 96, 98, 99, 101).

Marchitelli, Andrea e Giovanna Frigimelica. *OPAC*. Roma: AIB, 2012. (Cit. a p. 88).

Salarelli, Alberto. *Introduzione alla scienza dell'informazione*. Milano: Bibliografica, 2012. (Cit. alle pp. 86, 97).

Saracevic, Tefko. «Relevance: a review of and a framework for the thinking on the notion in information science». *Journal of the American Society for Information Science*. DOI: 10.1002/asi.4630260604 (1975): 321–343. (Cit. alle pp. 93, 96).

——.«Relevance: a review of the literature and a framework for thinking on the notion in information science. Part II: nature and manifestations of relevance». *Journal of the American Society for Information Science*. DOI: 10.1002/asi.20682 (2007): 1915–1933. (Cit. a p. 93).

——.«Relevance: a review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of relevance». *Journal of the American Society for Information Science*. DOI: 10.1002/asi.v58:13 (2007): 2126–2144. (Cit. a p. 93).

Serrai, Alfredo. *Dalla informazione alla bibliografia: la professione bibliotecaria*. Milano: Bibliografica, 1984. (Cit. a p. 86).

Svenonius, Elaine. *The intellectual foundation of information organization. Digital libraries and electronic publishing*. Cambridge: MIT Press, 2000. (Cit. a p. 95).

Vakkari, Pertti. «A theory of the task-based information retrieval process: A summary and generalization of a longitudinal study». *Journal of Documentation* (2001): 44–60. (Cit. a p. 99).

Vaughan, Jason. *Web Scale Discovery Services*. Chicago: ALA TechSource, 2011. (Cit. a p. 88).

Vickery, Brian C. *On retrieval system theory*. London: Butterworths, 1965. (Cit. a p. 95).

Wilson, Thomas D. «Models in information behaviour research». *Journal of Documentation*. DOI: 10.1108/EUM0000000007145 (1999): 249–270. (Cit. alle pp. 97, 99).

Xie, Iris. *Interactive Information Retrieval in Digital Environments*. New York: IGI Global, 2008. (Cit. alle pp. 97, 101).

Zhang, Yin e Athena Salaba. *Implementing FRBR in libraries: key issues and future directions*. New York: Neal-Schuman, 2009. (Cit. a p. 89).

ANTONELLA IACONO, Università La Sapienza di Roma.
antonella.iacono@fastwebnet.it

Iacono, A. "Verso un nuovo modello di OPAC. Dal recupero dell'informazione alla creazione di conoscenza ". *JLIS.it*. Vol. 4, n. 2 (Luglio/July 2013): Art: #8903. DOI: 10.4403/jlis.it-8903. Web.

ABSTRACT: The paper analyzes the evolution of libraries electronic catalogs evidencing the new technologies of the Semantic Web and *Linked Data*. The essay is divided into two distinct parts. In the first part starting from critical analysis of the new generation of OPAC functional model and discovery tools proposes the need for a revision of current development paradigm that accepting the theories of information behaviour can be based on the user, on its information needs, its behaviours, and on the analysis of components that takes part of search information process. In the second part of next publication will explore the possibility that Linked Data may be the most appropriate technology to build up new OPAC based on output of knowledge within the information process.

KEYWORDS: Electronic catalogue, OPAC, Linked data; semantic web, discovery tools, next generation catalogs, information behaviour.

Submission: 2013-03-14
Accettazione: 2013-04-08
Pubblicazione: 2013-07-01

JLIS.it

Development of a metadata schema describing Institutional Repository content objects enhanced by "LODE-BD" strategies

Iryna Solodovnik

The move toward Linked Data will be the most significant change in library data in these two centuries
(J. Zaino, *The future of libraries*)[1]

1 Introduction

Issues like handling metadata, cross-referencing them consistently with authority control and semantic vocabularies, licensing activities valorizing scope and usage of digital resources within Institutional Repository (IR) infrastructures will become certainly increasingly challenging in the future years. It is due to emerging models and actualities like: Repositories of research data and Data Management in research infrastructures; interoperability among Repositories, with CRIS (Current Research Information Systems), and external services and applications; capturing research context in connection to re-

[1] http://semanticweb.com/the-future-of-libraries-linked-data-and-schema-org-extensions_b35315 .

JLIS.it. Vol. 4, n. 2 (Luglio/July 2013).
DOI: 10.4403/jlis.it-8792

search output by Service Providers; application trends of Semantic Web technologies service-oriented frameworks for bibliographic data; metadata management across disciplines with wide re-use of Repository data and services, as well as necessity of reliable value-added services over Trusted digital Repositories.

With this in mind, due attention must be paid for the development of qualitative and updated – according to current standards, guidelines and best practices – metadata application profiles supported by standard and "good practices" compilation and encoding strategies.

To provide more visible and sharable data on the web, different communities are aligning their digital contents according to current best practices for publishing and consuming data on the web, formalized within Linked Data (LD, Web of Data, Web 3.0) paradigm, the first practical expression of the Semantic Web – declared useful, feasible and applicable to all forms of data. Digital contents published as LD sets are presented graphically within Linking Open Data (LOD) Cloud,[2] namely a visual historic landscape with the evidence of many different L(O)D packages covering actually more than an estimated 50 billion facts[3] from different knowledge domains. These facts are of varying quality and most of them (published under Open Licenses) can also be re-used (consumed and enriched) by different agents.

In the last years, also different bibliographic datasets - including digital collections, metadata, semantic and authority files (mono e multi-lingual vocabularies, classifications, thesauri) – have been published and re-used according to "Tim's 5 star deployment scheme"[4]

[2]CKAN: Linking Open Data Cloud, http://datahub.io/group/lodcloud.

[3]Linked Heritage Project "Best practice report on cultural heritage", http://www.linkedheritage.eu/getFile.php?id=229.

[4]Tim Berners-Lee, Up to Design Issues, 2006, http://www.w3.org/DesignIssues/LinkedData.html; "Tim's 5 star" Open Data plan with examples, http://5stardata.info; OCLC video: "Linked Data for Libraries": short introduction to the concepts

principles. Library Linked Data (LLD) Report and CKAN Registry section for LLD,[5] Linked Open Vocabularies (LOV) Service,[6] the "Global interoperability and Linked Data in libraries" international "know-how" exchanging meeting (*Global interoperability and Linked Data in libraries. Special issue of "JLIS.it"*) can be cited within the first most important "witnesses" reporting and describing proliferating of bibliographical LD activities at the global scale. The landscape of bibliographical information – treated according to LD methodologies – is already enough widespread. Just to mention some connected experiences:

1. German National Library (DNB) LD Service for authority bibliographical data linking;[7]

2. Library of Congress LD authority files;[8]

3. LD collections from "The Open Library", "The European Library", "Europeana" and "WorldCat.org"[9] web services;

4. Hungarian National Library OPAC and Digital Library published according to LD and SKOS (Simple Knowledge Organi-

and technology behind Linked Data, how it works, and some benefits it brings to libraries, http://www.youtube.com/watch?v=fWfEYcnk8Z8.

[5]Library Linked Data Incubator Group: Datasets Value Vocabularies, and Metadata Element Sets W3C Incubator Group Report 25 October 2011, http://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset-20111025; CKAN, Library Linked Data: http://datahub.io/group/lld.

[6]http://labs.mondeca.com/dataset/lov/index.html.

[7]http://openbiblio.net/2012/01/26/german-national-library-goes-lod-publishes-national-bibliography; http://files.d-nb.de/pdf/linked_data.pdf.

[8]http://authorities.loc.gov.

[9]http://openlibrary.org/; http://www.theeuropeanlibrary.org/tel4; http://pro.europeana.eu/linked-open-data;http://dataliberate.com/2012/06/oclc-worldcat-linked-data-release-significant-in-many-ways;https://www.oclc.org/data.en.html.

zation Systems)[10] formalisms;

5. British National Library bibliographical LD datasets connected to different LOD sets such as VIAF, LCSH, Lexvo, GeoNames, MARC country, "Dewey.info", RDF Book Mashup;[11]

6. LODUM, LOD service improving access to scientific and educational data at the University of Münster;[12]

7. "Burckhardtsource.org" and VOA3R digital infrastructures allowing enrichment, cross-relating and searching of cultural and scientific digital contents with LD technology support;[13]

8. "Data.bnf.fr" LD Project of the Bibliothèque nationale de France;[14]

9. LD at the Biblioteca Nacional de España;[15]

10. Public Library of Veroia in Web 3.0.[16]

An overview of consuming LD applications (faceted browsers,[17] LD browsing, LD search engine, On-the-fly *mashups* etc.) was recently good described in "Consuming Linked Data" document (Sequeda). To translate the initial success of Linked (Open) Data into a stable world-scale reality within bibliographical universe, encompassing

[10] http://iskouk.blogspot.com/2010/05/hungarian-national-library-opac-and.html.

[11] http://talis-linkeddata-libraries.s3.amazonaws.com/Linked%20Data%20Prototyping.pdf.

[12] http://code.google.com/p/lodum.

[13] http://burckhardtsource.org;http://voa3r.cc.uah.es.

[14] http://data.bnf.fr/docs/databnf-presentation-en.pdf.

[15] http://openbiblio.net/2012/02/02/linked-data-at-the-biblioteca-nacional-de-espana.

[16] http://gr.okfn.org/2012/10/libver/?lang=en.

[17] FAST (Faceted Application of Subject Terminology): http://fast.oclc.org, an Experimental OCLS Services for Controlled Vocabularies: http://tspilot.oclc.org/resources.

the Web 2.0 and commercial data alike, there are still several challenges to be addressed:

- "LD literacy" about benefits of publishing, re-using and integration of bibliographical resources as LD still needs to be widely promoted, directly (through standards) and indirectly (through "good practices");
- different requirements "to express metadata design patterns, both as templates for Linked-Data-compatible data formats and as reference points for creating and consuming coherent metadata within communities of discourse and practice"[18] according to a common *Resource Description Framework* (RDF, an international data exchange standard) should be re-evaluated;
- available strategies, e.g. "LODE-BD Recommendations" (Subirats and Zeng) regarding LD-enabling metadata encoding should be widely welcomed and implemented (De Robbio and Giacomazzi);
- processes for automatic alignment of metadata terms with LD-enabling sets should be better explored, formalized and shared as common models among different communities of practice;
- trust and common sense of LD are all still necessary: only trustworthy data patterns should be published as LD;[19]
- available *scientific data publication models* on top of LD (Bechhofer et al.) should be broadly transferred between research communities and exploited more deeply.

[18] DC-2013 "Linking to the Future" initiative, http://dublincore.org.

[19] It is the goal of LOD2 Project (FP7 Information and Communication Technologies Work Programme) to develop adaptive tools for searching, browsing, and testing authoring of LD, http://lod2.eu/Welcome.html.

Institutional Repositories (IRs) - as digital information systems promoting knowledge visibility on institutional digital research resources[20] - can be both publishers of their value datasets (e.g. metadata, vocabularies, collections) as well as consumers of available L(O)D sets.

For example, at Oregon State University ScholarsArchive@OSU both *Linked Dataset* covering University's *theses* and *dissertations* as well as *links* from this Dataset to external LD sets have been developed (Johnson and Boock). This activity has been started from converting MARC and Qualified Dublin Core metadata - describing the respective theses and dissertations - into LD through a RDF data model formalizing the expression of key data points for these resources. Afterwards, different relationships among IR's resources (with *handle* identifiers) have been described in a simple way (e.g. *rdfs:seeAlso*), as well as through complex semantics: mappings supported by internally and externally maintained LD datasets and controlled vocabularies for "Title", "Responsible Body", "Subject" entities. The querying of the entire *Linked Dataset* is possible via a *SPARQL* (*Protocol and RDF Query Language*) endpoint provided by the *Triple Store* that sits on top of the created knowledge LD base. Considering the importance of the above presented issues, this article is aiming at:

1. making a short overview of LD origins and its benefits for digital contents;

2. describing a role of controlled and semantic vocabularies in improving creation, access and retrieval of digital contents. A list of some important authority and semantic LD-enabling datasets will be provided;

[20] In the IR context the term "resource" can denote an article, monograph, thesis, conference paper, research report, presentation material, thesis, learning object, research data etc.

3. overviewing some approaches, documents and principles for creating metadata elements describing IR objects, focusing on the "Guidelines for metadata creation and management in Institutional Repositories" strategies (Conference of Rectors of Italian Universities (CRUI), Open Access Group, Italy);

4. presenting benefits of "LODE-BD" Recommendations (Subirats and Zeng), whose encoding *Decision Tree* strategies are devoted to support Repository metadata to become LD-enabled. Aside "literal" values for qualifying metadata properties, "LODE-BD" strategies are paying particular attention to assigning "non-literal" Uniform Resource Identifier (URI)[21] values. LD-enabling is also possible through mappings between Dublin Core (DC) metadata and more specific ontology-oriented metadata;

5. contributing with an extension to "Intellectual Property Rights" LODE-BD's *Decision Tree*, providing decision steps to licence choice. A list of some important licences - LD-enabling (identified by URIs) will be presented;

6. discussing briefly "Design-time" and "Run-time" LODE-BD implementation strategies and reporting thereupon some practice examples.

2 Linked (Open) Data: a brief reminder of its origins and benefits

In recent years, the concept *Linked Data* - referring to a set of best practices for publishing and connecting structured data on the Web - has

[21] The URI standard definition, RFC 2396: http://tools.ietf.org/html/rfc2396.

been already evolved as a high promising candidate into addressing one of the biggest challenges in the area of intelligent information management: the use of the Web as a platform for data and information integration in addition to document search. The term *Linked Data* (LD) was coined by Tim Berners-Lee in 2006 and formalized within already mentioned "Tim's 5 star deployment scheme", whose principles are being summarized as follows:

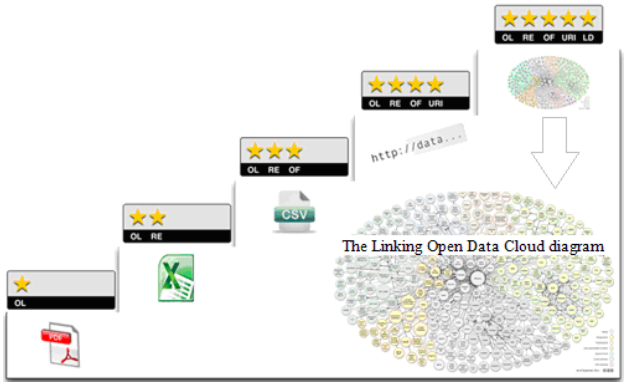


Figure 1: Tim's 5 star deployment scheme.
☆ Make datasets (contents) whatever format available on the Web under an *Open License*
☆☆ Make them available as structured data in RDF
☆☆☆Use *non-proprietary formats* (e.g. CSV instead of Excel)
☆☆☆☆ Use URIs to denote things, so that other agents can point at your datasets
☆☆☆☆☆*Link/combine* the data safely with other data in URIs global scheme to provide context

> [LD] isn't just about putting data on the web. It is about making links, so that a person or machine can explore the web of

> data. With Linked Data, when you have some of it, you can find other, related, data. (Berners-Lee)
>
> With Web advances to an era of *Open* and *Linked Data*, the traditional approach of sharing data within silos seems to have reached its end. From governments and international organizations to local cities and institutions, there is a widespread effort of *Opening* up and *Interlinking* their data. (Subirats and Zeng)
>
> Linked Data *does not of course in general have to be Open* - there is a lot of important use of Linked Data internally, and for personal and group-wide data. You can have *5-star Linked Data without it being Open*. However, if it claims to be Linked Open Data then it *does have to be Open*. (Berners-Lee)

Linked Open Data paradigm is a Linked Data strategy for global identity (Glaser and Halpin) of Open Data[22] (datasets published under *Open Licenses*) allowing re-use of LD datasets, freeing and enriching shared data between human and software agents. Below there are some benefits[23] that can derive from publishing and/or alignment of digital Repository content objects (resources, metadata,[24] research data) according to LD-enabling strategies:

1. possibility of linking, sharing, and querying (meta)data from different sources and formats. LD leads to organize a "silo" environment of disconnected resources from different Repositories in one space of structured connected data;

[22]The Open Data Handbook. Open Knowledge Foundation, 2010-2012, http://opendatahandbook.org/en.

[23]Benefits of the Linked Data Approach, http://www.w3.org/2005/Incubator/lld/wiki/Benefits; EC FP7 Support Action LOD-Around-The-Clock (LATC),http://5stardata.info.

[24]User Guide/ Publishing Metadata: "How to use DCMI Metadata as Linked Data. Publishing and Consuming Linked Data with RDFa", http://wiki.dublincore.org/index.php/User_Guide/Publishing_Metadata.

2. avoiding data redundancy (duplication) and keeping it updated;
3. cross-referencing to L(O)D authority and semantic files;
4. bookmarking of global encyclopedic cross-domain information (e.g. *DBpedia, Open Library data mirror in the Talis Platform, The Open Library, Freebase, GeoNames Semantic Web*) avalaible as LOD and reusing its parts;
5. directly processing data without being confined by the capabilities of any particular software, to perform data aggregation, calculations, visualisation, access, exporting, fine-granular control over the data items (e.g. load balancing, caching);
6. better measuring of data contributions in specific research disciplines. LD strategies can bring together the datasets living in disparate Repositories around the world that vary significantly by (or within) disciplines or even type of study;
7. new audience attracted by rich digital content developed on Repository LOD sets by means of APIs (Application Programming Interfaces) and web mash-ups often combining ”general“ APIs (Jarrar and Dikaiakos)
8. better user experience based on connected contextually relevant datasets. Users be more likely to visit again this or that Repository or Portal enhanced by LD-enabling strategies.

3 The importance of controlled vocabularies and semantic schemes

Controlled vocabularies and semantic schemes - such as lists of authority control including standard name identification, classification

systems, thesauri, topic maps, ontologies - are known generically as Knowledge Organization Systems (KOSs). KOSs provide a systematic way to better organize, access and retrieve knowledge inherent to information resources, through the mandate use of predefined, authorized and semantically expanded terms, with indication of different variations, spellings and misspellings, uppercase versus lowercase variants (Guerrini, Tillett, and Sardo). Without using KOSs in describing digital resources, both users and machines are stymied in their efforts to better access and aggregate them (Salo). Controlled vocabularies are maintained by an Authority (e.g. NACO Authority of the Library of Congress) ensuring that all terms are defined consistently and have well-defined relationships. In theory, any piece of information is amenable to authority control such as *personal* and *corporate names*, *uniform titles*, *series*, and *subjects*, trying to bring "structure and order" (to collocate materials that logically belong together but which present themselves differently) to the task of helping users to find information.

Assigning, for example, to an *author*, *subject*, *license* etc. a particular unique heading (term expressed by string or web address identifier), which is then used consistently, uniquely, and unambiguously to describe all references to that "piece" - can be combined into a database and called an *Authority File*. This files should be maintained and updated as well as "logical linkages" to other connected files/records should be provided by metadata practitioners and other information professionals. Different controlled and semantic vocabularies have been already published according to SKOS (Simple Knowledge Organization Systems)[25] RDF/S formalisms and released as L(O)D sets, in order to be comprehensible, shared and re-used among different actors on the web. Use of these KOSs to qualify certain metadata values could also facilitate LD-enabled linking in IRs, as

[25] http://www.w3.org/TR/2009/REC-skos-reference-20090818.

it was already demonstrated in the already mentioned *VOA3R* and *ScholarsArchive@OSU* Repositories.

Despite the availability of different SKOS/LD KOSs,

> Future work on Linked Data [in Institutional Repositories] should address gaps [...] Since most Thesis authors and many other Repository submitters do not appear in major Library Name Systems, these [controlled vocabularies published as LD] solutions are of limited help. What is needed is a Locally maintained Name database [...] for internal Name Authority based on the Simple Knowledge Organization System (SKOS) vocabulary (Johnson and Boock)

and possibly published as LD under Open licenses, which would allow for derivatives to be created (e.g. multilingual versions, connection with other LOD authority and semantic files). Normalized and semantically enriched (through URIs values) metadata terms could present a qualitative basis for high-tech navigation interface modules (e.g. faceted search[26]) to refine and expand search and retrieval results:

> applying Standard Subject Vocabularies and Classification Schemes is more expensive than assigning a few uncontrolled keywords [...] expenditures in development often result in greater efficiency and effectiveness for the end user. Use of a Standardized Subject Thesaurus or other Controlled Vocabulary, for example, can provide greater precision and recall in searching, and can enable future functionality, such as faceted subject browsing and dynamic searching of subject matter. (NISO Framework Working Group 58-59)

[26]EIFL, Knowledge without boundaries, http://www.eifl.net/faceted-search.

4 Some approaches, documents and principles for creating qualitative and extensible metadata elements describing IR objects

At various stages of an *information object's life cycle*,

> creators of digital objects should be encouraged to embed as much metadata as possible within the object before it is shared or distributed [...] Institutions should be aware that, depending upon the nature of their collections, a single Metadata Schema may not suffice for all their needs. Thus a judicious combination of metadata schemas may be the best solution for some materials.[27]

The metadata schema from CRUI Guidelines offers an extend use of 15 Unqualified (simple) DC metadata with additional refinements and elements. DC simple presents basic metadata elements to describe IR content objects, in order to support minimum interoperability among OAI-compliant Repositories by means of OAI-PMH protocol. Preferences to use DC metadata can be explained by its simplicity ("almost anyone can use it, or at least parts of it: hence, it is the metadata of choice for Institutional Repositories, where users upload their own works and create their own metadata"[28]), as well as by its high integrating capability (e.g. DC-Library Application Profile, Scholarly Works Application Profile, VOA3R AgRes AP Metadata Terms).
In order qualified metadata from Data Providers are not be flattened and depleted by harvesting OAI-PMH mechanisms, both Data and

[27] http://framework.niso.org/node/24.
[28] http://framework.niso.org/node/24.

Service Providers should support common and widely shared standards and protocols as well as qualitatively developed cross-walking schemes (mappings among schemas), limiting loss of data or their specificity.

It is a good practice when metadata elements motivated choice, along with their consistent compilation and encoding design approaches and requirements are declared in the appropriate IR Policies. These last are also important for the development of a widely-spread new trend for Repositories such as *Data Management Plans* (DMPs)[29] aiming to qualitatively support entire life cycle both of metadata and research data[30] complementing the context of deposited content objects.

With qualitatively programmed, encoded and widely cross -referenced metadata, "*Institutional Repositorie*s will be ultimately to form an *International Network* of indexed Repositories searchable from a single interface",[31] deploying a single virtual entry-point for exchanging and augmenting open bibliographic data improving the dissemination of research results in via Open Access. During the selection and development of metadata elements it would be appropriate to make the continuous confrontation with six NISO's (National Information Standards Organization)[32] principles for "good metadata". "*Good metadata*":

1. conforms to community Standards in a way that is appropriate

[29]Data Management Plans. Digital Curation Center, http://www.dcc.ac.uk/resources/data-management-plans.

[30]"Research Data", University of Bath, http://www.bath.ac.uk/research/data.

[31]Statement from the University of Oregon Libraries, http://library.uoregon.edu/diglib/irg/SB_Role.html.

[32]http://framework.niso.org/node/24. On February 2013 NISO launched a new initiative to develop Standard for "Open Access Metadata and Indicators" (standardized bibliographic metadata and visual indicators to describe the accessibility of Journal articles with respect to how "open" they are): http://www.niso.org/publications/newsline/2013/newslinefeb2013.html#report2.

to the materials in the collection, users of the collection, and current and potential future uses of the collection;

2. supports interoperability;

3. uses Authority Control and Content Standards to describe objects and collocate related objects;

4. includes a clear statement of the conditions and terms of use for the digital object;

5. supports the long-term curation and preservation of objects in collections;

6. are objects themselves and therefore should have the qualities of good objects, including authority, authenticity, archivability, persistence, and unique identification.

"Good" (qualitative) metadata requires an understanding of both data that is going to be described and standard/s by which such a description would be possible. The section "Metadata validations" (Conference of Rectors of Italian Universities (CRUI), Open Access Group, Italy 11-12) of CRUI Guidelines underlines that metadata quality, in turn, determines the quality of functions performed and services offered both by Repositories (Data Providers) and their aggregators (Service Providers), considering the context of interoperability within the OAI model. In creating "good metadata" elements, it is also worth referring to such an authoritative document as "User Guide/Creating Metadata" developed within DCMI Community.[33]

[33] http://wiki.dublincore.org/index.php/User_Guide/Creating_Metadata#Guidelines_for_the_creation_of_medium_content.

4.1 CRUI Guidelines: requirements for creation of qualitative IR metadata

To ensure metadata accuracy and their qualitative compilation during the self-archiving process of digital materials in the IR, CRUI Guidelines recommends to:

1. Assist users during self-archiving (based on the metadata insertion process) of their content objects. It may be possible through the establishment of facilities such as *metadata editors* with dynamic lists for auto-completion and capture/import of metadata values from different authoritative sources (e.g. internal and external authoritative files to control values of ”Responsible Body”, ”Subject”, ”Place” metadata).

2. Validate metadata inserted prior to its exposure to the final users and Service Providers. Effectiveness and efficiency of the metadata import/export are closely related to the use of Unique Identifiers (e.g. URI). It is a good practice when Unique Identifiers are assigned automatically within IRs platforms to research products and authors. Using Unique Identifiers as ”non-literal” data values describing metadata properties should reassure the stability of metadata elements they addressing to, as well as their interoperability among different systems. Moreover, the duplication of metadata values will be easily avoided, effective filters for the discovery of related resources (e.g. created by the same author), as well as efficient navigation tools can be developed. Unique Identifiers could be also of great importance in creating qualitative connections between research content and its evaluation processes (e.g. IRs as technical infrastructures for research management and assessment[34]). Effectively exploiting within networks the po-

[34]Institutional Repositories for Research Management and Assessment, on the

tential of Unique Identifiers assigned by IRs, alongside with CERIF (Common European Research Information Format) and other research data metadata standards as well as with applying of widely-accepted scientific disciplinary sector classifications, greater integration between Open Access Repositories (OAR) and Current Research Information Systems (CRIS e Euro-CRIS)[35] can be achieved. Currently,

> Many different research information systems (RIS) implement CERIF data model [which] has concepts of base relations and link relations (with role and temporal duration) [...] Several RIS providers had also published Web APIs using SOAP or REST technologies to support web applications and mash-ups with data from other systems. These APIs varied and were proprietary [...] Bringing "data islands" to a global, interconnected data space leveraging RDF, SPARQL, and OWL ontologies. In that context, reuse of well-established ontologies beyond FOAF, Dublin Core, and BIBO should be explored. (Jeffery and Corson-Rikert)

3. Provide each Repository with professional metadata support. Considering that the validation of metadata quality is an organizational management issue rather than a procedural one, it is a good practice to establish within each IR a support unit directed by metadata professionals.

In the near future

"Open Access scholarly Information Sourcebook" portal, http://openoasis.org/index.php?option=com_content&view=article&id=165&Itemid=335.

[35]The World Confederation of Open Access Repositories (COAR) and euroCRIS recently announced a strategic partnership. Specific attention will be paid to the domain of interoperability between different OA Repositories and CRIS to ensure appropriate management of research results, http://www.coar-repositories.org/news/eurocris-and-coar-join-forces-building-up-a-mutual-partnership-2.

> it is very likely that [all] local Repositories will be forced to employ a *quality metadata* content description and metadata harvesting system [as] Most leading citation databases consider metadata, or as the case may be metadata harvesting systems, conditional for integrating or monitoring the Repository [Moreover] In order to fulfil their mission and maintain a high quality Standard, these local Repositories have to seek and implement innovations in compliance with the latest technologies and information resources development so that their content can be unequivocally identified and meta-described with a view to content distribution. (Šimek 88)

The "metadata quality" concept recalls the concept of "trusted environment", which is being actively promoted within the frame of (certificated) *Trusted Digital Repositories*[36] developed in respect with the requirements of widely-accepted Standards, trusted recommendations and guidelines.

The core metadata elements presented by CRUI Guidelines are aiming to cover a basic description of the following types of digital content research objects: *Article, Patent, Book and Part of the book, Conference object, Paper of conference, Poster of conference, Annotation, Review, Doctoral Thesis, Master Thesis, Bachelor Thesis, Working Paper*. The Metadata schema that will be presented in the penultimate paragraph of this article will state:

- restructured and well defined metadata elements from CRUI Guidelines according to "LODE-BD" metadata groups of common properties;
- an extended number of metadata elements from "Guidelines" according to proposed "LODE-BD" mappings;

[36]Interesting contributions on this theme were released during International Conference 2012 "Cultural Heritage online – Trusted Digital Repositories", Florence, http://www.rinascimento-digitale.it/conference2012-culturalheritageonline-materials.phtml.

- choices for encoding of metadata elements from "Guidelines" according to "LODE-BD" strategies.

4.2 "LODE-BD" Recommendations

"LODE-BD" Recommendations are encompassing important components that a Data Provider may encounter when decides to produce sharable LOD-ready structured data describing bibliographic resources such as *Articles, Monographs, Theses, Conference Papers, Presentation Material, Research Reports, Learning objects*, etc. (Subirats and Zeng 4). "LODE-BD" aims at addressing two questions:

1. how data - hosted by diverse Open Repositories - can be better exchanged across Data Providers;

2. how to encode this data within LOD-enabled metadata.

"LODE-BD" provides a selected number of widely used metadata standards and the emerging LOD-enabled vocabularies. Metadata terms from the DCMES (dc:) and DCMI Metadata Terms (dcterms:) are the fundamentals, while metadata terms from other namespaces are supplemented when additional Repository needs should be met. These supplemented metadata are including the namespaces from BIBO Ontology, AGLS Metadata Standard of the Australian Government Locator Service, eprint (UKOLN Eprints Terms, SWAP), and MARCrel (MARC List for Relators). All metadata terms are presented in a crosswalk table. Based on different cross-referred metadata namespaces and controlled vocabularies, the descriptive metadata would of course benefit in terms of their consistency, extensibility, semantic and authority richness.
Referring to the development stage of metadata terms according to "LODE-BD", Repository managers should address the following issues:

- What kinds of entities and relations there should be involved in describing and accessing bibliographical resources?
- What properties should be considered for publishing meaningful/useful LOD-ready bibliographic data?
- What metadata terms are appropriate in any given property when producing LOD-ready bibliographical data from a local database? (Subirats and Zeng 5)

4.2.1 "LODE-BD" *Decisions Trees*. Between "literal" and "non-literal" metadata values

The real strength of "LODE-DB" development stages are Decision Trees (Figure 2) designed to facilitate the selection of the appropriate strategies adjustable to Data Providers according to their local needs, while all moving towards the goal of metadata exchange and re-use f their values on the Web of Data.

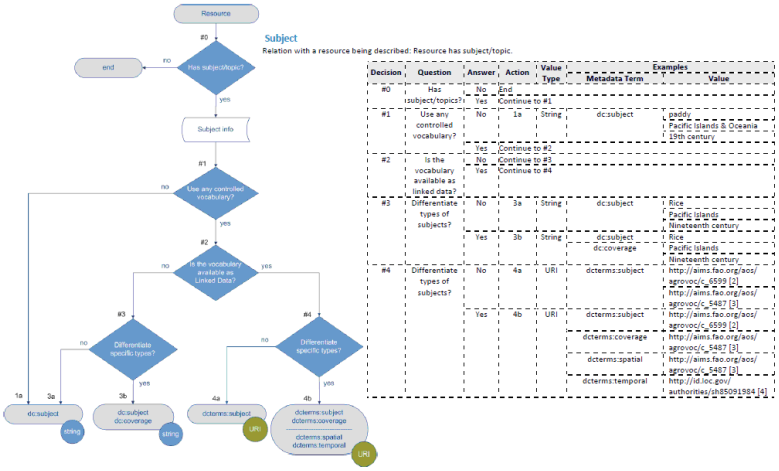


Figure 2: LODE- BD *Decision Tree* and explanation table to describe and encode "Subject" information (Subirats and Zeng 31-32)

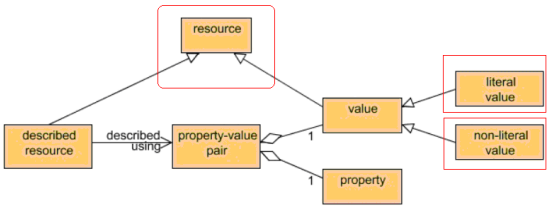
LODE-BD *Decision Trees* are developed to support description and encoding of the following metadata elements:

1."Title information"	Title/Alternative title
2. "Responsible Body"	Creator. Contributor. Publisher
3. "Physical Characteristics"	Date. Identifier. Language. Format/Medium. Edition/Version. Source
4. "Holding/Location information"	Location/Availability
5. "Subject Information"	Subject/Topic
5. "Description of Content"	Description/Abstract/Table of Contents. Type/Form/Genre
6. "Intellectual property rights"	Right Statements
7. "Usage"	Audience/literary indication/ education level
8. "Relation"	Relation between resources. Relation between agents

All *Decision Trees* are starting from the property describing a *Resource* instance and are delivered in flowcharts with various acting points, giving a "step-by-step" solutions for decisions to be made, further explained within text based tables, with notes, steps, and examples matching encoding suggestions, whenever essential. Within these explanations two types of metadata values - that can be chosen to qualify certain metadata properties - are provided:

1) **Literal value**. *This is typically a string of characters using a Unicode string as a lexical form, together with an optional language tag or data-type, to denote a "Resource".*

Examples of metadata namespace **dcterms:alternative** *"A Feast of Beans"* **dcterms:available** *"2006-07"*^^dcterms:W3CDTF ...



2) **Non literal value**. *This value presents physical, digital or conceptual entities indicated by Unique Identifiers.* LODE-BD "Decision Trees" help Data Provider to evaluate the existing gap between current use of literal values and their evolution to a LD approach (i.e. by using "non-literal" URI values from Controlled Vocabularies and other LD sets).	Examples of metadata namespaces: **dcterms:conformsTo** <http://www.w3.org/2001/XMLSchema> **dcterms:contributor gnd:135066719** gnd:135066719 *foaf:familyName "Elliott"; foaf:givenName "Missy" ; foaf:nick "Missy E"...*

Properties of some DC metadata namespaces ("dc:" and "dcterms:") – as it is demonstrated within LODE-BD explanatory tables and good described in the User Guide "*How to use DCMI Metadata as Linked Data*"[37] - may be qualified both by "literal" and "non-literal" values. However, to produce LD-enabled metadata that can be easily harvested by Service Providers on the web, the use of "dcterms:" namespace properties qualified by "non-literal" (URI) values is recommended.

The pragmatic relevance of LODE-BD *Decision Three*'s approach for producing LOD-enabled metadata is that each Data Provider can highlight within the concrete *Decision Tree* its own decision paths, marking the metadata terms to be used as well as choosing vocabularies and standards on their support. "LODE-BD" are not limited to subject-specific domains, thus being appropriate for use by any Data Provider accordingly to local needs. Nevertheless, "Decisions regarding what Standard(s) to adopt will directly impact the degree of LOD readiness of the bibliographic data" (Subirats and Zeng 1,3).

[37] http://wiki.dublincore.org/index.php/User_Guide/Publishing_Metadata#Properties_of_the_terms_namespace_used_only_with_non-literal_values.

4.2.2 "Intellectual Property Rights". Controlled vocabularies LD-enabling

Before a certain resource is published, it is important to decide under which License it will be presented to users. As it was already mentioned in connection with "Tim's 5 star deployment scheme", it is advisable to publish digital contents on the Web under an *Open License*, in order they can be freely: shared (copied, distributed and transmitted), remixed (adapted), used by any 3rd party (including commercial) to produce derivates, anyhow with attribution the work to the author or licensor. This should be applied even more to research resources produced in public domain (De Robbio). However, considering that some IR resources could be connected with issues of: "Embargoed access" (the resource is of Closed Access, until released for Open Access on a certain date), "Restricted access" (Open Access, but with restrictions) and "Closed access" (opposite of Open Access), aside "Open" also "Not open" licenses may be used to denote "dcterms:rightsHolder", "dcterms:licence" metadata properties. Authors can find useful informational support about *Intellectual Property Rights* and *Licenses* within good compiled services like SHERPA/Romeo "Publisher copyright policies & self-archiving"[38], "Diritto d'autore" (service offered by University of Padova Library System).[39]

Some decision steps to choose a particular License describing the use of resource are presented in Figure 3 on the next page, as an extension to "LODE-BD Decision Tree" referring to "Rights: Situations and best practices for encoding the data".

After a certain License is chosen, a value ("literal" and/or "non literal") identifying officially the License type should be encoded in the appropriate metadata property (ies), as according to "LODE-BD"

[38] http://www.sherpa.ac.uk/romeo.

[39] http://www.cab.unipd.it/servizi/diritto-dautore.

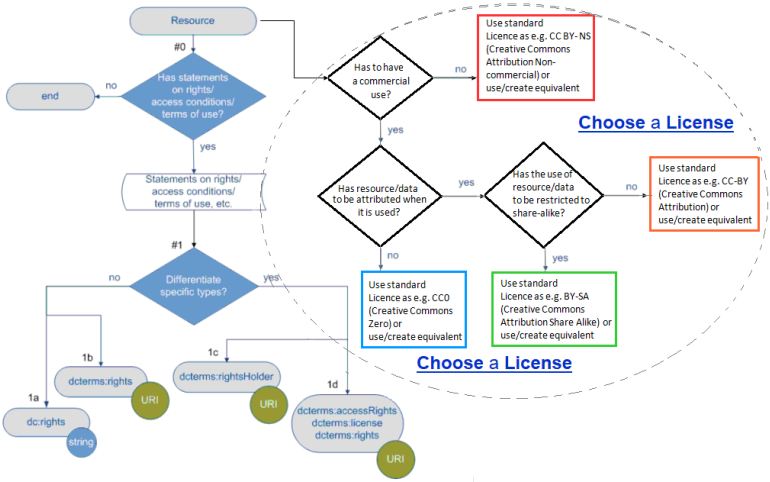
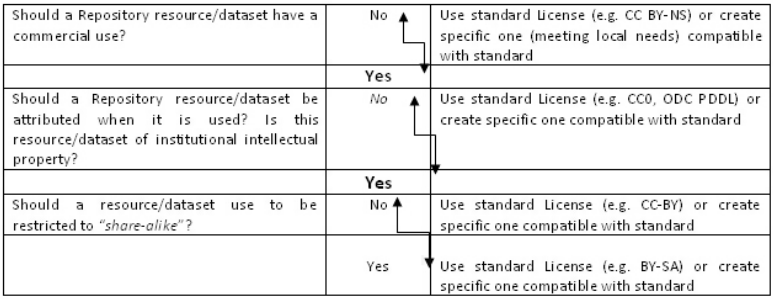


Figure 3: *Decision Tree*: a choice of a License for publishing Repository datasets

Figure 4: The extension provided to the LODE-BD Decision Tree in Figure 3 on the preceding page can be interpreted as follows:



encoding strategies. In Appendix … some wide-used (”Open” and ”Not-open”) Licenses are provided, together with their ”non literal” (URI) legal identifiers, which can help ”Intellectual Property Rights” metadata to become LD-enabled.

4.2.3 ”LODE-BD”: mapping of metadata with ”schema.org” mark-ups

In the ”LODE-BD” Appendix 4 cross-walks from certain metadata elements to ”schema.org” mark-ups are provided. ”Schema.org” mark-ups - natively relevant for webmasters[40] - (i.e., html tags used by webmasters to markup their pages in ways recognized by major search engines including *Bing*, *Google*, *Yahoo!* and *Yandex*) can be also used to improve representation and search of bibliographic information on the web. When you are exploring how data is inter-related on the web in order to learn more about patterns or things implicit in the data, is when it would be of benefit not only consider a RDF graph or LD view but also ”schema.org” mark-ups. This is

[40] http://schema.org; http://schema.org/docs/full.html.

particularly relevant to intelligence applications, scientific research and many other types of applications exposed on the web.

Different bibliographical data has already been supported by "schema.org" mark-ups in services such as, for example, *WorldCat.org*, *Data.bnf.fr*,[41] VOA3R Open Access Repository. The reason why "schema.org" is included in the "LODE-BD" is essentially through two reasons:

1. the benefit of creating micro-data by individual sources, e.g. webmasters or authors themselves when they publish data on the web, instead of going through a Repository and get exposures. It is another way to expose resources. It does not replace any metadata schema as, in case of "LODE-BD" proposed schema, it is to be complementary to DC metadata terms;

2. because it is multiple schemes, many of the properties used for bibliographic description also are used by other types of resources. Assuming there will be more resources use "schema.org", the chance of interoperability is high. Repositories also can harvest from those data which would have various benefits.

The "Schema Bib Extend Community Group"[42] within the "W3C Web Schemas Task Force" is preparing different proposals for extending "schema.org" vocabularies to improve representation and search of bibliographic data on the web.

[41] https://www.oclc.org/en-US/news/releases/2012/201238.html; http://data.bnf.fr/docs/databnf-presentation-en.pdf.

[42] http://www.w3.org/community/schemabibex.

5 Metadata schema for description of IR digital content objects

In the Metadata Schema as according to CRUI Guidelines, the mapping to OAI_DC metadata will be provided.
The metadata elements from CRUI Guidelines consider the Unicode encoding standard, important for the consistent representation and handling of text expressed in most of the world's digital writing systems, using XML schema as the primary medium based on "mix and match" method combining elements and sub-elements, related attributes, and controlled attribute values throughout the element sets. "LODE-BD" promotes the encoding of metadata elements within RDF/XML schemas to support their semantic consistency required today in most digital environments. Both CRUI Guidelines and "LODE-BD" assume that the metadata they provide could be more complex and structured, first of all in view of creating a more balanced framework that may allow to accommodate better different metadata models according to different Repository local needs for representation and management of their digital content objects.
The aim of alignment metadata elements from CRUI Guidelines according to "LODE-BD" is to show how metadata terms selected for the description of IR digital objects can be enhanced by encoding "LODE-BD" strategies. Summarily, such an aligning will lead to:

1. radically-improved metadata workflows. Data integration and reusability will save time for the development of new metadata indexes;

2. better IR resource description and discovery (searching and browsing) on the Web of Data. IRs will be able redirect their users straight from the Repository discovery interfaces to the connected knowledge DMSs (Data Management Systems)

Data Hubs provided by different related datasets in the LOD Cloud. The Repository contents will increase tremendously in their visibility and integration on the Web;

3. better data exchange through collectively shared data, based on common LD values;

4. the development of common search interface like "Institutional Repository WorldShare Platform" (see experience of "WorldCat Local"[43]) for search of IRs digital contents interconnected through LD-enabled metadata values worldwide;

5. creation, sharing and use of new applications enhancing the dissemination channels and accessibility of L)O)D sets through IR services, contributing qualitatively to Open Research Commons space[44] (White).

The metadata schema presented in Appendix can be considered an a tentative to create an application profile (AP) for IRs objects based on DC metadata (presented by CRUI Guidelines) and modeled according to "LODE-BD" (structuring metadata in categories; encoding strategies based on motivated use of "literal" and "non-literal" values; choice of cross-walking to more specific metadata terms and to "schema.org"). The aim of this presentation is also to show the usefulness of "LODE-BD" Recommendations to enhance expressive quality of IRs DC descriptive metadata.
The concept of AP was emerged within the DCMI as a way to declare which elements from which namespaces would be better to use in a particular application or project. Metadata elements can be

[43]Single-search access to 1.071+ billion items from your library and the world's library collections, http://www.oclc.org/worldcatlocal/default.htm.
[44]http://aims.fao.org/community/open-access/blogs/building-institutional-repositories-global-research-commons.

combined together by implementers in different ways, optimizing descriptive and system local needs.
The presented metadata profile can be considered as a part of "Design-time" implementation strategy defined by "LODE-BD". Both "Design-time" and "Run-time" LODE-BD strategies will be discussed in the next conclusive paragraph.

6 LODE-BD "Design-time" and "Run-time" implementation strategies

To align and implement descriptive metadata according to "LODE-BD" strategies, Data Provider may follow next two options (Subirats and Zeng 44) (Figure 5 on the following page):

1. "Design-time", i.e. changing current metadata model, replacing it with "LODE BD" proposals for selection and modeling of descriptive metadata. The choice of this strategy means also some changes to a current metadata database and services accessing it.

2. "Run-time" (on the fly)[45] option means that - while keeping the current metadata model and database structure unchanged - Data Provider should add a *conversion service* mapping and translating chosen metadata values from "literal" to "non-literal", following to "LODE-BD" Decision Tree's encoding strategy.

In Figure 5 on the next page due attention is given to the description of the "Run-time" strategy, pointing on conversion of "Subject"

[45]Example: those using OAI protocol, such as National Science Digital Library (NSDL): http://nsdl.org/contribute, here the wiki (for Contributors and Developers) can be followed to find the documents.

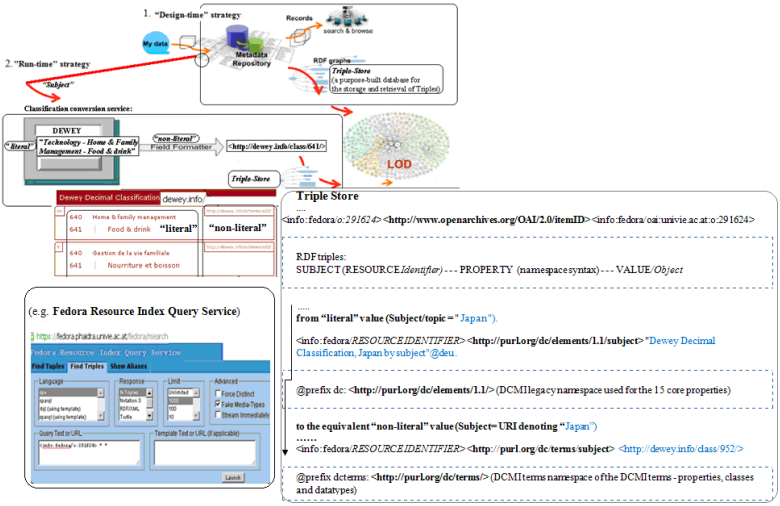


Figure 5: Metadata modelling according to LODE-BD ”Design-time” and ”Run-time” strategies.

metadata value from "literal" to "non-literal" value language. As is it shown, the "Subject" information can be described both by "literal" ("Japan" from Dewey Classification language @deu) and "non-literal" values (URI to equivalent concept from "Dewey.info" LD service). Both "literal" and "non-literal" values can be traced as graphs (Figure 6) in *Triple Store*.

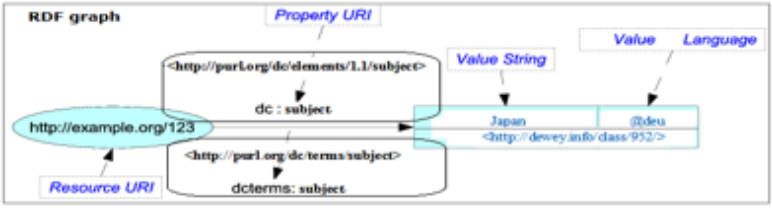


Figure 6: Exemplification of RDF graphs registered in *Triple Store*.

Triple store is a purpose-built database optimized for the storage and retrieval of triples (RDF), representing data entities composed of Subject (Resource[46]) - Property(Predicate) – Object (Value). Triple stores can be seen like the advantage for performance of Data Providers, also because all the information traced in a Triple store can be retrieved via a query language (e.g. a query language of Fedora Resource Index Query Service; Figure 5 on the preceding page). In addition to queries, triples can usually be imported/exported using RDF LD-enabling and other formats. A "non-literal" URI value denoting the "Subject" information in relation to a certain Repository resource – can be imported by other Providers (implementing Triple stores over querying of graph-based RDF models) using the same or related scheme(s) to qualify the "Subject" information:

[46] "To benefit from and increase the value of the World Wide Web, agents should provide URIs as identifiers for Resources", http://www.w3.org/TR/webarch/#uri-benefits

> if another party might reasonably want to create a hypertext link to it, make or refute assertions about it, retrieve or cache a representation of it, include all or part of it by reference into another representation, annotate it, or perform other operations on it.[47]

This assertion corresponds to the fourth and fifth stars of the mentioned "Tim's 5 star deployment scheme": (4) "use URIs to denote things, so that other agents can point at your datasets", (5) "*combine* the data safely with other data in URIs global scheme to provide context", thus contributing to the richness of content and context exchange within the global Linked Open Data space and, therefore, on the Web of Data. Anyhow, "triplifying" data by automatic script should be avoided as it is not the same as developing well-structured triples suitable for Repository applications. Proper data modeling is an essential first step in any implementation. Attempts to automatically generate billions of RDF "triples" and publish them on the Web is not the same as producing high quality data sets of properly modeled data, according to Standards, Recommendations and Guidelines.

> Simply transforming database schemas into RDF does not create Linked Data [...] To create automatic links between RDF triple stores on the web should be possible, otherwise there is a risk of creating RDF silos. The easiest way to facilitate the establishing of automatic linking between datasets is the use of Standard Vocabularies, including Standard Vocabularies for describing data/metadata elements and Standard Vocabularies for indicating values.[48]

The way how information content and context exchange can be obtained in an information service "on the fly", is good demon-

[47] http://www.w3.org/TR/webarch/#uri-benefits.
[48] http://aims.fao.org/linked-data/getting-started.

strated within the AGRIS searching platform[49] (Figure 7). Through "Open_AGRIS Beta" application, AGRIS platform supports the search and retrieval of resources described in its system by AGROVOC Thesaurus LOD set, included in AGRIS application through web service. Through AGROVOC "non-literal" (URIs) values, "Subject" metadata terms can be connected with different resources on the web, whose topic description is based on values of AGROVOC and related online datasets.

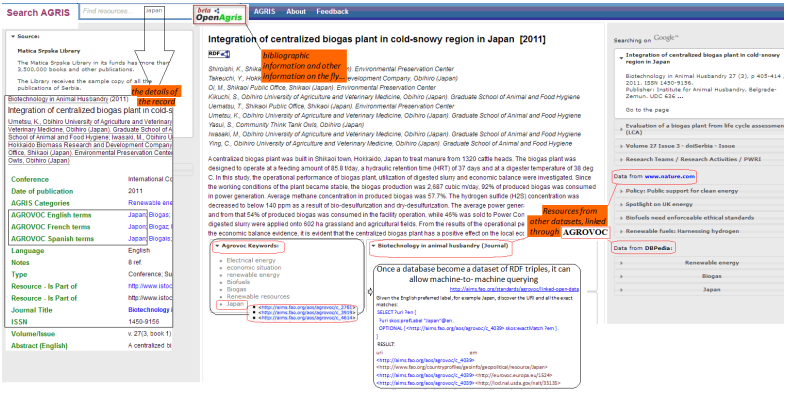


Figure 7: Search and retrieval of AGRIS resources and related datasets described by AGROVOC Thesaurus.

The search functionalities proposed by "AGRIS" can be considered one of the best examples presenting all components working together through "Subject" metadata values, supported by the controlled AGROVOC LD-enabled vocabulary using URIs to identify concepts and mapping concepts.

[49] International Information System for the Agricultural Sciences and Technology, http://agris.fao.org.

LODE-BD Decision Trees' metadata encoding strategies are based on the concept "usefulness to others". In the context of IR, their usefulness can be interpreted in terms of developing a rich IR LD-enabled metadata schema that can be re-used by different web actors, contributing to enhance visibility and semantic interoperability of IR digital content objects on the global scale.

References

Bechhofer, Sean, et al. "Why linked data is not enough for scientists". *Future Generation Computer Systems* 29.2 (2013): 599 –611. (Cit. on p. 113). Web. <http://www.sciencedirect.com/science/article/pii/S0167739X11001439>.

Berners-Lee, Tim. "Linked Data. Designed Issues". (2006). (Cit. on p. 117). Web. <http://www.w3.org/DesignIssues/LinkedData.html>.

Conference of Rectors of Italian Universities (CRUI), Open Access Group, Italy. "Linee guida per la creazione e la gestione di metadati nei Repository Istituzionali". (2012). (Cit. on pp. 115, 123). Web. <http://www.crui.it/HomePage.aspx?ref=2066>.

De Robbio, Antonella. "Forme e gradi di apertura dei dati. I nuovi alfabeti dell'Open Biblio tra scienza e società". *Biblioteche oggi* 30.6 (2012). (Cit. on p. 131). Web. <http://www.bibliotecheoggi.it/content/201200601101.pdf>.

De Robbio, Antonella and Silvia Giacomazzi. "Dati aperti con LODe". *Bibliotime* 10.2 (2011). (Cit. on p. 113). Web. <http://eprints.rclis.org/16440>.

Glaser, Hugh and Harry Halpin. "The linked data strategy for global identity". *IEEE Internet Computing* 16.2 (2012): 68–71. (Cit. on p. 117). Web. <http://eprints.soton.ac.uk/333924/>.

Global interoperability and Linked Data in libraries. Special issue of "JLIS.it". 2013. (Cit. on p. 111). Web. <http://leo.cilea.it/index.php/jlis/issue/view/536>.

Guerrini, Mauro, Barbara Tillett, and Lucia Sardo. *Authority control. Definizioni ed esperienze internazionali. Atti del convegno internazionale, Firenze, 10-12 febbraio 2003*. Firenze: Firenze University Press, Associazione Italiana Biblioteche, 2003. (Cit. on p. 119). Web. <http://www.fupress.com/Archivio/pdf/4383.pdf>.

Šimek, Pavel. "Using Metadata Description for Agriculture and Aquaculture Papers". *Agris on-line Papers in Economics and Informatics* 4.4 (2012). (Cit. on p. 126). Web. <http://online.agris.cz/files/2012/agris_on-line_2012_4_simek_vanek_ocenasek_stoces_vogeltanzova.pdf>.

Jeffery, Keith G. and Jon Corson-Rikert. "euroCRIS and VIVO. Part II Cooperation as Strategic Partners". (2012). (Cit. on p. 125). Web. <http://www.vivoweb.org/files/presentations/12Fri/euroCRIS_LOD_and_\%20VIVO.pdf>.

Johnson, Thomas and Michael Boock. "Linked Data Services for Theses and Dissertations". *Proceedings of the 15th International Symposium on Electronic Theses and Dissertations*. Lima. 2012. (Cit. on pp. 114, 120). Web. <http://hdl.handle.net/1957/32977>.

NISO Framework Working Group. *A Framework of Guidance for Building Good Digital Collections*. 3rd ed. Baltimore: NISO, 2007. (Cit. on p. 120). Web. <http://www.niso.org/publications/rp/framework3.pdf>.

Salo, Dorothea. "Name Authority Control in Institutional Repositories". *Cataloging and Classification Quarterly* 47.3/4 (2009). (Cit. on p. 119). Web. <http://minds.wisconsin.edu/handle/1793/31735>.

Sequeda, Juan F. "Consuming Linked Data". Proc. of Semantic Technology Conference, 2011. (Cit. on p. 112). Web. <http://www.slideshare.net/juansequeda/consuming-linked-data>.

Subirats, Imma and Marcia L. Zeng. "LODE-BD Recommendations 2.0 : How to select appropriate encoding strategies for producing Linked Open Data (LOD)-enabled bibliographic data". (2012). (Cit. on pp. 113, 115, 117, 127, 128, 130, 137). Web. <http://aims.fao.org/lode/bd>.

White, Wendy. "Institutional repositories: contributing to institutional knowledge management and the global research commons". *4th International Open Repositories Conference*. 2009. (Cit. on p. 136). Web. <http://eprints.soton.ac.uk/48552/>.

IRYNA SOLODOVNIK, Scuola Dottorale Internazionale degli Studi Umanistici (SDISU), Università della Calabria.
iryna.solodovnik@gmail.com

Solodovnik, I. "Development of a metadata schema describing Institutional Repository content objects enhanced by "LODE-BD" strategies". *JLIS.it*. Vol. 4, n. 2 (Luglio/July 2013): Art: #8792. DOI: 10.4403/jlis.it-8792. Web.

ABSTRACT: Based on "Guidelines for metadata creation and management in the Institutional Repositories" (CRUI, Italy, 2012) and "LODE-BD Recommendations" (AIMS, 2012), and exploring other principles and strategies for qualitative development of metadata representing contents and properties of digital contents, this article presents the specific metadata profile for description of Institutional Repository information resources. This profile is allocated within a metadata schema provided by well-defined metadata terms, compilation specifications, and alignment (mapping) strategies to more specific metadata terms and value properties enabling basic metadata to become more efficient in authority control context, richer in their semantic profiles and more accessible and usable on the web by means of Linked Data sets. Developing and implementing metadata schemas aligned completely or partially with Linked Data paradigm will provide metadata exchange among different Linked Data-enabling repositories, potentiate semantic relationship browsing and querying of their contents, enable their participation in the Linked Open Data cloud and contribution to an open research commons space.

KEYWORDS: Authority file; Institutional repositories; Knowledge management; Linked data; Metadata; Persistent identifiers.

ACKNOWLEDGMENT: The author would like to thank for some information support from AIMS members: Imma Subirats and Marcia Lei Zeng.

Submitted: 2013-02-23
Accepted: 2013-04-25
Published: 2013-07-01

JLIS.it

Usage of Reference Management Software at the University of Torino

Enrico Francese

1 Introduction

1.1 The practice of reference management in digital libraries

In today's scientific research and production, the practice of bibliographic citation management and "backward chaining" (Palmer, Teffeau, and Pirmann 11) can be managed by dedicated software tools, commonly known as 'Personal bibliographic softwares', 'Bibliographic Citation Management Software', 'Citation managers'. Following the Telstar definition,[1] the term "Reference Management Software" will be used (from now on shortened in RMS). According to Telstar's definition, RMS have two main functions:

1. building a database of citations to organize the documents useful for one's research;

2. formatting bibliographies and citations when writing papers through plug-ins or add-ons for Word processing software.

[1] http://www.jisc.ac.uk/whatwedo/programmes/institutionalinnovation/telstar.aspx.

JLIS.it. Vol. 4, n. 2 (Luglio/July 2013).
DOI: 10.4403/jlis.it-8679

Today's packages offer advanced features which vary form software to software, from the PDF storage and organization to including ways for annotation and sharing of data. The most prominent feature relates to the very nature of a "global information infrastructure" (Borgman) as a place of continuous and seamless interaction and integration: citations are shared, discussed, commented, suggested within members of the scientific community. RMS can act as a virtual research environment, or a platform for a "collaboratory" (Bos; Voss and Procter), sometimes adopting the features of virtual web collaboration networks, such as academic social bookmarking (Alhoori and Furuta; Fourie).

1.2 Research Questions

This study was conducted in 2012 within a master thesis project, whose topic was the inquiry of the role of RMS in a large academic institution such as the University of Torino, Italy. The research questions are:

1. what level of awareness about RMS exists in the members of the University of Torino?
2. what are the major trends in the usage of the RMS among the scholars?

The research's specific aims are:

- to explore and to understand the measurements about the actual awareness and usage of RMS;
- to understand the context in which scholars operate when dealing with citations and literature management;
- to provide evidence-based information upon which libraries can base their strategies about services, assistance, training.

1.3 The stage of the research: the University of Torino

The University of Torino (Università degli Studi di Torino, from now on shortened in UniTo) is one of the largest public universities in Italy, counting a population of 70000 students and 2000 faculties.[2]

In January 2008 UniTo, after solicitations by a group of professors, purchased 347 licenses for the software EndNote X1, to be distributed among those faculties who expressed a declaration of interest. The software was already known and used, but it was purchased by individuals, not by the institution. The largest group of users was constituted by the Biomedical Faculties (40%) followed by scientific areas (18%). All other disciplinary fields have been covered by less than 10%. Training sessions on the software gained a moderate participation ($\frac{1}{3}$ of the people involved in the distribution opted for a training session). In 2010 the licenses were not renewed due to two reasons: the lack of money and the technical difficulties posed by the new versions of the softwares (lack of compatibility with older operating systems, difficulties in upgrade, bugs, etc.).

2 Literature Review

Literature about RMS focuses mostly on the technical analysis of the features offered by the software packages[3](Gilmour and Cobus-Kuo; Childress; Butros and Taylor; Hensley and Kern). An important look is given at the reliability of these tools and the proper training needed by users: the papers by Fitzgibbons and Meert ("Are Bibliographic Management Software Search Interfaces Reliable? A

[2]The data exposed in this section are taken from the University's Programming Plan 2007-2012, and are updated at the academic year 2010-2011. See: http://www.unito.it.

[3]http://www.burioni.it/forum/dellorso/bms-dasp/text/index.html.

Comparison Between Search Results Obtained Using Database Interfaces and the EndNote Online Search Function") and Van Ullen and Kessler ("Citation Generators: Generating Bibliographies for the Next Generation") point at the role of reference librarians in providing information and support on managing bibliographies and citations. RMS can be looked at from the perspective of the users' behaviour and their relationship with other digital research tools, such as virtual environments. In their paper about the approach to digital libraries by researchers, Hull, Pettifer and Kell consider RMS as instruments that could enhance both personalization, social networking and collaboration, integration and accessibility (Hull, Pettifer, and Kell).

Giglia and Hane both point out the novelty and the potentials of the social networking solutions specifically addressed to the academics: "some social networks have been created and tailored to scientists' needs, in order to make them find researchers with similar interests or expertise, to keep in touch with their peers, to share their information" (Giglia). Haglund & Olsson ("The Impact on University Libraries of Changes in Information Behavior Among Academic Researchers: A Multiple Case Study") find that Swedish researchers do not have deep knowledge of the up-to-date digital tools that could enhance research and information management. A similar lack of awareness is shown by Ollé & Borrego: according to their study at Catalan Universities, researchers «described their techniques as "primitive" or "rudimentary"». Only 25% of their sample use some kind of personal bibliographic tool (Ollé and Borrego 51). In their survey conducted in 5 American universities, Niu et al. find that "information-seeking and information-handling habits of researchers are very personal" and inconsistent behaviours can emerge, even though the usage of a RMS is widespread. The activity of "sharing information within laboratories or groups or

among multisite collaborations", using tools like RMS, is seen as a potential evolving practice (Niu).

The studies above state that the usage of specific reference management tools is scarce and inconsistent. Yet few quantitative data are provided: Steele claims that "citation management softwares have existed since 1980 and are widely used today" (Steele 463), but doesn't give any reference for that. A survey at the Tallin University, Estonia, in 2011 (Francese) showed that the usage of these tools is low and not supported by a proper knowledge: scholars seem to be not fully aware about the potentials and the features of the RMS. Several papers indicate the active role that libraries can take about this subject (East; Siegler and Simboli; Martin). Childress considers the RMS in a practical perspective, studying them within the researchers' needs and workflows, and reflects about the supporting role that libraries can have (Cooke). According to East, the big effort in support and training given by his institution would be a key strategy for the future, and it will require big investment in staff resources: "the role of the academic librarian in general is evolving into a much broader function, particularly as regards the new and emerging information technologies" (East 70). The potential role of libraries is also confirmed by Crowley and Spencer: "Libraries also need to make their [i.e. the researchers'] research management and collaboration tools such as EndNote, EndNote Web, Zotero and RefWorks easily available, and ensure that all search interfaces incorporate a straightforward citation export function" (Crowley and Spencer 216). McMinn finally pushes for deeper studies on the topic: "There are a number of reasons why it is important to examine the different approaches research libraries take in providing similar services: ensuring that the services provided are consistent with those of peer institutions; determining how services have been tailored to meet the unique needs of different institutions; determining the

level of support and optimum allocation of resources" (McMinn 279).

3 Methodology and Method

This study aims to provide new essential informations on a relatively unexplored subject, with the goal of providing background for future understanding and comparison. To do so, the chosen method is a descriptive survey performed with a qualitative approach. Data were collected in two ways. An online questionnaire composed of 17 questions collected the measurable quantitative informations. 13 interviews were then conducted on a sample of the population to deepen, enlighten and circumscribe the data collected through the questionnaires with the aid of qualitative informations. Interviews were designed as "guided interviews" (Patton 202), or, to use the more precise terminology adopted by Corbetta, "semi-structured interviews" (Corbetta 198): a list of 8 "threads" was prepared, each of it being expressed through one or more questions. Interviews were performed in presence, face to face, recorded and transcribed. Data were collected anonymously: each respondent was labelled by a number, and no connection between the data and its identity was made. This study adopts the "constant comparative analysis" method (Strauss and Corbin). Concepts were identified as the data were being collected and linked together under 7 topics discussed in the end. The dimensions and the variety of the population of the University of Torino, required the choice of a focused disciplinary area. The literature review seemed to suggest that the health sciences and bio-medical areas are the most sensitive to the RMS features (Lawrence and Ashwell). Different key informants within UniTo confirmed this. The questionnaire was therefore addressed to professors, researchers and PhD students from the STM departments of

UniTo. An email with an introduction to the research and the link to the online form was sent to a mailing list of 1031 addresses. To select the interviewees, the availability was asked within the questionnaire, then a snow-ball chaining was performed across each respondent. In the end, 13 interviews were collected; respondents represent all the scientific areas questioned.

3.1 Limitations and caveat

The selected sample is only a subset of all the disciplinary fields covered at UniTo. Therefore its globality and heterogeneity are not wholly represented. The participation rate also can hide some clue to the faculties' interest or awareness about RMS. Although this cannot be proved, it is very likely that people very interested in the topic are more eager to participate in the survey, not to mention the interviews. This should be taken in account when reaching the final conclusions. Suggestions or expectations expressed in the interviews came mostly from active RMS users than from supposed non-users. This can become evident when cross-referencing the results from the questionnaire with those from the interviews. Conclusions risk to be unbalanced due to the nature of the sample.

4 Data results

4.1 The questionnaire

4.1.1 Response rate

The questionnaire collected 187 responses, reaching a response-rate of the 18,13% of the initial recipients. The academic roles are equivalently divided among researchers and professors (42% and 38%), with a 6% of PhD students and 15% of other roles (postdoc, research

fellow, lab assistant). The age of the respondents is also quite equilibrate: the majority is represented by people between 35 and 45 (37%).

4.1.2 Awareness and usage

The first important result is the general awareness about reference tools (figure 1) only 8% of the respondents declare to not know any software.

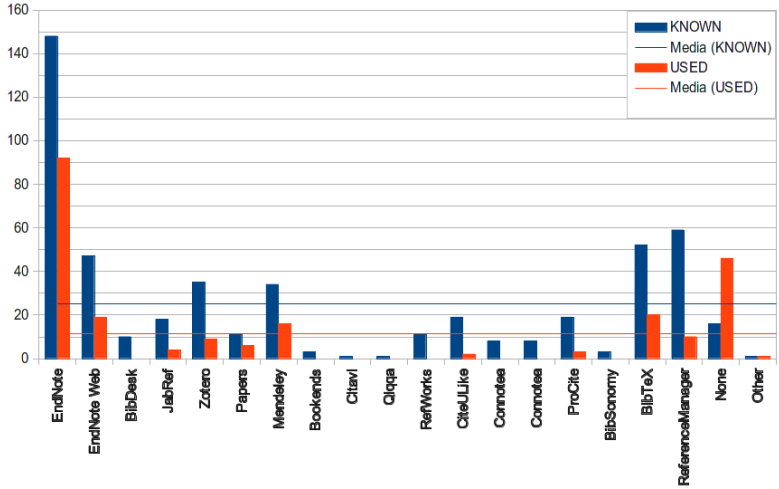


Figure 1: Knowledge and usage of softwares.

EndNote proves to be the best-known software: 79% of respondents know or heard about it, and among these, the 25% know about its web counterpart EndNote Web. Only 2 other softwares reached the 25%: BibTex (28%) and Reference Manager (32%). All the others seem to be mostly ignored; Zotero and Mendeley obtain 19% and 18% respectively, and the rest are from 10% under.

Data about usage show a more extreme trend. The non usage is relevant: 24%, almost a quarter of the sample. Usage of EndNote doesn't reach the half of the sample: barely 49% is the number of actual users, and just 10% also use EndNote Web. Of all the other softwares, only two are around 10% (BibTeX 11%, Mendeley 9%). It is remarkable the narrower set of softwares indicated in this answer: most tools obtain 0 responses.

The software distribution among age-ranges (figure 2) show how the percentage of non-usage is higher among older scholars (42% for the over 55), and very low among younger (9% among people from 26 to 35).

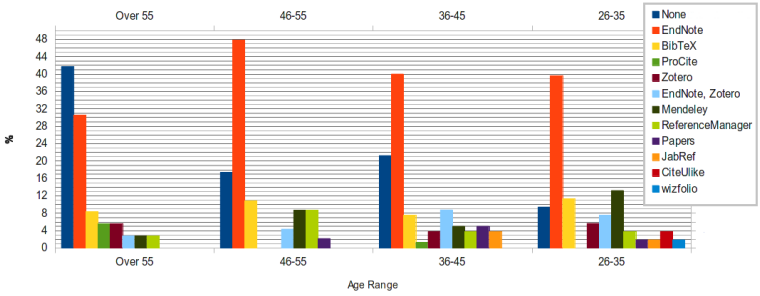


Figure 2: Percentage of software distribution per age (under 26 is exluded due to low sample).

The largest slice of members 53 respondents, 28%) is of long-time users (figure 3 on the following page).

4.1.3 Reasons and behaviour

Informations about user behaviour and the reasons behind are analysed through the interviews to be better understood. From a numeric point of view, we see that the most relevant reasons behind the

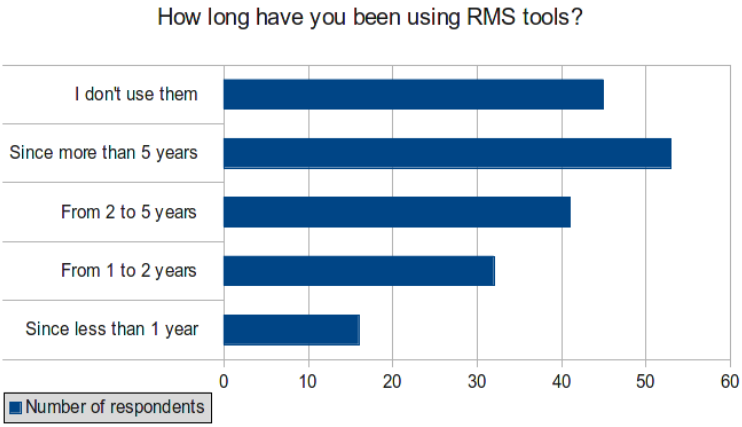


Figure 3

choice of a software indicate a sort of passive behaviour (figure 4 on the next page): softwares are mostly used because provided by the institution (33%) or used by the rest of the community (41%). While the community has a strong role, external information hasn't: only 2% chose a software after reading about it in journals or magazines.

Gratuity and open-source collect different responses: while the 16% pays attention to the freedom-of-cost, only the 7% cares about the license behind it.

From a quantitative point of view, usage of RMS varies: the number of citations saved ranges from less than 50 to more than 1000, with the highest numbers on the middle range (figure 5 on page 156). Figure 6 shows interesting data about the general approach to the tool (figure 6 on page 156). The most used features are the basic ones: editing (55%) and pasting (66%) the citations when writing the paper. Sharing citations is not a relevant activity (13%). Almost non existing is the usage of the RMS as a way to discover new references

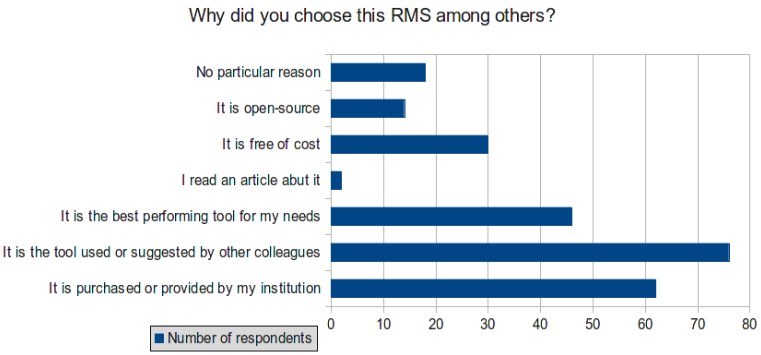


Figure 4

(2%) or connecting to other colleagues on the web (0%).

4.1.4 Training and support

Only 6% of respondents declared to have followed training sessions (figure 7 on page 157). The library seems external to these needs: only 13% of respondents state that they received help by the library in using the RMS, and they generally refer to the EndNote distribution of 2008 (figure 8 on page 157).

Of the 162 "no", 28 provided details, admitting that they just "never asked", or "never heard about any initiatives". When asked if they ever suggested the tool to other colleagues, the majority replied "yes" (63% against 37%: see figure 9 on page 158). The opposite happened towards the students: only 38% of respondents declare to have suggested a RMS to students (figure 10 on page 158). This answer comes from any type of academic role (professors, researchers, postdoc, research fellows, etc.)

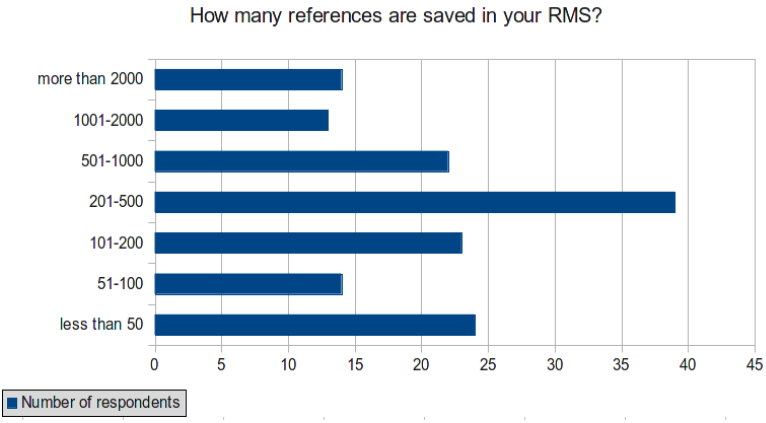


Figure 5

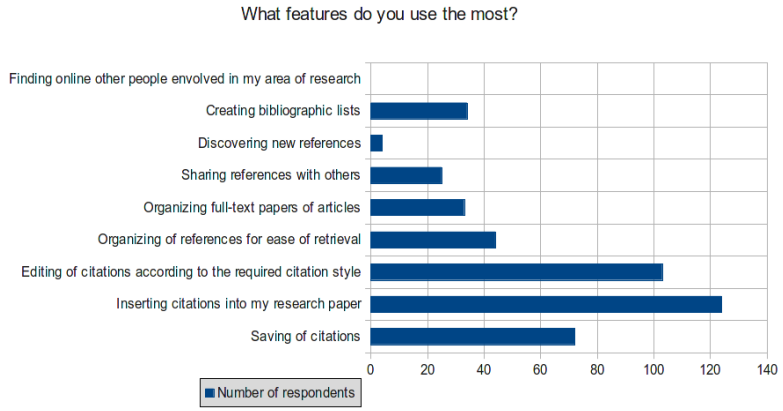


Figure 6

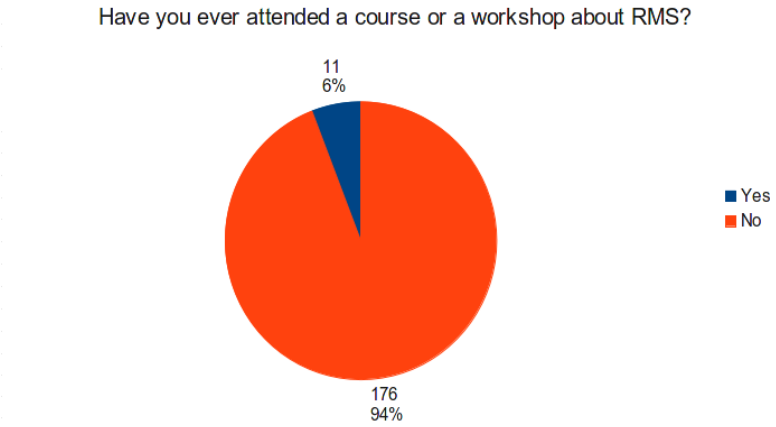


Figure 7

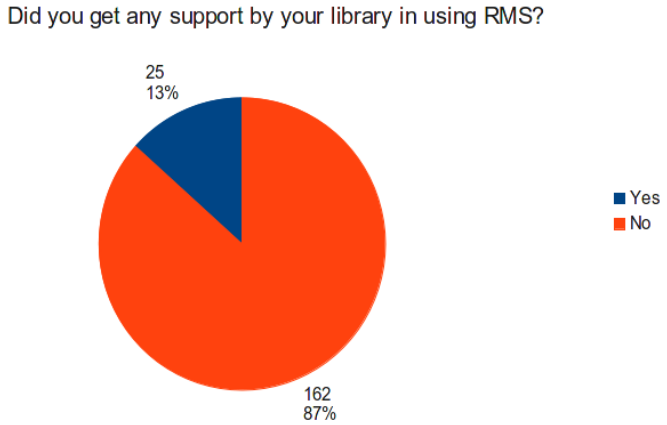


Figure 8

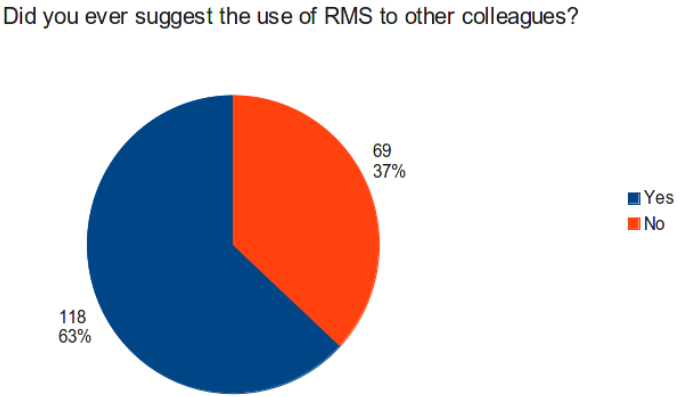


Figure 9

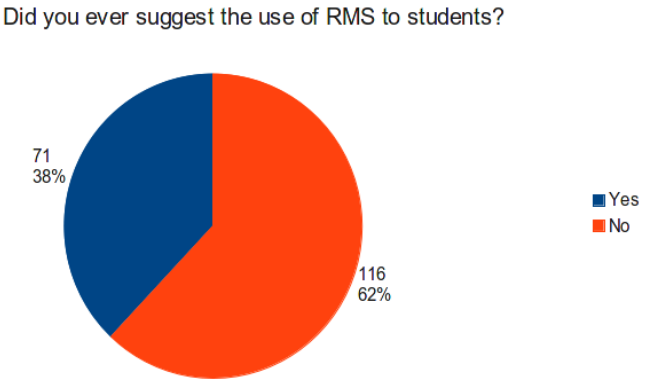


Figure 10

4.2 The interviews

4.2.1 What are your general knowledge and experience with RMS?

Respondents admit to use softwares in a very practical way to satisfy a very present need. Users never explore the most sophisticated features. Almost all respondents point out how crucial is the time factor in their work; for this reason there is no way to develop a strong mastery in the software. Some call it "laziness": "there is a particular laziness in every researcher: if something works, you don't feel the need for something else" (resp. 03). Easy and rapidity of use is highly valued: "When you are used to a certain product, you need a substitute with a very low learning curve" (resp. 05); "The reason I never change is one: habit. Upgrades are too much of a burden" (resp. 07).

Most respondents make clear that they seldom move away from a used and known product to discover or try a new one, and when they do they rarely feel satisfied. "I downloaded and installed 2 products, but I uninstalled them very soon: I did not understand them, I felt them unfriendly, they did not do what I needed" (resp. 10). Sometimes change is hard even when the software used has evident flaws: "EndNote is good and very powerful, but it has a lot of problems, it's heavy on your system, it needs high requirements and is extremely hard to move it across different platforms" (resp. 05).

A software is often chosen because already used or suggested by other colleagues (resp. 04: "I started using Reference Manager because it was used by a colleague with whom I worked when I was in another university") or because it's dominant in the community (resp. 03: "I use EndNote and ReferenceManager, because it is already available to all of us here in the lab, or because it is acquired

by your superior"). The technological context is also a key factor: according to the operating systems and word processor used, the most compliant software is adopted: "In physics, we all use LaTeX, so BibTeX comes naturally" (resp. 00).

It is interesting in this matter to note how EndNote was often already in use before the institutional purchase made in 2008. The institutional purchase just allowed the diffusion of more legal copies, removing the cost burden from the individuals and the departments.

Only two respondents adopted a different choice from what they found already available: instead of taking advantage of the available copies of EndNote, they sought different solutions. One because of the compliance with her system: "I decided to use Zotero because it works on Linux" (resp. 06); the other because of reasons related to the proprietary nature of EndNote: "I believe that in the university world we should use non proprietary software, so I looked for open-source - or at least free of cost - alternatives" (resp. 12).

4.2.2 What is your research workflow, and how does the RMS fit into it?

With the obvious differences due to disciplinary fields and community practices, all respondents show how strongly the RMS is related to the research and writing work-flow.

The research work-flow doesn't vary much among the respondents: a team leader usually wraps up all the contributions by the different collaborators and edit the final draft to submit to a journal. This sometimes explains the reasons behind the non-usage: when a researcher is not the project coordinator he doesn't take part in the bibliography editing: "Generally the supervisor actually writes the final paper, while we just run the experiments and collect the data" (resp. 02).

4.2.3 How do you consider virtual collaboration?

The approach towards systems of virtual collaboration is almost non existent. The only forms of virtual collaboration happen in a very traditional way: through email, sometimes through some sort of peer-to-peer communication system (such as Skype). Scholars often use cloud-based shared folders systems, like Dropbox, to share journal papers. One first reason is the lack of knowledge about the possibility itself to virtually collaborate with other colleagues. An interviewed (resp. 12) was a declared Mendeley user, but strangely he did not know about the social features which are the very heart of Mendeley.

Respondents showed a sort of diffidence about building an online presence to connect with other colleagues around the world through dedicated scientific networks. One is very clear: "It is impossible: in science, when working on the same subject, you either cooperate, or you compete. If you collaborate, it comes naturally to work with daily tools; if you compete, you are very careful not to put reveal, anticipate, or share your data" (resp. 03). Another one has the same opinion, even though a little more open to possibility: "There is no such thing as a virtual Alexandria. Data exchange is daily done within small groups" (resp. 09).

Only one respondent gave a very different opinion: "Collaboration is fundamental: our job is always been based on collaboration on a international level. An online tool, cloud-based, through which I can invite other people to contribute to an online list of references, would be of utter importance" (resp. 10). Another respondent shares this conception of science as an international collaborative endeavour, and looks positively to web platform that can act as a showcase for the scientific production: "I believe these sort of things - social networks, forums for mutual assistance - are very useful and interesting. It is very useful to be present and visible on the web

to communicate, to share informations, to ask questions to more experienced people" (resp. 11).

4.2.4 How much do you consider RMS as a fundamental tool for the academic work?

This direct question, aimed to probe the perceived importance of the tool across all the different ways of usage, gained a wide range of responses. One interviewed was a complete non-user of RMS, and provided an interesting chemistry-related metaphor to explain the reason: "I apply a principle of chemical kinetics: if you want to speed up a reaction, you have to modify the slow part of the process. I believe that in writing a research paper, the 'slow part' is not the reference management" (resp. 09). On the opposite scale, another respondent said: "They are absolutely essentials. It is crucial for the credibility of a paper to display properly formatted citations. I insist a lot on the reference check in the works of my group" (resp. 11). A middle-ground response is the following: "Nothing is really 100% essential; on a scale of importance from 1 to 10, I would deem the RMS as 8" (resp. 12). Ultimately, the importance of the instrument, considering these nuances, is generally considered high.

4.2.5 Have you ever suggested its usage to colleagues or students?

This two-sided question was made to understand the correlation between the user and his context, and see how the RMS is a node in a network. We saw that most people use a software because suggested by other colleagues, or because part of the work-flow of the labs or departments where they worked some moment in their career. The reverse action happens in the same way. Some suggest the usage of a tool when they have to coordinate a research group:

"I always suggest to use EndNote to those who collaborate with me, because I don't want to waste time in fixing citations" (resp. 10). There is always a very practical reason behind the behaviour: the topic is brought in when the need arises among people who need to collaborate; it is a rare topic of discussion among them outside the practical activity.

Somehow different is the approach towards students: respondents seem to be cut in two groups, those who consider essentials for students to learn how to use a RMS in the beginning of their career, and those who think that, before the PhD, students do not need such knowledge. "I always suggest to use a RMS, but I leave liberty of choice. This is helpful because it saves a lot of work to do in the end" (resp. 06); "You learn by doing. Generally I think it is useful to learn at the early stage of the career, as a sort of literacy" (resp. 03).

Other responses underline an opposite vision. "Master students already have many difficulties to face; introducing a sophisticated software like a RMS would be adding more trouble" (8). "I tried to explain the functions of RMS to students, but with no success" (4). This is often related to the actual need: "for a master thesis students don't need to handle so many citations to justify the learning of a specific tool" (5). Another response considers RMS as less important than other tools: "I would rather spend time to explain how PubMed works" (resp. 07).

4.2.6 What kind of support does the library give you? What kind of role do you expect from it?

Since one of the purposes of the present study is to help libraries to understand what libraries can do to assist their members in reference management, this was a key topic touched in the interviews.

When asked about their relationship with their academic library,

answers were generally like: "Libraries have disappeared from my life" (resp. 11); "I never go to the library: I do everything online, and if I need something more, I ask to colleagues and friends from other universities" (resp. 03). "We turn to the library staff only when we need documents not available online" (resp. 01). However, the importance of the digital library infrastructure is recognized: "I don't step into the library any more, but the library provides access to all the online resources I need" (resp. 09). Sometimes this infrastructure is invisible, and scholars do not realize what is there behind the online access sometimes the content available online is considered as just "free" (resp. 01).

About the RMS, some libraries just provided the licensed copy of EndNote, some support information, and not much more. This doesn't mean that the library service is judged negatively: a researcher explained: "I never asked for assistance, even though I know they are very kind and professional; I just prefer to overcome the difficulties by myself" (resp. 04).

About the possible role for libraries, a respondent gave an interesting answer: "Libraries cannot be just the keepers of knowledge any more. I know a lot of librarians who are willing and able to assume a more active role in the research process. But the institution must support this with proper funding and resources" (resp. 05). Other respondents wish for a more active role by the library: "Library could be very important in setting a standard within the institution; so far it never had this role, but it would be important if it starts having it" (resp. 10).

Other respondents show how low is the level of acquaintance with the library staff, therefore how distant are the libraries to some of their community members. This sort of "assumption" is confirmed by other respondents, who say "I don't know if they can provide help" (resp. 08) or "It's something I never thought about" (resp. 09).

4.2.7 Are you interested in training, support or information initiatives?

Most respondents reveal to be self-taught about these tools. Some consider themselves fine with this, and do not feel a deep need for special training about RMS. Others recognize their need for a specific and structured training. In general, training sessions, support initiatives and any sort of communications are considered welcome by almost all, despite their role or experience.

A common point of view emerged from all the respondents: training and information sessions have to be extremely practical and to-the-point. Nobody is interested in introductory sessions, generic informations, or such. They need to learn how to do things, how to solve the problems they face in their work. Their amount of time to dedicate is too small. This is the only factor which is always underlined, often with dramatic tones.

4.2.8 How do you value open-source when selecting such a software?

The RMS landscape shows a sort of competition between commercial-closed and open-free products [4] The main difference perceived is about money, not about technology. When asked if they ever consider an open-source software, respondents always interpreted it as choice between an expensive software and a free one. Sometimes the interviewed did not even seem to have a clear distinction between the two concepts of open-source and free-of-cost. Reliability and ease of use are the main aspects considered for a software: "It must be stable and performing, otherwise it makes no sense" (resp. 08);

[4] See for example the legal dispute between EndNote and Zotero: http://www.ncbi.nlm.nih.gov/pubmed/18843308 and http://quintessenceofham.org/2009/06/04/thomson-reuters-lawsuit-dismissed.

"Free software are interesting also because they are easy: easy to obtain, to distribute, to copy, to patch, to upgrade. Limits put by commercial licenses push towards piracy" (resp. 10).

Other respondents gave extremely clear replies about the importance of open-source: "I believe the university must move on the open-source ground not for economic reasons, but because it's in its nature. What counts is the sharing and participation culture. I think that a researcher must have a wider vision of things: I always try put a conscience in what I do, thinking about the cause and impact of my actions" (resp. 12). "I consider open-source, not only because it's free of cost. I consider it as an element of evaluation: I never suggest to use closed products, because it creates difficulties in sharing products, data, contents" (resp. 06).

It is nice to see some clarity about the link between sharing science and sharing technology: "I am interested in the idea of open-source: I like the fact that people cooperate as a community without a business view, especially in the academics where knowledge has to be shared" (resp. 04); "We work in scientific research without commercial purposes: it is hard for me to accept the idea of producing knowledge for an economic payback" (resp. 03).

4.3 Analysis and discussion

4.3.1 Awareness and usage distribution

Awareness is relatively high in terms of quantity: 92% of people know about RMS. It is low in terms of quality: very few are the known softwares. The percentage of actual users is a little lower: 75% of the respondents are active users. It can be openly said that RMS are widely used across UniTo, although the 25% of non-users constitute a sack of resistance not to be underestimated. The questionnaire clearly declares EndNote as the most used software, fol-

lowed by a very low range of alternatives: Mendeley, BibTeX, Zotero, Reference Manager, all of them with incomparable low numbers.

4.3.2 Basic practical approach

RMS are used when needed (when writing a paper which requires a reasonable number of references) and they are used in their basic functions. This explains the numbers emerged in the questionnaire, which shows a very basic need underlining its usage. Participants in the survey do not show interest in the technological implications of the tool, as long as it works fine. This leads to be closed against additional extended features, or to paradigm changes: the ignorance about the world of virtual science and networking collaboration explains how little today scholars are aware of the opportunities provided to scientists by the web environment.

4.3.3 Time factor

One concept emerges very strongly from the interviews: time is a crucial factor in everything. Everything in the process must speed researcher's work and save time. This applies to all the aspects: choice of a software and discovery, deep knowledge of its functionalities, training and learning sessions. This also explains the numbers of the questionnaire: few softwares known or used, basic functionalities used, little contact with the library asked or desired etc. It is worth noticing that although citation management is often rooted in the research process, it is often perceived as an element of minor importance. It is also true, on the other hand, that a more proper training on RMS could help saving time: some interviewees point to this when considering the benefits of such skills.

4.3.4 Habit

A general laziness, or force of habit, prevents change. This attitude, openly admitted but the respondents, prevents scholars to discover new products or new features. When a RMS is used, generally it's because a former experience by some colleagues proves it useful: it seems unlikely that someone is willing to experiment something new on his own. When this happens, it generally leads to frustrating and unsuccessful experiences. The numeric data are made stronger by the responses to the interviews, which show how low are the range of softwares actually used and the curiosity for different alternatives, due to the time and need factors discussed above. Finally, the fact that the University acquired and distributed licenses of EndNote made the faculties stick with this software without worrying about other alternatives. Now that the licenses are not purchased any more, it will be interesting to see how scholars will change their approach.

4.3.5 Economic issues

Economic issues are always important, even when selecting a software. Everything that can save money is welcome: this applies to softwares as well.

Yet this seems true more on the intentions than in the practice: the economic issue is stressed by all the interviewees, but only 16% of the participants in the questionnaire actually indicate it as a reason of choice. The habit of already-in-use tools is stronger than the need to move on better instruments. Often the economic constraint is not strong enough to push people to experiment alternatives. More general implications of a software license, such as long term costs and technological impact, are not considered, as shown by the general lack of awareness on the open-source topic.

4.3.6 Training and literacy

If we compare the answer to the questionnaire, which says that 87% never received or asked any support, with the interviews responses, which show how basic is the general knowledge of the tools and their functionalities, it is clear how much impact has the lack of specific training. Even if not stated explicitly, there is need for training and literacy. Results clearly show how low is the awareness because scholars do not know RMS at all and do not have time to go deeper and improve their skills beyond the self-taught basics.

There are no common practices in the training to RMS: the usage of a RMS is more part of a "tacit knowledge" present in the research environment, rather than a conscious part of the set of skills and methods of a researcher.

It is remarkable how every concept examined so far - shallow knowledge, time constraints, economic awareness - can be considered within a set of aimed training initiatives.

Given this, any kind of training must be tailored to the actual needs. If RMS serve the purpose of facilitating the research process and saving time, any training on it must not go in the opposite direction.

Students might benefit from a specific training in RMS as part of their academic information literacy. The strong stress given by some interviewees about this, nevertheless, doesn't match with the percentage of those who actually suggest a RMS to their students (38%).

4.3.7 Library role

Librarians, as information experts, must have a more active role in RMS support. But this role must considered in the more general context of the library impact in a community. The survey shows that

library staff skills are mostly not perceived, therefore scholars are alone when they face reference management issues. This creates a separation between the library and the academics instead of bringing a mutual dependence.

There is a lot of room for the library to be active in this process. Responses let emerge needs such as: information, training, guidance. Library is not the keeper of resources any more, but also the keeper of bibliographic tools. RMS require a lot of time and skills that researcher seldom have; a professional expert in these tools could help the scholars guiding them across the wide range of packages, across the basic functions, focusing on problem-solving activities. This could be an extremely cost-benefit effective initiative. If the library assumes the role of information assistants and technology experts, it can be the link between the world of technological information solutions and researchers' needs.

This considerations confirm what is said in the literature. East already noted the relationship between bibliographic support and reference management training. He recognizes "the well-established role of the library in training researchers in searching electronic databases and downloading retrieved references. From here it was only a short step to beginning to train researchers in the management of those references" (East 65). This has not happened yet at the University of Torino, but the survey suggests that it should, and that a loud call for a new commitment is given.

4.4 Conclusions

This survey confirms the wide presence of RMS in scholars' research work-flows, but it points out the lack of information about it. RMS features, from the basic reference management to the advanced virtual collaboration, are adopted much below their potentials. Scholars rely on common practice and word of mouth rather than on

specific training upon the tools. The general users' behaviour confirms what other studies proved: information-handling habits of researchers are personal and often rudimentary. In this area, information professionals in libraries have a great chance to assume the role of assistants to improve researcher's efficiency in managing the literature.

The data collected refer to a specific area of a single Italian university. Further studies should perform the same type of inquiry in different settings and provide a wider cross-institutional analysis.

Also, having proved that habit is a strong factor, searching for patterns of behaviour among different age ranges could lead to important understanding on how the phenomenon is likely to change in the next future. The technological development could have a key role in this, and it should be monitored closely.

References

Alhoori, Hamed and Richard Furuta. "Understanding the Dynamic Scholarly Research Needs and Behavior as Applied to Social Reference Management". *Research and Advanced Technology for Digital Libraries International Conference on Theory and Practice of Digital Libraries, TPDL 2011, Berlin, Germany, September 26-28, 2011*. (Cit. on p. 146). Print.

Borgman, Christine L. *From Gutenberg to the Global Information Infrastructure: Access to Information in the Networked World*. Cambridge, MA: MIT Press, 2003. (Cit. on p. 146). Print.

Bos, Nathan. "From Shared Databases to Communities of Practice: Taxonomy of Collaboratories". *Journal of Computed-Mediated Communication* 12.2 (2007). (Cit. on p. 146). Print.

Butros, Amy and Sally Taylor. "Managing Information: Evaluating and Selecting Citation Management Software, a Look at EndNote, RefWorks, Mendeley and Zotero". *Netting Knowledge: Two Hemispheres/One World: Proceedings of the 36th IAMSLIC Annual Conference*. 2010. (Cit. on p. 147). Print.

Childress, Dawn. "Citation Tools in Academic Libraries: Best Practices for Reference and Instruction". *Reference & User Services Quarterly* 51.2 (2011): 143–152. (Cit. on p. 147). Print.

Cooke, Nicole A. "Internet Resources - Citation Management 2.0". *Public Services Quarterly* 6.4 (2010): 360–372. (Cit. on p. 149). Print.

Corbetta, Piergiorgio. *Metodologia e Tecniche Della Ricerca Sociale*. Bologna: Il Mulino, 1999. (Cit. on p. 150). Print.

Crowley, Emma and Chris Spencer. "Library Resources: Procurement, Innovation and Exploitation in a Digital World". *University Libraries and Digital Learning Environment*. Ed. Penny Dale, Jil Beard, and Matt Holland. Farnham: Ashgate, 2011. 215–238. (Cit. on p. 149). Print.

East, John W. "Academic Libraries and the Provision of Support for Users of Personal Bibliographic Software". *LASIE: Library Automated Systems Information Exchange* 32.1 (2001): 64–70. (Cit. on pp. 149, 170). Print.

Fitzgibbons, Megan and Deborah Meert. "Are Bibliographic Management Software Search Interfaces Reliable? A Comparison Between Search Results Obtained Using Database Interfaces and the EndNote Online Search Function". *The Journal of Academic Librarianship* 36.2 (2010): 144–150. (Cit. on p. 147). Print.

Fourie, Ina. "Librarians Alert: How Can We Exploit What Is Happening with Personal Information Management (PIM), Reference Management and Related Issues?" *Library Hi Tech* 29.3 (2011): 550–556. (Cit. on p. 146). Print.

Francese, Enrico. "The Usage of Reference Management Software (RMS) in an Academic Environment?: A Survey at Tallinn University". *Advances on Information Processing and Management* 1 (2012): 293–296. (Cit. on p. 149). Print.

Giglia, Elena. "Academic Social Networks: It's Time to Change the Way We Do Research". *European journal of physical and rehabilitation medicine* 47.2 (2011): 345–350. (Cit. on p. 148). Print.

Gilmour, Ron and Laura Cobus-Kuo. "Reference Management Software: a Comparative Analysis of Four Products". *Issues in Science and Technology Librarianship* 66.6 (2011): 63–75. (Cit. on p. 147). Print.

Haglund, Lotta and Per Olsson. "The Impact on University Libraries of Changes in Information Behavior Among Academic Researchers: A Multiple Case Study". *The Journal of Academic Librarianship* 34.1 (2008): 52–59. (Cit. on p. 148). Print.

Hensley, Merinda Kaye and M. Kathleen Kern. "Citation Management Software: Features and Futures". *Reference & User Services Quarterly* 50.3 (2011): 204–208. (Cit. on p. 147). Print.

Hull, Duncan, Steve R. Pettifer, and Douglas B Kell. "Defrosting the Digital Library: Bibliographic Tools for the Next Generation Web". *PLoS computational biology* 4.10 (2008). (Cit. on p. 148). Print.

Lawrence, D. and S. Ashwell. "Reference Management Software. Libraries Can Help You. . . and They Do". *BMJ. British Medical Journal* 307.6903 (1993): 569. (Cit. on p. 150). Print.

Martin, Justine L. "Course Instructor Perceptions of Computer-generated Bibliographic Citations". *Reference Services Review* 37.3 (2009): 304–312. (Cit. on p. 149). Print.

McMinn, H. Stephen. "Library Support of Bibliographic Management Tools: a Review". *Reference Services Review* 39.2 (2011): 278–302. (Cit. on p. 150). Print.

Niu, Xi. "National Study of Information Seeking Behavior of Academic Researchers in the United States". *Journal of the American Society for Information Science and Technology* 61.5 (2010): 869–890. (Cit. on p. 149). Print.

Ollé, Candela and Angel Borrego. "Librarians' Perceptions on the Use of Electronic Resources at Catalan Academic Libraries: Results of a Focus Group". *New Library World* 111.1-2 (2010): 46–54. (Cit. on p. 148). Print.

Palmer, Carole L., Lauren C. Teffeau, and Carrie M. Pirmann. *Scholarly Information Practices in the Online Environment*. Dublin: OH, 2009. (Cit. on p. 145). Print.

Patton, Michael Quinn. *Qualitative Research and Evaluation Methods*. 3rd ed. Thousand Oaks, CA: Sage, 2002. (Cit. on p. 150). Print.

Siegler, Sharon and Brian Simboli. "EndNote at Lehigh". *Issues in Science and Technology Librarianship* 34 (2002). (Cit. on p. 149). Print.

Steele, Susan E. "Bibliographic Citation Management Software as a Tool for Building Knowledge". *Journal of Wound Ostomy & Continence Nursing* 35.5 (2008): 463–468. (Cit. on p. 149). Print.

Strauss, Anselm and Juliet Corbin. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theorys*. London: Sage, 1998. (Cit. on p. 150). Print.

Van Ullen, Mary K. and Jane Kessler. "Citation Generators: Generating Bibliographies for the Next Generation". *Journal of academic librarianship* 31.4 (2005): 310–316. (Cit. on p. 148). Print.

Voss, Alexander and Rob Procter. "Virtual Research Environments in Scholarly Work and Communications". *Library Hi Tech* 27.2 (2009): 174–190. (Cit. on p. 146). Print.

ENRICO FRANCESE, Università degli Studi di Torino.
efrancese@gmail.com

Francese, E. "Usage of Reference Management Software at the University of Torino". *JLIS.it*. Vol. 4, n. 2 (Luglio/July 2013): Art: #8679. DOI: 10.4403/jlis.it-8679. Web.

ABSTRACT: The present research, originally a master thesis, aims to investigate the popularity and usage of Reference Management Softwares among researchers and scholars of the University of Torino, Italy, and the role that university libraries can assume about the subject. This study, based upon a qualitative approach, is a descriptive survey composed of an online questionnaire and direct interviews addressed to the population of professors and researchers of the STM areas at the University of Torino. A qualitative analysis was made across the 187 responses from the questionnaire and the 13 interviews performed. 7 key concepts were outlined and discussed.The knowledge of Reference Manage Softwares is high among the respondents, but their adoption is not. EndNote is the most known and used software, while other alternatives are more scarcely considered. Scholars, hindered by time issues, rely on old habits and are very unlikely to discover new ways to manage the literature they need. Virtual collaboration is absent from the common research practice. The research gives light on the users' behaviour in a large Italian university, confirming the results provided by the literature. Librarians should assist scholars by providing informations and support about the proper tools to improve the research process.

KEYWORDS: Academic libraries; Citation management; Reference management software; User behaviour.

ACKNOWLEDGMENT: This research was originally performed as part of the Master Thesis for the Digital Lirbary and Learning (DILL) Master Program 2010-2012. Special thanks go to prof. Pat Dixon.

Submitted: 2012-11-30
Accepted: 2013-02-01
Published: 2013-07-01

JLIS.it

Obtaining the Dewey Decimal Classification Number from other databases: a catalog clean-up project

Stefano Bargioni, Michele Caputo,
Alberto Gambardella, Luigi Gentile

1 Introduction

The Library of the Pontifical University Santa Croce[1] is a research library that belongs to the URBE (Roman Union of Ecclesiastical Libraries) Network.[2] It possesses approximately 167,000 volumes corresponding to 145,000 bibliographical records cataloged in MARC21 format. In order to manage the library there have been three Integrated Library Systems (ILS): Aleph 300, Amicus 3.5.4, and the current Koha 3.2.7. With the implementation of the open source ILS, Koha,[3] authority records were then introduced thanks to the software's advanced productivity. Moreover, Koha's flexibility has enabled the opening of new avenues for experimentation which is ordinarily impossible with a commercial ILS. With the intention of providing users with greater tools for catalog research from a semantic point of view, and bearing in mind that subject cataloging

[1] http://www.pusc.it/bib.
[2] http://www.urbe.it.
[3] http://koha-community.org.

JLIS.it. Vol. 4, n. 2 (Luglio/July 2013).
DOI: 10.4403/jlis.it-8766

based on the Nuovo Soggettario of the Central National Library of Florence is fairly recent, the decision was made to develop the potential inherent in the Dewey decimal classification,[4] which had already been partially implemented in the library for about ten years and was assigned to approximately 25% of the documents in the library's patrimony. Thus, the idea developed to increase the use of the Dewey decimal classification in the bibliographical records by importing the relevant information from other databases,[5] using the International Standard Book Number (ISBN)[6] as a key for the retrieval of missing numbers. We began by identifying the sources (databases) that would significantly meet our needs, both in terms of quality and quantity. The practice of copy cataloging — one of Koha's strengths — was fundamental in this regard. Once both the national and international resources were determined, methods were identified through which it was possible to access the information therein by means of a program. As the various institutions use different modalities to publish their data, it became necessary to diversify the query methods in order to systematically gain access to the relevant information. These ranged from the more modern example of OCLC, which gave rise to Classify,[7] a specific experimental web service for classification, to less simple cases of information retrieval from HTML pages. In order to ensure the quality of the

[4] http://dewey.info.

[5] The importation of data from other bibliographical sources is justified by the"principle of sharing", which one finds in public catalogs. This principle establishes the exchange of information through OPAC, Z39.50, web interfaces, etc. It also has as its objective the comparison and mutual control of registration and identification of the library information source, confirmed by field 035 in MARC21, for example. The importation of data occurred in accordance with the possible conditions or warnings expressed on the webpages of the queried sites. The case of a commercial use of the information retrieved could be different.

[6] http://www.isbn.org/standards/home/index.asp.

[7] http://classify.oclc.org.

Dewey classification numbers obtained, a special algorithm was created, which is described in the section "Quality Control". The process of searching for and importing data was also analyzed under the stress it incurred both for the system that was the source of the data as well as for our Koha system. Queries of the servers cannot occur at an excessive pace, and that is why some of them expressly issue warnings to any possible software, such as crawlers or web robots, which access them.

2 Identifying the records to be modified

The records to be enhanced contained an ISBN (tag 020), but lacked a Dewey number (tag 082). They may be identified in Koha through an SQL query (listing 1) that is specific to the MySQL database and which is applied to the `marcxml`[8] field of the `biblioitems` table.[9]

Listing 1: Query for record identification in Koha.

```
SELECT biblionumber, ISBNlist
FROM biblioitems
WHERE isbn_present
AND dewey_absent
AND language_008='...'
```

Since it was not an index driven search, the retrieval occurred via a record by record analysis of the database. This is an aspect of the project which depends on the computing power of the server hosting the ILS. Other ILSes allow to find the system number and the ISBN of a record without a Dewey classification number in a

[8]The field biblioitems.marcxml contains a display of the bibliographical record in MARCXML format (http://www.loc.gov/standards/marcxml, http://en.wikipedia.org/wiki/MARC_standards#MARCXML.)

[9]Main elements of the query are described in Table 9 on page 194.

manner quite different from that of Koha, due to the data structure used to store bibliographical data and the tools available to access it.

3 Sources

The ISBNs of each record, extracted from the query, were used to search seven different databases. The sources selected are listed in Table 1 in the temporal order of the query.

1	Classify	OCLC Classify
2	LC	Library of Congress
3	BNF	Bibliothèque nationale de France
4	DNB	Deutsche Nationalbibliothek
5	BNCF	Biblioteca Nazionale Centrale di Firenze
6	BNCR	Biblioteca Nazionale Centrale di Roma
7	BNB	British National Bibliography

Table 1: Dewey classification sources queried.

As the purpose of our work was essentially practical, no attempt was made to query each source with the same ISBN. In the event that a Dewey Decimal number should be retrieved and saved in a record, it was decided that each particular source would take priority over those following, so that the record would not be further processed. This way seemed more efficient to us than the other two possibilities, i.e. to query all sources with the same ISBN, either simultaneously or in succession. Moreover, in several cases the search was limited to the predominant language of the source queried, both to avoid an excessive number of searches and because it was deemed more reliable. Among the languages present in the catalog, Spanish language was not incorporated due to the absence of databases we consider adequate for this purpose. The method adopted does not allow for

comparisons among different sources on equal terms. However, a statistical analysis regarding the use of the Dewey number in different sources is still made possible, as will be seen later.
Figure 1 shows the address, the type of data returned, the type of service contacted for each source, and the language involved.
Sources other than the web provide connection data on their re-

	Source	Type	Connection	Query	Data	Language
1	Classify	REST	http://classify.oclc.org/classify2/Classify?summary=false&isbn=*ISBN*		XML	All
2	LC	Z39.50	lx2.loc.gov:210/LCDB	find @attr 1=7 *ISBN*	MARC	All
3	BNF	Z39.50	z3950.bnf.fr:2211/TOUT-UTF8 user ID: Z3950/password: Z3950_BNF	find @attr 1=7 *ISBN*	MARC	All
4	DNB	web	https://portal.dnb.de/opac.htm?query=isbn%3D*ISBN*&method=simpleSearch		HTML	ger
5	BNCF	web	http://opac.bncf.firenze.sbn.it/opac/controller.jsp?action=search_avanzatasearch&query_fieldname_1=keywords&fieldname=identifkw&query_querystring_1=identifkw%40%3A*ISBN*		HTML	ita
6	BNCR	web	http://193.206.215.17/BVE/result.php?textexpert=is%3DISBN&comebackb=1&dove=completa&numschede=10&ordschede=4+ia&lastRefinedQueryRPN=is%3DISBN&cercato=is%3DISBN&numresults=1&startp=esperta&query4usr=is%3D*ISBN*&formatoAna=3&vaiformat=Esequi		HTML	ita
7	BNB	Z39.50	z3950cat.bl.uk:9909/BNB03U requires credentials, given upon request	find @attr 1=7 *ISBN*	MARC	eng

Figure 1: Characteristics of the queried sources containing a Dewey decimal classification number.

spective pages explaining the service. For web sources, however, connection and query are generally collected from the Advanced Search screen of the catalog. In order to identify the parameters to be sent, including the ISBN, one may proceed in any of the ways listed in the second paragraph of the Appendix. In the case of web pages, the technique adopted for the extraction of data is very specific. One

must apply what is normally referred to as web scraping,[10] screen scraping, or more commonly data scraping. Essentially, it is necessary to understand whether one has a method for locating and extracting the data of interest from within the HTML code obtained. This operation is easier and more standardized when responses provided are structured data. Web 2.0 and even more, linked data, lead to the hope that the data sources might offer not only web interfaces, essentially intended for human use, but above all, interfaces with standard structured responses that are serviceable by other machines and stable in time.
The logic used in the programs to query the data sources can be explained in the algorithm represented in figure 2.

```
open the connection to the bibliographical database
obtain the ISBN from records without a Dewey number
open the connection to the data source, if Z39.50
for each ISBN
      query the data source according to the current ISBN
      if a Dewey number is available in the response
            if the Dewey number passes quality control
                  update the bibliographical record
      wait to avoid overloading
close the connection to the data source, if Z39.50
close the connection to the bibliographical database
```

Figure 2: Representation of the logic used in the programs to query the data sources.

There is an exception in the case of Classify, which has already been mentioned above. The process of "querying the data source by the current ISBN" must be followed by:
The third paragraph of the Appendix provides examples for each of the three types of data obtained as a response: XML, MARC, and HTML.

[10] http://en.wikipedia.org/wiki/Web_scraping.

```
if the ISBN corresponds to several works
      repeat the query in relation to the first work
```

Figure 3: Representation of exception in the case of Classify.

The response in Classify[11] typically falls into four categories, as per the table 2.

Response code	Meaning
2	ISBN corresponds to a single work
4	ISBN corresponds to several works
101	ISBN incorrect
102	ISBN not found

Table 2: Categories of responses in Classify.

In the event of the response "ISBN corresponds to several works," Classify[12] provides a list of OCLC# identifiers for related works. The first of these was preferred while locating the detailed record through its OCLC# with another query such as: http://classify.oclc.org/classify2/Classify?summary=false&swid=OCLC#.
This generates a response in the form of code 2, "ISBN corresponds to a single work".
Classify's response in the case of a single work (an example of which can be seen in paragraph 3 of the Appendix) reports both

[11]Classify APIs are described at http://classify.oclc.org/classify2/api_docs/index.html and may be tested through the Classify API Explorer at http://classify.oclc.org/classify2/api_docs/classify.html.

[12]The aggregations in Classify occur through the application of FRBR.
At http://www.oclc.org/research/activities/classify.html it states: "Bibliographic records are grouped using the OCLC FRBR Work-Set algorithm http://www.oclc.org/research/activities/frbralgorithm.html to form a work-level summary of the class numbers and subject headings assigned to a work. You can retrieve a summary by ISBN, ISSN, UPC, OCLC number, author/title, or subject heading."

the combination of the Dewey number and the LCC classification assigned to a particular work by the many catalogs which contribute to OCLC, as well as a list of editions containing the classification number. It seemed preferable to import the number from the first edition in the list as it was often more complete in comparison with the others.
Z39.50 sources essentially require extracting the tag value of the Dewey number, according to the rules of the corresponding MARC format, as shown in Table 3.

MARC format	tag	code subfield	edition subfield
MARC21	082	a	2
InterMARC or UNIMARC	676	a	v

Table 3: Tags for the Dewey classification number in some MARC formats.

4 Quality Control

Before the project, Dewey decimal numbers referencing editions 19 through 23 populated the catalog. The decision not to introduce classification numbers from the abridged version or classification numbers of Dewey editions below 19 meant having to give up several classification numbers found, as reported in the statistics in Table 7 on page 188. Priority was assigned to quality rather than quantity in order to effect an enhancement that is more suitably aligned with the cataloging approach. In reality, beyond limiting the edition to 19 or higher, classification numbers with indicators 1 and

2 — different from "0 0" and "0 4"[13] — were discarded.
Classification codes containing non-numerical characters or lacking an edition were also discarded. Finally, classification numbers were standardized before being saved into the record.

5 Tag 035

While updating the record, it seemed appropriate to keep track of the details from which the imported Dewey classification number was obtained with the help of tag 035 in MARC21, as in the following example:

Listing 2: Example of the use of MARC21 tag 035.

```
00872nam a2200265 i 4500
001 000000035650
003 IT-RoPUS
005 20121121122621.0
008 041027r19851982xxk u000 u eng c
020 $a 0198247761
035 $a (OCoLC)007946090
040 $a IT-RoPUS $b ita
082 04 $a 111.85 $2 19
100 1 $a Savile, Anthony. $9 70779
245 14 $a The test of time : $b an essay in philosophical
    aesthetics / $c Anthony Savile.
...
```

[13] According to MARC21, the first indicator of field 082 with a "0" value signals use of the complete Dewey edition; the second indicator with a "0" value points to a Dewey number assigned by the Library of Congress, while the value "4" corresponds to a notation assigned by an agency other than the Library of Congress.

In the case of a non-MARC21 source, or one without a MARC Organization Code,[14] it was decided to assign the most logical code possible, as shown in Table 4.

Table 4: Institution codes in 035.

1	Classify di OCLC	OCoLC	Official
2	Library of Congress	DLC	Official
3	Bibliothèque nationale de France	FR-PaBFM	Official
4	Deutsche Nationalbibliothek	DE-101	Official[a]
5	Biblioteca Nazionale Centrale di Firenze	BNCF	Unofficial
6	Biblioteca Nazionale Centrale di Roma	BNCR	Unofficial
7	British National Bibliography	BNB	Unofficial

[a] http://dispatch.opac.d-nb.de/DB=1.2/LNG=EN.

The ID was derived from the record, which in each case appeared in different locations. For Z39.50 sources it is located in tag 001, while the Library of Congress makes use of tag 010. Classify also specifically reports this ID in the XML record, while ID retrieval from records in HTML format is particularly complex. This decision enables the linking of the bibliographical record to that of an external catalog. This is useful for creating a link of interest whether at the level of OPAC (figure 4 on page 186) or linked data.

A link in the OPAC is created — for every occurrence of tag 035 — on the basis of the links from Table 5 on the facing page. The permanence of some is ensured (permalink). In other cases, the link — which is essentially unstable — can be created from the simple view of each individual record offered in the catalog.

[14] http://www.loc.gov/marc/organizations.

Table 5: Creation of a link in OPAC from an occurrence of tag 035.

Classify di OCLC - WorldCat	`http://www.worldcat.org/search?q=no%3A`*ID*	permalink[a]
Library of Congress	`http://lccn.loc.gov/`*ID*	permalink[b]
Bibliothèque nationale de France	`http://catalogue.bnf.fr/servlet/biblio ?idNoeud=1&SN1=0&SN2=0&host=catalogue& ID=`*ID*	
Deutsche Nationalbibliothek	`http://d-nb.info/`*ID*	permalink[c]
Biblioteca Nazionale Centrale di Firenze	`http://opac.bncf.firenze.sbn.it/opac/ controller.jsp? action=notizia_view&notizia_idn=`*ID*	
Biblioteca Nazionale Centrale di Roma	`http://193.206.215.17/BVE/ricercaEsperta. php?dove=esperta &cerca=Avvia+la+ricerca& textexpert=di%3D`*ID*	
British National Bibliography	`http://search.bl.uk/primo_library/libweb /action/search.do?vid=BLBNB&fn =search&vl%28freeText0%29=`*ID*	

[a] http://www.oclc.org/worldcatorg/linking/how.htm#oclc-number.
[b] http://lccn.loc.gov/lccnperm-faq.html.
[c] Concluded from the simple view of a single record at the end of any type of search.

Figure 4: A record in the OPAC, enriched with Dewey and a link to DNB.

6 Delay while searching sources

As mentioned in the Introduction, continual use may burden the queried server. This happens quite easily in the case of automated searches. Web pages such as "Terms and Conditions" allow the sources' terms of use to be regulated. For example, the Library of Congress explicitly[15] requires that crawlers use the Z39.50 server at a rate below 10 queries per minute. The Z39.50 server of the Bibliothèque nationale de France shuts down the connection after the tenth query. The program must then reopen the connection with the same frequency. It is not possible for the website of the Biblioteca nazionale centrale di Firenze to be accessed uninterruptedly, since it seems to be overloaded almost immediately. It is also opportune to verify, through the sources queried via http protocol, whether or not there are indications for crawlers in the file `/robots.txt`. At times restrictions on the frequency of access[16] can also be found.

Therefore, a wait time of 4 to 6 seconds between queries was es-

[15] http://lccn.loc.gov/lccnperm-faq.html#n12.

[16] http://en.wikipedia.org/wiki/Robots_exclusion_standard#Crawl-delay_directive.

tablished for all sources. These intervals prevented our catalog from being overloaded as well. In fact, after every record modification, the Zebra[17] search engine used by Koha and the independent search engine[18] for ordered lists update their indexes and may slow down both the consultation of OPAC and regular usage. This is one aspect which must be assessed on the basis of the available processing power.

The pace imposed by these intervals actually prolongs the import process by hours if not days, depending upon the quantity of ISBNs to be processed. Such a pace may require adjustments to the program, e.g., by setting up parameters so that it operates only on a certain timetable.

7 Log

The import process was monitored in order to collect statistics on the work carried out. The types of log records listed in table 6 on the next page were recorded.

[17] http://www.indexdata.dk/zebra.

[18] At present, scrollable index searches – also known as browse searches - are not available in Koha. It was possible to add this feature to our installation of Koha through an application based on Solr [http://www.lucene.apache.org/solr/] and developed by our library. This browse feature was presented at the international meeting of Koha users held in Edinburgh in June 2012 [http://www.wiki.koha-community.org/wiki/KohaCon12_Schedule#Adding_browse_to_Koha_using_Solr_.2815-20_min.29], and will be integrated into incoming versions of Koha, particularly when Solr will be an alternative to, or substitute for, Zebra.

1	*System number*	*ISBN*		ISBN not found
2	*System number*	*ISBN*		ISBN incorrect
3	*System number*	*ISBN*		ISBN related to several works
4	*System number*	*ISBN*		Dewey number not found
5	*System number*	*ISBN*	Classification # and edition found	Unsatisfactory
6	*System number*	*ISBN*	Classification # and edition found	Record modified

Table 6: Types of log records; types 2 and 3 are only related to Classify.

8 Statistics

The generated logs facilitated the creation of the following tables and certain comparisons between the different sources.

Fonte	Language	Record scanned	Record modified	ISBN not found	Dewey # not found	Dewey # discarded	Several works with same ISBN	ISBN incorrect
Classify	all	42387	10267	5321	6607	20059	8240	133
LC	all	31999	1252	21195	8562	1011		
BNF	all	30903	2253	21327	7268	55		
DNB	ger	4193	163	3867	163	0		
BNCF	ita	12017	4088	3643	3542	744		
BNCR	ita	7549	1515	3003	2978	53		
BNB	eng	6215	193	5449	55	518		
Total			19710					

Table 7: Calculations.

Source	Samples	Ed. 19 (%)	Ed. 20 (%)	Ed. 21 (%)	Ed. 22 (%)	Ed. 23 (%)
Classify	10267	19,86	23,03	36,18	20,13	0,79
LC	1231	28,11	25,83	24,29	19,58	2,19
BNF	2253	0,00	0,09	0,36	99,56	0,00
DNB	163	0,00	0,00	0,00	100,00	0,00
BNCF	4088	9,10	23,46	55,04	12,40	0,00
BNCR	1515	2,38	9,70	87,92	0,00	0,00
BNB	193	16,58	19,69	26,42	28,50	8,81
Total	19710					

Table 8: Distribution of editions related to the classification numbers found.

Table 8 has been reproduced in the graphs compiled in Figure 5,

one for each source.

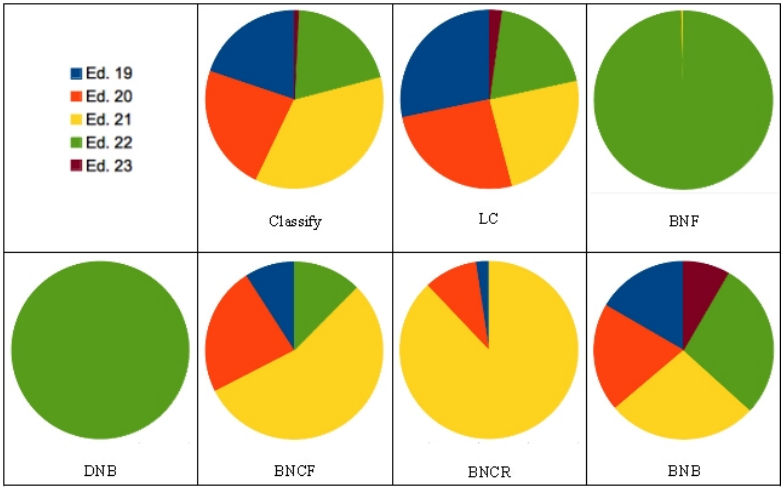


Figure 5: Distribution of editions.

We see here some definite choices (e.g. BNF, DNB, and BNCR), which give priority to a single edition. On the other hand, with the amount of different editions reported by Classify, those who have used the Dewey number for some time do not seem to have allowed for an update of Dewey notations in the catalog. Certainly, this can be ascribed to the complexity of the operation. Finally, we note the (still) low usage of edition 23. As previously stated, the catalog has been increased by 19,710 new Dewey classification numbers in just as many bibliographical records. The increase amounted to 47.8%, given that earlier records with tag 082 totaled 41,255. The current distribution of Dewey numbers, shown in figure 7 on page 191, outline a profile of the library's holdings, reflecting the areas of interest in the University's different schools as well as the library's growth.

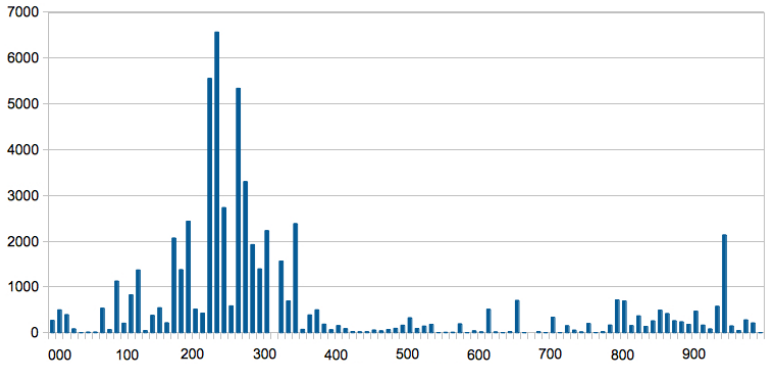


Figure 6: Distribution of library works as a function of Dewey classification subdivisions.

Figure 7 on the facing page represents the distribution of Dewey editions in the catalog. The absence of editions for a significant number of bibliographical records is a case of inhomogeneous cataloging. To address this situation, a method of retrieval very similar to the one outlined in the present work could be used.

9 The Dewey Index in the OPAC

Through scrolling indexes, shown in Figure 8 on page 192 and mentioned in note 21, it is possible in the OPAC to offer a path of semantic search based upon the Dewey classification number. The search counts performed by users show that the index of greatest use is precisely the Dewey decimal classification, even higher than the name index, which is of particular importance for references of ancient authors as well as popes.

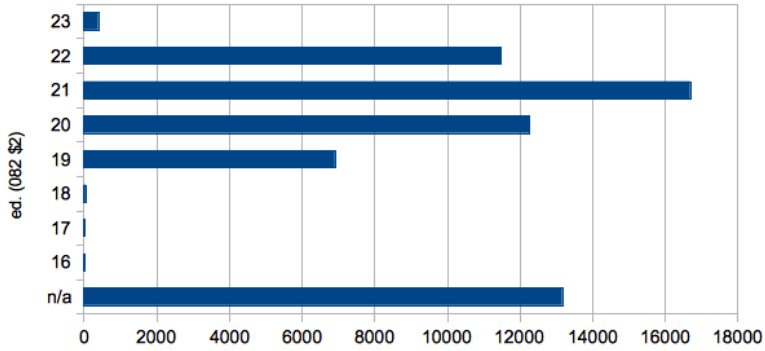


Figure 7: Distribution of Dewey classification editions.

10 Software used

The seven query programs were written in Perl language, making use of the Koha API and the following libraries:[19] LWP for HTTP connections, ZOOM for Z39.50 connections, DBI for connections to the MySQL database, XML::XPath for XML data processing, WWW::Scraper for HTML data processing, and MARC::Record for the processing of MARC records.

11 Conclusions

The present work helped us to understand the value and problems of retrieving information online, which can contribute to the improvement of bibliographical catalogs. Generally, copy cataloging is considered important in obtaining the entire record, but — through

[19] Each library is documented and available at http://search.cpan.org.

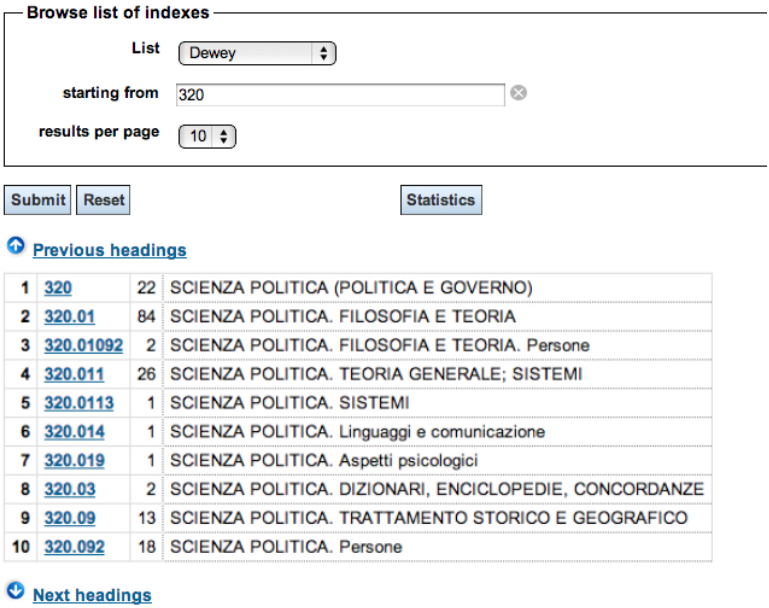


Figure 8: Browsing Dewey in Koha.

unique identifiers such as the ISBN etc. — it is possible to find partial or "atomic" information by means of which several objectives may be achieved:

- improving the catalog from a static perspective, as in the case presented
- dynamically enhancing the OPAC by the retrieval of data while viewing a record
- increasing navigability for better utilization of the OPAC
- helping to clean up the catalog
- performing quality checks
- providing support tools for cataloging
- increasing the number of unique identifiers in the catalog
- comparing databases.

12 Appendix

12.1 Query elements for the selection of records without a Dewey classification number

The function ExtractValue,[20] which is present in MySQL 5.1.5 or higher, allows querying of XML data, specifying the field to be examined and an XPath expression as parameters.[21]

[20] http://dev.mysql.com/doc/refman/5.1/en/xml-functions.html.
[21] http://en.wikipedia.org/wiki/XPath.

`biblionumber`	The system number of the bibliographical record
`ISBNlist`	`ExtractValue(marcxml,'//datafield[@tag="020"]/subfield [@code="a"]')` This is a list of occurrences of subfield $a in tag 020, which are separated by a space; normally, an occurrence is unique.
`isbn_present`	`ExtractValue(marcxml,'count(//datafield[@tag="020"] /subfield[@code="a"])>0')` At least one occurrence of 020$a
`dewey_absent`	`ExtractValue(marcxml,'count(//datafield [@tag="082"]/subfield[@code="a"])=0')` No occurrence of 082$a
`language_008`	substr(ExtractValue(marcxml,'//controlfield[\@tag=\"008 \"]'),36,3) = 'language_code'

Table 9: Principle elements of the query for the selection of bibliographical records to be reviewed.

12.2 Parameters for web searches

To identify the parameters that compose the search URL, including the ISBN, it is possible to proceed in one of the following ways:

- run the query and note the response URL. If it does not contain parameters, i.e. in the event of a form where `method="post"`, change its `method` to the value `get` through "Inspect Element" (contained in various browsers), by pressing right-click on the `form`, and then run the query;

- analyze the `http` request sent from the query by means of a plugin for traffic analysis, or a special feature in the browser.

12.3 Examples of some responses

An example of XML response obtained from Classify[22] is the following:

Listing 3: XML

```
<?xml version="1.0" encoding="UTF-8"?>
<classify xmlns="http://classify.oclc.org">
  <response code="2"/>
  <!-- Classify is a product of OCLC Online Computer Library
       Center: http://classify.oclc.org -->
  <work author="Beaucamp, Evode" editions="5" format="Book"
       holdings="69" itemtype="itemtype-book" title="Israel en
       prière : des Psaumes au Notre Père">014271167</work>
  <orderBy>hold desc</orderBy>
  <input type="isbn">2204022659</input>
  <start>0</start>
  <maxRecs>25</maxRecs>
  <editions>
    <edition author="Beaucamp, Evode" format="Book" holdings="
         40" itemtype="itemtype-book" language="fre"
         oclc="014271167" title="Israel en prière : des Psaumes
          au Notre Père">
      <classifications>
        <class edition="19" ind1="0" ind2="4" sf2="19"
             sfa="220.6" tag="082"/>
        <class ind1="0" ind2="4" sfa="BS680.P64" tag="050"/>
      </classifications>
    </edition>
    <edition author="Beaucamp, Evode" format="Book" holdings="
         21" itemtype="itemtype-book" language="fre" oclc="
         299394640" title="Israel en priere : des psaumes au
         Notre Pere">
```

[22] http://classify.oclc.org/classify2/Classify?summary=false&isbn=2204022659.

```
      <classifications>
        <class ind1="1" ind2="4" sfa="200" tag="082"/>
        <class ind1=" " ind2="4" sfa="BX2033B42 1985" tag="050"
            />
      </classifications>
    </edition>
    <edition author="Beaucamp, Evode" format="Book" holdings="
        5" itemtype="itemtype-book" language="fre" oclc="
        246374613" title="Israel en prière : des psaumes au
        Notre Père"/>
    <edition author="Beaucamp, Evode" format="Book" holdings="
        2" itemtype="itemtype-book" language="fre" oclc="
        442622354" title="Israel en prière : des Psaumes au
        Notre Père"/>
    <edition author="Beaucamp, Evode" format="Book" holdings="
        1" itemtype="itemtype-book" language="fre" oclc="
        718332441" title="Israel en prière : des Psaumes au
        Notre Père"/>
  </editions>
  <recommendations>
    [ ... ]
  </recommendations>
</classify>
```

An example of a Z39.50[23] (MARC21) response in readable format:

Listing 4: MARC21

```
00932cam 2200253 a 4500
001 500315
005 20050929180451.0
008 851021s1986 nyua 000 0 eng
035 $9 (DLC) 85073338
010 $a 85073338
```

[23]From Library of Congress, lx2.loc.gov:210/LCDB, find @attr 1=7 0874472466.

```
020 $a 0874472466 (pbk.) : $c $8.95
040 $a DLC $c DLC $d DLC
050 00 $a LB2353.57 $b .A16 1986
082 00 $a 371.2/6 $2 19
245 00 $a 10 SATs : $b the actual and [...] prepare for it.
250 $a 2nd ed.
260 $a New York : $b College Entrance Examination Board : $b
    ...
300 $a 304 p. : $b ill. ; $c 28 cm.
[ ... ]
```

An example in HTML code :[24]

Listing 5: HTML

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN"
"http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="de" lang=
    "de" dir="ltr">
<head>
<title>DNB, Katalog der Deutschen Nationalbibliothek</title>
[ ... ]
</head>

<body onload="doLoad()">
[ ... ]
  <tr>
    <td width="25%" >
      <strong>Link zu diesem Datensatz</strong>
    </td>
    <td>http://d-nb.info/977758214</td>
  </tr>
```

[24] https://portal.dnb.de/opac.htm?query=isbn%3D9783525563427&method=simpleSearch.

```
[ ... ]
  <tr>
    <td width="25%"><strong>DDC-Notation</strong></td>
    <td>231.6 [DDC22ger]</td>
  </tr>
  <tr>
[ ... ]
</body>
</html>
```

The browser version is shown in Figure 9.

Link zu diesem Datensatz	http://d-nb.info/977758214 ID
Titel	Fides caritate formata : das Verhältnis v
Person(en)	Rose, Miriam
Verleger	Göttingen : Vandenhoeck & Ruprecht
Erscheinungsjahr	2007
Umfang/Format	303 S. ; 24 cm
Gesamttitel	Forschungen zur systematischen und ö
Hochschulschrift	Zugl.: München, Univ., Diss., 2004/200
ISBN/Einband/Preis	978-3-525-56342-7 Pp. : EUR 59.90 3-525-56342-6 Pp. : EUR 59.90
EAN	9783525563427
Sprache(n)	Deutsch (ger)
Schlagwörter	Thomas <de Aquino>: Summa theologi
DDC-Notation	231.6 [DDC22ger]
Sachgruppe(n)	230 Theologie, Christentum
Links	Inhaltsverzeichnis

Figure 9: Result of a search by ISBN in the Deutsche Nationalbibliothek catalog.

STEFANO BARGIONI, Biblioteca della Pontificia Università della Santa Croce, bargioni@pusc.it
MICHELE CAPUTO, Biblioteca della Pontificia Università della Santa Croce, caputo@pusc.it
ALBERTO GAMBARDELLA, Biblioteca della Pontificia Università della Santa Croce, gambardella@pusc.it
LUIGI GENTILE, Biblioteca della Pontificia Università della Santa Croce, gentile@pusc.it

---

Bargioni, S., M. Caputo, A. Gambardella, et al. "Obtaining the Dewey Decimal Classification Number from other databases: a catalog clean-up project". *JLIS.it*. Vol. 4, n. 2 (Luglio/July 2013): Art. #8766, p. 1–200. DOI: 10.4403/jlis.it-8766. Web.

ABSTRACT: The increasing availability of online catalogs and bibliographical databases allows not only for copy cataloging, but also for the retrieval of atomic information useful within the catalog. To this end, Dewey decimal numbers were imported from national and international sources by means of the unique identifier ISBN. Technical specifications have been developed to locate the records to be enhanced, to query external databases, to extract the Dewey decimal classification numbers and add them to the catalog. The exceptionally large amount of Dewey numbers added to the catalog has improved the semantic usability of the OPAC. The procedure established has also facilitated the collection of information on the use of the Dewey Decimal System in the various databases used and allowed to make certain comparisons between them. The tools employed can be used analogously for data-retrieval operations in the catalog, as an aid in the cataloging process, or to improve the OPAC in either a static or dynamic manner. Taking into account its virtually exclusive practical purpose, this work is characterized by practical rather than theoretical choices. However, the experience acquired opens up areas even in the field of academic research.

KEYWORDS: Dewey Decimal Classification; DDC; Information retrieval; ISBN; Koha; OPAC; REST; Z39.50.

Submitted: 2013-02-04
Accepted: 2013-03-28
Published: 2013-07-01

JLIS.it

Libraries and law firms in Italy

Ewelina Melnarowicz, Federica Vignati

1 Introduction

At the core of this study on law firm libraries in Italy is a professional experience of the authors in a law firm to reorganise and manage the library service. At this juncture, looking for examples and solutions adopted in a similar context, there is a dearth of literature in Italian referred to libraries, law firms and corporate libraries generally. Regarding legal public libraries, see the contribution of Rosa Maiello at the IFLA 2009 Conference ("Law Libraries in Italy"). Searches carried out in major LIS databases have not given any relevant publications on this topic in Italy. Instead some interesting starting points can be found in online contributions of lawyers and managers of knowledge management services. For example, see the contribution of Gaia Bassani Antivari presented at the 11° Meeting ACEF ("Gestire La Conoscenza Per Migliorare L'efficienza e Garantire Continuità Al Cliente"). Therefore, in the first instance we have considered the importance of verifying the spread of libraries in Italian law firms by analyzing resources, tools and services through a questionnaire

JLIS.it. Vol. 4, n. 2 (Luglio/July 2013).
DOI: 10.4403/jlis.it-8800

submitted to the 100 major Italian law firms.[1] Particular attention was given to the human resources in charge of the library and to their professional profile. In this study, we have often used the term *knowledge management* referring in a rather inclusive way to all the activities connected to the retrieval, usage and communication of information. More specifically, we distinguished between the term library, used here as the management of all the knowledge generated externally to the firm, and the term knowledge management, as the management of the knowledge generated in the firm, in order to avoid ambiguity and for a better comprehension of the activities involved. You may see Profili and Capitani (*Il Knowledge Management: Approcci Teorici e Strumenti Gestionali*; *Il knowledge management*) to gain familiarity with knowledge management. All the answers were firstly collected and then mapped against the information provided by the law firms on their web sites. The publication of the results of this study is intended to be a first approach to provide insight into the spread and the organization of the Italian law libraries and to initiate a discussion on the professionals involved in their management.

2 The study

2.1 Sampling

When defining the sample, it has been chosen to take in consideration the major Italian law firms by revenue and the number of

[1] The journal Top Legal publishes annually a list of the first 100 Italian law firms by revenue, calculated on the previous year. In addition to revenue the article reports the number of lawyers and other economic data on the state of the legal market. The reference for this study is the list for the year 2012, based on the data of the 2011 (Di Carlo). The list comprises of 102 firms, two of which closed down their practice. The sample therefore consist of 100 law firms.

lawyers employed, since this could guarantee greater possibilities to find structures dedicated to knowledge management. As pointed out previously, the population consists of the 100 Italian law firms classified annually by the journal Top Legal. The choice of this source, being a household name in the legal community, could be a valid standard for the respondents to the questionnaire.

2.2 Methods

In this investigation, we decided to use an online questionnaire to reach a numerous and geographically dispersed population, even if limited to Italy. For more information on this research methodology see (Pitrone; Agnoli). The preliminary phase of the survey focused on collecting e-mail addresses in order to send the invitation for participating in the questionnaire, in the following order: in the first instance if available, email address of the library manager or of the head of knowledge management; then, if unavailable, an email address of the firm; at last, if when there was no answer, email address of one of the associate lawyers. Apart from retrieving email addresses, we undertook an analysis of the law firms web sites with the purpose to identify a library, knowledge management services and publishing activities.[2] The analysis of the web sites has been the fundamental step to verify and validate the results of the questionnaire. This recipients of the questionnaire received an email containing a link to the questionnaire and a short presentation of the research and its purposes. For the management of the questionnaire

[2]Starting from the list cited previously, between August 2012 and January 2013 we analyzed 96 law firms' web sites (from the initial sample of 102, 6 law firms were excluded, because have stopped activity or without a web site, or with a web site under maintenance).

we utilized a dedicated software[3] in order to be able to manage email invitations and reminders in an automatic way and to control the data during and after the data collection. Moreover the choice of this platform allowed to trace which email addresses effectively compiled the questionnaire, therefore allowing for further verification of the data. The questionnaire was opened for one month (7 November - 7 December 2012). Unless there was an answer, or when a questionnaire was partially completed, three reminders were sent also to different email addresses. The questionnaire was opened 57 times, and compiled 35.

2.3 Limits

The problems encountered regard primarily a low response rate (35 out of 100) and the difficulty to validate the collected data: typical problems of surveys carried out through electronic questionnaires. Regarding questionnaires see Pickard and Caselli (*La Ricerca in Biblioteca: Come Migliorare i Servizi Attraverso Gli Studi Sull'utenza*; *Indagare col questionario. Introduzione alla ricerca sociale di tipo standard*) on self-completion questionnaires without a face to face contact with the researcher. Indeed when there is no direct contact between the researcher and the interview ee it is in fact much more difficult to avoid misunderstanding of the questions, ascertain trustfulness of the responses, and limit the abandonment of the compilation (Denscombe). However the functionalities of the tool utilized for distributing the questionnaires allowed to compare the answers of the law firms with the data provided in the Top 100, hence providing a double-check against the answers regarding the revenues, the number of the lawyers and the location of the headquarters. Once

[3]We chose Sondaggiofacile, http://www.sondaggiofacile.com/, free web site for creating surveys. It enabled a good personalization of the design and of the question structure.

the quality of the responses has been ascertained, although the number of the law firms surveyed was negligible, the data showed that their composition by revenue, location and number of the lawyers replicate quite faithfully the composition of the original sample (see questions 1 on the following page, 2 on page 207 and 3 on page 207). This partial overlap therefore allows to obtain from the collected data, if not an exhaustive description of the Italian situation, at least some common characteristics.

2.4 Questions

The questionnaire consists of 14 closed and open-ended questions distributed in 4 macro areas: *profile questions* (revenue, number of lawyers, location of the headquarters, name of the firm), *general information* (publications, presence of a library at the headquarters and in the branch offices), *resources, tools and services* (type of resources, tools and services offered by the law firm), *staff* (professional profile and qualifications). Lastly, the final open question offered a possibility to provide additional information on the library and the library personnel and to express personal opinions.

3 Results and analysis of the data

As anticipated, there are 35 law compiled the questionnaire, of which 8 partially. Where a comparison could be useful, the answers given to the questionnaire were cross-tabulated with the relative data on the population of the study provided by the journal Top Legal or the data collected in the web site analysis. What follows are the questions of the questionnaire and the collected data.

3.1 Profile questions (1-3)

These questions, in the opening part of the questionnaire, addressed the need to delineate the composition of the sample utilizing objective and easily verifiable data. The answers of every law firm were compared with the information provided in the Top Legal list, and, if necessary, were corrected and integrated. Therefore, the first three questions report the real data of the 35 participating law firms and are compared with the first 100 Italian law firms.[4]

Table 1: The headquarters of the law firm is (Question 1):

	sample 35		**sample 100**	
in Italy	21	60%	69	69%
abroad	14	40%	31	31%
	35	100%	100	100%

The first question (Table 1) gathered responses from all of the participants. The majority of the participating law firms have their head office in Italy (60%), confirming therefore a substantial overlap with the sample of reference (69%).

[4]The two groups of reference will be referred to herein as *sample 100* (Top Legal list) and *sample 35* (participants in the survey).

Table 2: Number of the active lawyers in the Italian head office/s in 2011 were (Question 2):

	sample 35		**sample 100**	
Less than 20	7	20%	25	25%
Between 20 and 50	10	29%	41	41%
Between 50 and 100	12	34%	20	20%
Between 100 and 200	4	11%	9	9%
More than 200	2	6%	5	5%
	35	100%	100	100%

Among the respondents to this question the prevailing number of the lawyers is in the mid range with more than 60% of the law firms comprised between the 20 and 100 lawyers. In this case the composition of the sample of the participants is not totally overlapping with the reference sample, mainly with regard to the ranges 20-50 and 50-100.

Table 3: The revenue of the law firm in 2011 (in millions of €) (Question 3):

	sample 35		**sample 100**	
Less than 10	15	43%	45	45%
Less than 20	9	25.5%	29	29%
Less than 50	9	25.5%	19	19%
More than 50	2	6%	7	7%
	35	100%	100	100%

The two samples are also similar in the revenue breakdown. Also in this case there are differences in the mid range, but they are of negligible entity.

Name of the Law Firm (optional) (Question 4):
This question, indicated as optional, that sought to determine the name of the firm, received responses in 54% of the cases. As far as the answer to this question was not necessary in order to trace the identity of the participating law firms, it is however interesting to notice that nearly half of the sample preferred not to associate the name of the law firm to the provided answers.

3.2 General information (5-7)

Table 4: The law firm has publications in paper or digital format? (Question 5):

Yes	25	78%
No	5	16%
N/A	2	6%
	32	100%

The majority of the law firms have been involved in publishing. This emphasizes that the law firms are also places where knowledge is not only used but also communicated. The collected data are supported by the web site analysis, from which it turned out that 59% of the sample, equal to 57 firms out of 96, communicates the presence of several publishing activities, from monographs to thematic portals with legal information. Many of the publications from the law firms are moreover freely available and accessible on-line in full text.

Table 5: The head office of the firm has a library? (Question 6)

Yes	33	100%
No	0	0%
N/A	0	0%
	33	100%

For this question (Table 5 on the facing page) as for the next one (Table 6) the meaning of the term library has not been specified, however implying physical aspects of the library, a place for studying and for the physical collection of the library, leaving to the next questions the task to identify resources and services on offer. Interestingly, the results (100% of respondents have a library) emphasize that the library is a necessary service to the practice of the surveyed law firms. However, due to the self-selection bias, it is possible that in the survey participated only law firms that have a library service.

Table 6: Information about the services provided by the library (Source of the data: web sites):

Library, reference, research centre	22	23%
Knowledge management	7	7%
Pictures of the library	20	21%
N/A	47	49%
	96	100%

From the web site analysis, as summarised in the above table, it is possible to gain additional insight. In particular, nearly half of the law firms surveyed explicitly referred to have a library, a research centre or knowledge management service and/or pictures of the library.

Table 7: Branch offices of the firm have a library? (Question 7):

Yes	26	79%
No	4	12%
The law firm does not have branch offices	3	9%
N/A	0	0%
	33	100%

The answers indicate that for the majority of the respondents with more than one office the information services are widespread. At this moment it has not been addressed in depth the accessibility of the local collection or rather services and resources, both from the branch offices and the headquarters (such as for example reference service, databases or e-journal).

3.3 Resources, tools and services (8-10)

Table 8: What kind of resources are available to the lawyers? (Question 8):

Databases	31	97%
Books	30	94%
Print journals	29	91%
E- journals	29	91%
Daily newspapers	27	84%
EBooks	6	19%
N/A	1	3%
Other (please, specify)	2	6%
-centro studi		
-prestiti interbibliotecari tramite catalogo ACNP		
Sample size	**32**	

According to the answers there are several resources available to the lawyers. In addition to the printed resources, although widespread (books 94%, journals 91%, newspapers 84%), there are databases (the most common, they are indeed used by 97% of the respondents) and e-journals. To find more about the variety of resources used by the corporate lawyers see Breslin ("Research and Resources for Corporate Lawyers").
Definitely the less common are eBooks, still they are used by 6 out

of 32 law firms. In the future it will be interesting to see if the possibilities brought in by the mobile devices and the adaptation of the legal publishing market would bring in an increase of this type of publications in legal libraries.[5]
The notice about an interlibrary loan through ACNP catalogue[6] offered the opportunity to reflect on the one hand on the provision of the services and on the other hand on the relationship between *corporate libraries* and other libraries. In fact, although the document delivery services are one of the most common (see the next question), this is however the only case in which the tool used (ACNP - Italian Catalogue of journals) to provide the service is mentioned. It is not, unfortunately, specified if the library uses exclusively the document delivery service through email or if it joined ACNP and/or uses automated exchange services like NILDE. In addition, the use of ACNP denotes the necessity regarding legal libraries to draw from a wealth of shared tools, used at the national level. On this basis, and considering a substantial affinity between legal libraries and academic libraries with regard to resources and services (see questions 9 on the following page and 10 on page 213) it is possible to foresee conditions for a potential collaboration between public and private libraries.

[5]Some of the publishers particularly active in the legal market, such as Wolters Kluwer Italy and Giuffrè, have currently on offer digital libraries, called respective "La mia biblioteca" and "Biblioteca volumi" that allow to access monographs also on mobile devices. Other publishers in the legal market, such as for example il gruppo Sole 24 ore offer new publications also in digital version with DRM.

[6]Given the type of catalogue, perhaps the intention was to refer to a document delivery service. In any case it is more correct to relate this answer to the next question, regarding services.

Table 9: What kind of resources are available to the lawyers? (Question 9):

Lending services of books and other resources	27	84%
Document delivery	26	81%
Courses on using databases and/or other resources	19	59%
Other (please, specify)	7	22%
reference digitale		
corsi di marketing e di aggiornamento professionale		
servizio di knowledge management		
Centro Studi		
ricerche giuridiche fatte da uno stagista laureando in giurisprudenza		
partecipazioni seminari e/o convegni		
Seminari per formazione professionale continua.		
Sample size	**32**	

Among the available services, book lending and reference service are the most common (available respectively in 84 and 81% of the law firms). The most interesting aspect as evidenced in this question turns out to be the involvement of the library in educational activities, such as, in the first instance, training on using databases and other information tools (59% of the law firms). Additional comments emphasized that the law firm library is truly at the core of learning for lawyers, as the library is committed to the provision of continuing professional education courses[7] and capacity building (marketing), both the administration of the attendance of lawyers in seminars or conferences organized by external bodies.
Additional information is also provided on the reference service:

[7]For the lawyers, like for other professions, starting from 2007 has been introduced a required continuing professional education. Lawyers must therefore participate in continuing professional education courses in order obtain required credits. The law firms can provide these courses themselves upon registration from the Italian Lawyers' Council. For further information see the website of the Italian Lawyers' Council in the section dedicated to continuing professional education: http://www.consiglionazionaleforense.it/site/home/formazione/formazione-continua.html.

one law firm emphasized that the bibliographical research is carried out by a graduate majoring in law, therefore paying special attention to the specific knowledge of the field. Interestingly, the reported here digital reference service is usually found in academic libraries and public libraries. Also in this case it could be interesting to learn more about it in particular with regard to the tools used for the distribution of the service (email? specific software?).

Table 10: What kind of resources are available to the lawyers? (Question 10):

Management software (for example Easylex)	27	87%
Library catalogue	22	71%
Knowledge Management System or Customer Relationship Management	13	42%
Internal wiki	8	26%
N/A	0	0%
Other (please, specify)	2	6%
-centro studi		
-intranet di studio con data base interni ed esterni, informazioni e procedure		
Sample size	**31**	

In this question we asked about some of the most common management software in law firms to understand the context of the library. The use of this software (modules such as customer records, accounting, etc.) is widespread and, interestingly, some of them already include functionalities designed to manage the library. It should be however stressed that there is a numerical mismatch between the number of libraries and library catalogues, it is therefore reasonable to assume that the library collection is not always managed in a

structured way. On the other hand Wikis, Knowledge management systems and Customer Relationship Management Systems represent advanced tools for the sharing of internal knowledge and even if at the moment they are less common, the data pointed out that a significant number of them are used among the respondents. The additional comments also pointed out the use of Intranet in law practice: in particular the Intranet represents a privileged point of access to the internal and external sources of information. Regarding the use of Wiki and Intranet in law firms see for example (Sarkanen260–265; Rudman250–253). Finally, one law firm responded in the same way to the questions regarding resources, tools and services: *Centro Studi*. Perhaps the answers provided to these questions do not reflect the internal structure of the firm, or maybe it was meant to indicate that there is an independent structure combining such responsibilities as the research, management and communication of information.

3.4 Library staff

Table 11: Who is in charge of the library and/or information services and knowledge management? (Question 11):

One dedicated person	11	35%
A person who also carries out other duties	11	35%
A dedicated team	7	23%
N/A	0	0%
Other (please, specify)	2	6%
-alcuni collaboratori dello Studio		
-i professionisti stessi		
	31	100%

In the law firms surveyed the library is managed equally by a dedicated person or a person that carries out also other duties (both 35%).

Also significant is the percentage of firms with a team in support of the internal information services. This aspect deserves further investigation, in particular with regard to the composition of the management teams (see question 14). The additional comments reported the management role carried out by lawyers and are reflected in the answers to the other questions, in particular number 13 on the following page, which acknowledged that the management of the library is mostly done by graduates in law. Another firm indicated, in a vague way, that some lawyers of the firm are managers of the library services.

Table 12: Which is the highest qualification of the person in charge of the law firm library and/or of bibliographical services and knowledge management? (Question 12):

High School Diploma	5	17%
Post High School Education	0	0%
Bachelor degree	21	70%
Master's degree	1	3%
PhD	0	0%
N/A	2	7%
Other (please, specify)	1	3%
-avvocato		
	30	100%

In most cases (70%) the person in charge of the knowledge management services holds a Bachelor degree, and in one case a Master's degree. Second most common are respondents with a High School Diploma equal to 17%. The only additional comment pointed out that a lawyer is in charge of the library (see questions 13 on the next page and 14).

Table 13: If the person in charge of the law firm library and/or the bibliographical services holds at least a bachelor degree what is the specific field of study? (Question 13):

Librarianship or similar	3	11%
Humanities	5	19%
Law	15	56%
Sciences	1	4%
N/A	3	11%
Other (please, specify)	0	0%
	27	100%

This question sought to gain insight into what extent the library staff has a background in library and information science, therefore, how important is this qualification in the selection of the candidates. The answers, however, show that the majority of the library managers in law firms earned their title in law, following are the humanities with 19% of the respondents and then librarianship, with only 3 out of 27 respondents. Probably still much weighs the subject while specific and wide ranging competences are less relevant, although applicable to fully profit in even completely different professional domains.

Other information on the library (optional) (Question 14):
La biblioteca è gestita da una bibliotecaria dedicata con il supporto di un "comitato Biblioteca" formato da professionisti di Studio, mentre per quanto concerne la "gestione della conoscenza" un comitato formato da professionisti di Studio propone aggiornamenti al sistema documentale.

Although only one firm provided additional information, the data turned out to be very interesting for our study. In particular it offers an overview of some aspects of the knowledge management in the firm, clearly separating out the management of the library from the

Knowledge management. In both areas, however, the organization has a team of lawyers, managed by a designated staff as far as the library is concerned. It is possible to hypothesize that lawyers coming from various practice areas are part of the committees in order to represent all the areas of practice of the firm. The skills of the lawyers can be spent to take decisions about the acquisition, management and the review of book collections and documents and to support the bibliographical research.

4 Conclusions

This study, although carried out on a limited sample, has showed that the law firms involved in the survey present wide-ranging characteristics regardless of revenue and location of the firm. The firms showed pronounced patterns typical of most of the respondents, including the publishing activities, the spread of the library, which is regarded as one of the fundamental support services and a wide range of libraries in branch offices, and a wide use of resources, including databases, books and journals. The services and tools show a diversified picture, in which there are *standard* services such as lending and reference (document delivery) with the support of a management tool (catalogue) together with courses on using databases and continuing professional education. This forms a close relationship between resources, tools and services. Moreover, there is a wide variety of legal tools in use, also taking into consideration the distribution of advanced software like wiki, KMS, CRM. Regarding professional qualifications the majority of the participants seems to entrust the knowledge management to staff graduated in law. In this regard, it is necessary to reflect on the profile and recognition that the library professionals have elsewhere than in the public sector. Since the specific *information science* skills can be spent elsewhere,

it is important to increase and to improve training courses to prepare the professionals to cater for the demand of the private sector. On the sidelines of this study it is possible to emphasize that the available resources and services are similar to those offered in academic libraries, in particular the reference, interlibrary loan and the use of databases. This enables to imagine a collaboration between public and private libraries (collaboration in some cases already started), supported by the interest of the law firms to distribute a part of their internal knowledge (as pointed out by the publishing activities).

5 Future studies

This study can be a valid point of reference for further research on law firm libraries. The results showed that more than one of the major law firms has a library, while the data from the web sites suggest an even greater spread of libraries and knowledge management services. The first point of interest could be to investigate how these support services are conveyed to the public, and in particular to the customers.[8] Moreover, it could be useful to investigate more in depth (through interviews or focus group with stakeholders) some aspects of the organization of the law firms with such characteristic, as for example the management by a team or the relationships between library services and knowledge management.[9] Finally, it

[8]The analysis of the web sites has not been thorough, but enough to evidence some recurring characteristics. In particular, the use of pictures or references to the library as a prestigious symbol rather than part of practice. In fact what is lacking is the description of library services, or its long term effects on the performance of the lawyers.

[9]The web site analysis pointed out that knowledge management and information services are only referred to in a general manner, and they are variously located on the web sites: in some cases mentioned along with the publishing activities, in others as one of the services for lawyers (under recruiting), in other cases in the "about us"

could be useful to compare professional profiles in Italy and abroad, for example in the English speaking countries.[10] This would allow to know solutions adopted in contexts with similar features and issues.

This study is only a+ first step in order to know the state of the art of the Italian law libraries. Similar studies with a purpose to investigate in depth and by involving a wider and a more representative sample, and using various methods for data collection could give interesting and comparable results with the data of this study.

section of the firm. Therefore it would be necessary to actually check in what way these services are provided.

[10]The web site analysis pointed out only a small number of specific professionals in the staff directory. The international law firms websites contain different profiles, such as "research librarian" and "knowledge management lawyer" pertaining respectively to the technical and legal staff.

References

Agnoli, Maria Stella. *Il Disegno Della Ricerca Sociale*. Roma: Carocci, 2004. (Cit. on p. 203). Print.

Bassani Antivari, Gaia. "Gestire La Conoscenza Per Migliorare L'efficienza e Garantire Continuità Al Cliente". *11° Meeting ACEF (R)INNOVARE LO STUDIO dell'Avvocato*. Milano, Italia, 2011. (Cit. on p. 201). Print.

Breslin, Jas. "Research and Resources for Corporate Lawyers". *Legal Information Management* 11.1. (2011): 65–68. (Cit. on p. 210). Print.

Capitani, Paola. *Il knowledge management*. Milano: FrancoAngeli, 2006. (Cit. on p. 202). Print.

Caselli, Marco. *Indagare col questionario. Introduzione alla ricerca sociale di tipo standard*. Milano: Vita e Pensiero, 2007. (Cit. on p. 204). Print.

Denscombe, Martyn. *The Good Research Guide: For Small-scale Social Research Projects*. Maidenhead: Open University Press, 2003. (Cit. on p. 204). Print.

Di Carlo, Amalia. "Stesso Podio sull'Olimpo". *Top Legal* 6. (2012): 40–63. (Cit. on p. 202). Print.

Maiello, Rosa. "Law Libraries in Italy". *World Library and Information Congress: 75th IFLA General Conference and Council 23-27 August 2009*. Milano, Italia, 2009. (Cit. on p. 201). Print.

Pickard, Alison J. *La Ricerca in Biblioteca: Come Migliorare i Servizi Attraverso Gli Studi Sull'utenza*. Milano: Bibliografica, 2010. (Cit. on p. 204). Print.

Pitrone, Maria Concetta. "L'intervista Con Questionario". *Ricerca Sociale: Dal Progetto Dell'indagine Alla Costruzione Degli Indici*. Carocci, 2007. (Cit. on p. 203). Print.

Profili, Silvia. *Il Knowledge Management: Approcci Teorici e Strumenti Gestionali*. Milano: FrancoAngeli, 2004. (Cit. on p. 202). Print.

Rudman, Sarah. "Knowledge Management and the Intranet at Field Fisher Waterhouse". *Legal Information Management* 9.4. (2009): 250–253. (Cit. on p. 214). Print.

Sarkanen, Anneli. "Using Wikis as Cost Saving Tools at Field Fisher Waterhouse". *Legal Information Management* 10.4. (2010): 260–265. (Cit. on p. 214). Print.

EWELINA MELNAROWICZ, La Scala Studio Legale.
emelnarowicz@gmail.com

FEDERICA VIGNATI, Università degli Studi di Milano.
federica.vignati@gmail.com

Melnarowicz, E., F. Vignati. "Libraries and law firms in Italy". *JLIS.it*. Vol. 4, n. 2 (Luglio/July 2013): Art: #8800. DOI: 10.4403/jlis.it-8800. Web.

ABSTRACT: This essay presents the results of the first study on law firm libraries in Italy. To date there are only few publications on this topic, therefore it has been considered interesting to determine the distribution of law libraries and professionals involved. The study, which was based on a sample population of 100 largest law firms in Italy by revenue, employed an online questionnaire and the data were collected through web site analysis aiming to identify a library, a knowledge management service and editorial activities. Besides variables regarding revenue, location and number of lawyers were extracted from the data of the sample population in order to validate the results. The study determined, even though the response rate was quite low (35% of the population) thus the sample is partially representative, that most of the law firm libraries show characteristics in common including resources, services and tools which may represent a large-scale standard.

KEYWORDS: Italy; Knowledge management; Law firms; Organizational culture; Quantitative research; Special libraries

Submitted: 2013-02-28
Accepted: 2013-04-30
Published: 2013-07-01

JLIS.it

La gestione dei diritti nei progetti di digitalizzazione: il pubblico dominio e le opere orfane.

Maria Cassella

Da alcuni anni il tema dei diritti è diventato cruciale per la gestione quotidiana delle attività della biblioteca digitale: dai repository istituzionali, alla valutazione della ricerca, dai contratti di licenza di uso fino ai progetti di digitalizzazione il tema dei diritti tocca in modo trasversale e profondo la biblioteca digitale, i suoi contenuti e i suoi *output*. Il mezzo digitale ha messo in crisi il tradizionale sistema di condivisione e pubblicazione di contenuti scientifici basato sull'intermediazione degli editori: consente, infatti, una più efficace gestione della conoscenza in rete, un'ampia e immediata disseminazione dei contenuti e dei servizi, rinvigorendo il mito della biblioteca di Alessandria. D'altro canto, proprio grazie al digitale, si assiste alla nascita di nuove forme di controllo dell'informazione. I principali ostacoli alla condivisione pubblica della conoscenza in rete sono:

- il sistema privatistico delle licenze di uso e dei Digital Rights Management (DRM)[1] imposto dagli editori che controllano il

[1]Si tratta delle informazioni crittografate e dei metadati che vengono aggiunti ad un file digitale per limitarne il ri-utilizzo.

JLIS.it. Vol. 4, n. 2 (Luglio/July 2013).
DOI: 10.4403/jlis.it-8797

mercato editoriale scientifico secondo logiche oligopolistiche e fortemente centralizzate (Caso). In virtù di queste logiche privatistiche e in nome del profitto gli editori scientifici chiudono l'accesso alla conoscenza moderna ed al suo riutilizzo. Internet, "bene comune libertario", infrastruttura comune e democratica ha paradossalmente favorito il potenziamento sul mercato degli editori scientifici, accrescendo a dismisura la loro visibilità nazionale ed internazionale, ampliando le possibilità di sfruttamento economico, favorendo la crescita dei prezzi e rafforzandone il *brand*;

- il sistema legislativo e i limiti imposti alla condivisione dei contenuti dalla tutela della proprietà intellettuale: il diritto di autore o, per i paesi anglosassoni, il sistema di copyright, la tutela dei marchi e dei brevetti.

> «Nella dimensione mondiale della globalizzazione assistiamo ad una creazione incessante di nuovi beni, la conoscenza prima di tutto, rispetto ai quali la scarsità non è l'effetto di dati naturali, ma di politiche deliberate, di usi impropri del brevetto e del copyright, che stanno determinando un movimento di chiusura simile a quello che in Inghilterra portò alla recinzione delle terre comuni, prima liberamente accessibili. Dobbiamo concludere che la tecnologia apre le porte e il capitale le chiude? Certo è che intorno al destino di nuovi e vecchi beni comuni si gioca una partita decisiva per la libertà e l'uguaglianza»(Rodotà).

Il modello dell'accesso aperto si confronta ormai da dieci anni con altalenanti fortune con questo complesso scenario. Nella comunicazione scientifica l'accesso aperto è un movimento di idee che promuove il libero accesso e riutilizzo in rete della letteratura scientifica

tutelata dal copyright. La gestione dei diritti di autore ha, in realtà, un impatto rilevante anche sulla fruizione ed il libero riutilizzo in rete dell'intero patrimonio culturale, posseduto e conservato dalle biblioteche e composto da opere tutelate dal copyright (in commercio e fuori commercio),[2] opere in pubblico dominio e opere orfane. In questo lavoro affronteremo le problematiche legali connesse con la digitalizzazione del patrimonio culturale, con particolare riferimento alle opere in pubblico dominio e alle opere orfane, mettendo in evidenza i principali nodi e le criticità che le biblioteche si trovano ad affrontare nel proporre e realizzare progetti di digitalizzazione.

1 L'Unione Europea e la digitalizzazione del patrimonio culturale

La digitalizzazione del patrimonio culturale è una delle attività strategiche per le biblioteche del Ventunesimo secolo. È un'attività trasversale a biblioteche, archivi e musei. Coinvolge tutte le tipologie di biblioteche: accademiche, di ricerca, specialistiche, universitarie, pubbliche. Nonostante gli alti costi di avviamento e gestione di un progetto di digitalizzazione e la necessità di ingenti investimenti economici,[3] l'impatto economico e sociale della digitalizzazione del patrimonio culturale, per il settore pubblico così come per quello privato, appare ormai un dato acquisito. Non è un caso che nel

[2] In termini di tutela giuridica non vi è alcuna differenza tra opere in commercio e fuori commercio a meno che queste ultime non rientrino per scadenza dei termini di tutela nel pubblico dominio. Attanasio («La gestione dei diritti di autore nelle biblioteche digitali: il caso ARROW») sottolinea che tra le opere fuori commercio è utile fare un'ulteriore suddivisione tra opere con titolari attivi e quelle con titolari non attivi

[3] Nel rapporto finale del Comité des Sages viene stimata una cifra pari a 100 bilioni di euro per digitalizzare il patrimonio complessivo delle biblioteche, archivi e musei in Europa.

corso degli ultimi dieci anni il pubblico settore, in Europa e negli Stati Uniti, e un certo numero di entranti privati (tra gli altri: Google, Microsoft, ProQuest),[4] si siano inseriti strategicamente, anche se con tempi e modalità diverse, nelle politiche di digitalizzazione del patrimonio culturale dando vita a progetti di digitalizzazione di massa o, comunque, di ampia portata: Europeana,[5] HathiTrust Digital Library, Digital Public Library of America, Google Books, Open Content Alliance, per citare solo i più noti. La Commissione Europea ha sottolineato da subito l'importanza delle attività di digitalizzazione del patrimonio culturale facendone uno degli obiettivi chiave della propria azione e dell'Agenda digitale europea. Nella sua comunicazione *i2010: digital libraries* del 30 settembre 2005 (Comunicazione della Commissione al Parlamento europeo, al Consiglio, al Comitato economico e sociale europeo e al Comitato delle regioni - i2010: le biblioteche digitali) la Commissione tracciava la sua strategia per la digitalizzazione, accessibilità in rete e conservazione della memoria collettiva in Europa, evidenziando le sfide economiche, organizzative, sociali e legali dei progetti di digitalizzazione in Europa. Dedicata alla digitalizzazione del patrimonio culturale è la successiva Raccomandazione 2006/585/CE del 24 Agosto 2006 sulla "Digitalizzazione e l'accessibilità online del materiale culturale e sua conservazione". Nel documento la Commissione Europea sottolineava i vantaggi della digitalizzazione del patrimonio culturale europeo – monografie, riviste, quotidiani, materiale museale, archivistico, audiovisivo – per la popolazione degli Stati membri, legata a lingue e tradizioni diverse. La digitalizzazione infatti:

[4]Particolarmente apprezzate nell'ambito della digitalizzazione sono le partnership pubblico-privato. I privati sono, infatti, finanziatori di numerosi progetti di digitalizzazione di ampia portata.

[5]Europeana è la biblioteca digitale europea inaugurata il 20 novembre 2008.

- «makes that material broadly accessible to a large proportion of the population in the European Union (EU);
- makes the EU's multilingual and diverse heritage clearly visible;
- preserves that collective memory long-term for the benefit of future generations».

La Commissione sollecitava la creazione di centri di competenze sulla digitalizzazione in Europa, la cooperazione tra gli Stati membri e la definizione di strategie nazionali e di piani di azione per l'accesso e la conservazione a lungo termine del materiale in formato digitale. Ad ottobre 2011, alcuni mesi dopo la pubblicazione del rapporto finale sul nuovo rinascimento digitale del Comité des Sages,[6] riprendendone le conclusioni, la Commissione Europea pubblicava una nuova raccomandazione sulla "Digitalizzazione e l'accessibilità in rete dei materiali culturali e sulla conservazione digitale" (2011/7579/CE del 27 Ottobre 2011).

In questa seconda raccomandazione la Commissione invitava gli Stati membri a:

1. «**sviluppare** ulteriormente la pianificazione e il monitoraggio della digitalizzazione di libri, riviste scientifiche, giornali, fotografie, oggetti museali, documenti d'archivio, materiali sonori e audiovisivi, monumenti e siti archeologici (nel prosieguo denominati "materiali culturali"):

[6]Il Comité des Sages (Comitato di saggi) viene costituito nell'aprile 2010 dalla Vicepresidente della Commissione europea per la Digital Agenda Neelie Kroes allo scopo di esplorare i limiti e le opportunità della digitalizzazione in Europa

- fissando chiari obiettivi quantitativi per la digitalizzazione dei materiali culturali [...];
- creando una visione d'insieme dei materiali culturali digitalizzati e contribuendo alle attività di collaborazione mirate a creare una tale visione a livello europeo con dati raffrontabili;

2. **incoraggiare** i partenariati fra le istituzioni culturali e il settore privato al fine di creare nuovi modi per finanziare la digitalizzazione dei materiali culturali e promuovere usi innovativi di questi ultimi, garantendo nel contempo che i partenariati pubblico-privato nel settore della digitalizzazione siano equi ed equilibrati nonché coerenti [...];

3. **avvalersi** dei fondi strutturali dell'UE, ove possibile, per cofinanziare le attività di digitalizzazione nell'ambito delle politiche regionali d'innovazione per la specializzazione intelligente;

4. **tenere** conto di metodi per ottimizzare l'uso della capacità di digitalizzazione e conseguire economie di scala [...]»

Nella raccomandazione del 2011 la Commissione sottolineava, inoltre, la necessità di rendere accessibili i materiali digitalizzati attraverso il portale Europeana, di garantire l'uso di norme comuni per la digitalizzazione per favorire l'interoperabilità, di rendere disponibili a titolo gratuito i metadati esistenti, con l'obiettivo finale di rendere accessibili in Europeana tutte le principali opere in pubblico dominio pubblicate in Europa a partire dal 2015.

2 La digitalizzazione del patrimonio culturale: il pubblico dominio

Le biblioteche sono le depositarie di un ingente patrimonio culturale composto di: opere in pubblico dominio, opere orfane, opere tutelate dai diritti. Per ognuna di queste tipologie di opere si pongono in essere problematiche di vario tipo in merito all'attività di digitalizzazione. Il pubblico dominio "comprende tutte le conoscenze e informazioni, ad esempio libri, immagini e opere audiovisive che non dispongono di protezione tramite copyright e possono essere utilizzate senza limitazioni, nonostante in alcuni paesi siano soggetti ai diritti morali dell'autore".[7] Il pubblico dominio è il donatore universale che garantisce l'accesso alla conoscenza. È uno dei prerequisiti a garanzia del principio espresso dall'articolo 27, comma 1 della "Dichiarazione Universale dei Diritti dell'Uomo" che recita: "Ogni individuo ha diritto di prendere parte liberamente alla vita culturale della comunità, di godere delle arti e di partecipare al progresso scientifico ed ai suoi benefici." Nell'interpretazione ampia e internazionalmente orientata di Communia, il progetto di Network tematico europeo sul pubblico dominio finanziato dalla Commissione Europea,[8] il pubblico dominio riguarda tre differenti classi di opere:

- opere che rientrano nel pubblico dominio essendo scaduti i termini della tutela del copyright.[9] Il riferimento è alla normativa sul copyright nel suo senso più ampio includendo sia i

[7] Definizione tratta da" Lo Statuto per il pubblico dominio di Europeana", aprile 2010.

[8] http://www.communia-project.eu.

[9] In questo lavoro utilizzeremo i termini "copyright" e "diritto di autore" come sinonimi. Tuttavia è bene ricordare che il primo termine – letteralmente diritto di copia– fa riferimento agli ordinamenti normativi di matrice anglosassone, mentre il secondo appartiene alla tradizione normativa dell'Europa continentale.

diritti di sfruttamento economico dell'opera e che i diritti morali. Nella maggior parte degli Stati europei i diritti di autore scadono 70 anni dopo la morte dell'autore che ha vissuto più a lungo. Nella legislazione italiana i diritti morali e patrimoniali sono trattati diversamente. Infatti, in base all'articolo 22 comma 1 della legge italiana sul diritto di autore (legge 633/1941 e successive modificazioni) i diritti morali sono inalienabili e solo quelli patrimoniali possono essere ceduti;

- opere per le quali i titolari dei diritti hanno volontariamente scelto di condividere il proprio lavoro e renderlo riutilizzabile, ad esempio adottando licenze di tipo Creative Commons o dedicando l'opera al pubblico dominio;[10]
- opere che non sono tutelate dal copyright in quanto prive di originalità intellettuale.

Sono, inoltre, in pubblico dominio: le idee, i processi, i fatti, i sistemi, i metodi, i concetti a prescindere dalla forma in cui vengono descritti. Il diritto di autore tutela, infatti, la forma ma non il contenuto di un'opera. Si ispirano, infine, al concetto di pubblico dominio tutte le eccezioni e limitazioni al copyright che "assicurano l'esistenza di un sufficiente accesso alla cultura e alla conoscenza condivise, consentendo il funzionamento delle istituzioni sociali essenziali e la partecipazione sociale di individui con necessità particolari".[11] Il pubblico dominio è il "bene comune della conoscenza". Ha sia un valore sociale che un valore economico: si pensi al caso del software libero o al valore economico dei dati aperti, pubblicati in rete e riutilizzabili (Frosio). Internet, i progetti di digitalizzazione di massa, il successo del web 2.0 nonché la mancanza in Europa di un quadro legislativo consistente e pienamente armonizzato in tema di

[10] È il caso, ad esempio, del software libero.
[11] Citazione tratta da il "Manifesto del Pubblico Dominio" di Communia.

diritto di autore[12] hanno inasprito negli ultimi anni la tensione, da sempre esistente, tra diritto di autore e pubblico dominio. In merito ai progetti di digitalizzazione il caso delle opere in pubblico dominio è, in apparenza, un caso semplice. Le opere in pubblico dominio possono essere liberamente riprodotte e, quindi, digitalizzate. Non è un caso che gran parte dei progetti di digitalizzazione avviati dalle biblioteche a partire dalla metà degli anni Novanta – LiberLiber, il progetto Manuzio, il Gutenberg Project,[13] il progetto francese Gallica[14] – si siano concentrati su questo tipo di materiale che non pone problemi in merito al copyright ed al processo di selezione.[15] La scelta, inizialmente quasi obbligata, di concentrarsi sul materiale di pubblico dominio ha pesantemente condizionato negli ultimi dieci anni i progetti di digitalizzazione avviati dalle biblioteche: gran parte del materiale utile alla ricerca scientifica è rimasto, infatti, escluso, e lo è tuttora, dall'attività di digitalizzazione.[16] In merito alle opere

[12]Ciò nonostante l'adozione della Direttiva 2001/29 EC del 22 maggio 2001 sull'armonizzazione di taluni aspetti del diritto di autore e dei diritti connessi nella società dell'informazione, recepita con G.U. n. 167 del 22 giugno 2001 http://www.interlex.it/testi/01_29ce.htm.

[13]L'idea del Project Gutenberg viene lanciata negli Stati Uniti nel lontano 1971 da Michael Hart. Fratello ufficiale del Project Gutenberg è il Project Gutenberg Australia.

[14]Il progetto Gallica prevede, in realtà, anche la digitalizzazione di materiale protetto dal diritto di autore.

[15]Non così i progetti di digitalizzazione di massa. Ad esempio, il progetto Google Book Search. De Robbio («La gestione dei diritti nelle digitalizzazioni di massa: un'analisi alla luce del caso Google Book Search») fornisce le seguenti percentuali sul materiale digitalizzato da Google: 70% del materiale rientra nella categoria delle opere fuori commercio, comprese le opere orfane, il 20% fa parte di opere in pubblico dominio e il 10% deriva da libri in commercio protetti da copyright. Nell'accordo tra Google e Università di Oxford, tuttavia, è stata prevista la digitalizzazione di sole opere in pubblico dominio.

[16]Smith («Copyright risk management: principles and strategies for large-scale digitization projects in special collections»), ad esempio, discute il disagio percepito dai bibliotecari statunitensi di fronte all'impossibilità di completare la digitalizzazione delle proprie collezioni a causa dei limiti imposti dal copyright

di pubblico dominio esistono due nodi fondamentali da sciogliere in relazione al risultato (*output*) dell'attività di digitalizzazione:

- il primo nodo è se l'attività di digitalizzazione crei o meno nuovi diritti sull'opera;

- il secondo nodo, strettamente collegato con il primo, con il tipo di licenza adottata per le opere digitalizzate e con gli eventuali accordi conclusi con terzi, è la possibilità di riutilizzare liberamente i contenuti digitalizzati.

Quanto al primo quesito il tema appare ancora controverso. Esistono, infatti, due differenti interpretazioni in merito ai diritti che scaturiscono dall'attività di digitalizzazione: da un lato la tesi ormai prevalente di chi sostiene che il semplice procedimento di digitalizzazione non crea alcun nuovo diritto d'autore o diritto connesso alla versione digitale; dall'altro la teoria di chi ipotizza la creazione di nuovi diritti sull'opera derivanti dalla digitalizzazione, dal procedimento di riconoscimento ottico dei caratteri e, soprattutto, dall'arricchimento tramite metadati descrittivi e strutturali dell'opera digitalizzata.

Savenjie e Beunen («Cultural Heritage and the Public Domain») osservano che l'ipotesi di considerare i metadati come valore aggiunto intellettuale all'opera non solo condiziona la libera riutilizzazione delle opere di pubblico dominio ma rischia di avallare le pretese commerciali delle organizzazioni che finanziano o agiscono come sponsor nella digitalizzazione di questo tipo di opere. Quando l'opera in pubblico dominio viene arricchita nella fase di postproduzione da note, commenti, apparati critici che aggiungono un reale valore intellettuale al testo allora l'opera digitalizzata rientra nel caso delle edizioni critiche che vengono tutelate dall'articolo 85-quater della vigente legge italiana sul diritto di autore per venti anni dalla loro

realizzazione.[17] Ancora diverso è il caso di chi realizza in Europa e in Italia con un investimento rilevante sotto il profilo quantitativo e qualitativo una banca dati elettronica a partire da opere in pubblico dominio. In questo caso, infatti, pur essendo il contenuto in pubblico dominio, la banca dati nel suo insieme finisce per rientrare nel diritto *sui generis* relativo alla tutela giuridica delle banche dati, così come previsto dalla Direttiva 96/9/EC dell'11 Marzo 1996.[18] Si tratta di un diritto ideato *ad hoc* per favorire gli investimenti delle imprese europee nel settore dell'*information technology*. Sul piano concettuale il diritto *sui generis* è ben distante dalla tutela giuridica della creatività dell'opera. Trattasi, infatti, come scrive Pascuzzi (*Il diritto nell'era digitale*), di "un diritto di privativa in capo all'investitore sull'estrazione e reimpiego dei dati ". Quanto alla durata della tutela l'art. 10 della Direttiva 96/9 fissa in quindici anni la durata del diritto *sui generis*.[19] Per accelerare il processo di selezione delle opere da digitalizzare è fondamentale poter verificare preventivamente se un'opera è rientrata nel pubblico dominio. In Europa uno strumento utile a tal scopo è il Public Domain Calculator,[20] svilup-

[17] Legge 633/1941, Art. 85-quater: 1. Senza pregiudizio dei diritti morali dell'autore, a colui il quale pubblica, in qualunque modo o con qualsiasi mezzo, edizioni critiche e scientifiche di opere di pubblico dominio spettano i diritti esclusivi di utilizzazione economica dell'opera, quale risulta dall'attività di revisione critica e scientifica; 2. Fermi restando i rapporti contrattuali con il titolare del diritti di utilizzazione economica di cui al comma 1, spetta al curatore della edizione critica e scientifica il diritto alla indicazione del nome. 3. La durata dei diritti esclusivi di cui al comma 1 è di venti anni a partire dalla prima lecita pubblicazione, in qualunque modo o con qualsiasi mezzo effettuata.

[18] Direttiva recepita in Italia con d.lgs 6 maggio 1999 n. 169. La Direttiva 96/9 definisce una banca dati come "una raccolta di opere, dati o altri elementi indipendenti sistematicamente o metodicamente disposti ed individualmente accessibili grazie a mezzi elettronici o in altro modo".

[19] Sul diritto *sui generis* si legga Di Cataldo («Banche dati e diritto sui generis: la fattispecie costitutiva»). Sulla durata del diritto *sui generis* si legga Cardarelli («Il diritto sui generis: la durata»).

[20] http://outofcopyright.eu

pato nel 2009 dalla Biblioteca Nazionale Austriaca nell'ambito del network Europeana Connect, creato *ad hoc* per le esigenze di sviluppo di Europeana. Per verificare se un'opera pubblicata negli Stati Uniti è in pubblico domino va, invece, utilizzato il Copyright Term Calculator.[21] I due calcolatori sono semplici strumenti di verifica. Non sono stati concepiti come registri di opere in pubblico dominio. Per usufruirne al meglio è, quindi, necessario avere effettuato una preventiva accurata ricerca sull'opera.

3 Il ri-utilizzo delle opere in pubblico dominio: le licenze

In linea di principio ciò che è in pubblico dominio dovrebbe restare tale.[22] È quanto ribadisce lo Statuto per il Pubblico Dominio di Europeana:[23]

> «Il controllo esclusivo sulle opere di dominio pubblico non può essere ristabilito rivendicando diritti esclusivi di riproduzione tecnica delle opere o utilizzando misure tecniche e/o contrattuali per limitare l'accesso alle riproduzioni tecniche di tali opere. Le opere presenti nel dominio pubblico in forma analogica continuano ad essere di dominio pubblico anche dopo la digitalizzazione.»

Così anche il rapporto del Comité des Sages che invita le istituzioni culturali a considerare modalità di recupero dei costi sostenuti per la digitalizzazione di opere in pubblico dominio alternative alla

[21] http://www.publicdomainsherpa.com/calculator.html

[22] Ciò vale anche nell'ipotesi che la digitalizzazione crei nuovi diritti sull'opera. Anche in questo caso, infatti, l'ente che digitalizza può scegliere di rilasciare la copia digitalizzata in pubblico dominio.

[23] http://pro.europeana.eu/web/europeana-project/public-domain-charter-it.

commercializzazione e sconsiglia l'apposizione di filigrane intrusive e di altri dispositivi visivi di protezione sull'opera digitalizzata. Nel mondo digitale il sistema adottato per normare il rapporto legale con gli utenti è quello delle licenze: dal software commerciale a quello Open Source,[24] dai contratti di licenza di uso delle risorse elettroniche commerciali alle pubblicazioni scientifiche[25] l'uso delle licenze si è esteso a tal punto da mettere in discussione la centralità stessa della legge ovvero, per citare ancora Pascuzzi (*Il diritto nell'era digitale*):

> «la rivoluzione digitale mette in primo piano il contratto e la tecnologia mentre la legge perde la sua centralità e diventa uno strumento che, al limite, serve solo a rafforzare il controllo basato sui primi due strumenti normativi».

Rispetto al pubblico dominio esistono diversi tipi di licenze: la Public Domain Mark 1.0, la CC0 1.0 Universal, come tipo di licenze Creative Commons, e la Open Data Commons Public Domain Dedication and License della Open Knowledge Foundation. Per quanto simili le tre licenze di pubblico dominio non hanno esattamente la stessa funzione e, come vedremo, il loro ambito di applicazione, pur sovrapponendosi, resta in parte differente. Il Public Domain Mark 1.0 (Marchio di Pubblico Dominio) è stato rilasciato nel 2010 da Creative Commons allo scopo di "etichettare quelle opere che si ritengano non più soggette alle limitazioni previste dalle norme di diritto d'autore (ed eventuali diritti connessi) quali opere entrate in pubblico dominio."[26] L'adozione del Marchio di Pubblico Dominio consente di comunicare chiaramente al pubblico l'accessibilità ai contenuti di un'opera che è già in pubblico dominio e di dichiarare la riutilizzabilità degli stessi. Creative Commons non raccomanda,

[24] Si pensi alla licenza di tipo GNU.
[25] Si pensi alle licenze di tipo Creative Commons.
[26] http://www.creativecommons.it.

tuttavia, l'utilizzo del Marchio di Pubblico Dominio per quelle opere il cui status secondo il diritto d'autore differisca da giurisdizione a giurisdizione. La licenza CC0 1.0 Universal Public Domain Dedication è anch'essa una licenza di pubblico dominio ma è stata concepita da Creative Commons per le necessità di quegli autori che dedicano il proprio lavoro al pubblico dominio, rinunciando volontariamente alla tutela del copyright.[27] Creative Commons consiglia di non utilizzare la licenza CC0 per opere protette rientrate in pubblico dominio. La licenza CC0 è, altresì, l'unica licenza CC la cui adozione viene consigliata per consentire il libero riuso dei dati.[28] Nel 2012 Europeana ha adottato la licenza CC0 per rendere disponibili i metadati del portale (nel 2012 Europeana conteneva 20 milioni di documenti provenienti da più di 2.000 istituzioni culturali), consentire il riuso dei dati come Linked Open Data (LOD) e l'utilizzo delle Application Programming Interface (API) da parte dei siti web dei partner anche commerciali. Infine l'adozione della licenza CC0 da parte di Europeana permetterà di condividere i dati del portale europeo con Wikipedia. L'Open Data Commons Public Domain Dedication and License (PDDL) è stata concepita dall'Open Data Commons (ODC). Il progetto ODC nasceva negli Stati Uniti nel 2007 per realizzare strumenti legali per la condivisione di dati in rete. La PDDL viene rilasciata nel 2008. Scaturisce dall'esigenza di creare una licenza di pubblico dominio utile a pubblicare in pubblico dominio un database o il suo contenuto o entrambi. Si legge, infatti, nel preambolo della licenza:

> «Many databases are covered by copyright. Some jurisdictions, mainly in Europe, have specific special rights

[27] http://creativecommons.org/publicdomain/zero/1.0

[28] E', infatti, tra le licenze consigliate dai Principles on Open Bibliographica Data http://openbiblio.net/principles; la traduzione italiana: Principi per i dati bibliografici aperti è disponibile alla URL http://openbiblio.net/principles/it.

> that cover databases called the "sui generis" database right. Both of these sets of rights, as well as other legal rights used to protect databases and data, can create uncertainty or practical difficulty for those wishing to share databases and their underlying data but retain a limited amount of rights under a "some rights reserved" approach to licensing as outlined in the Science Commons Protocol for Implementing Open Access Data. As a result, this waiver and licence tries to the fullest extent possible to eliminate or fully license any rights that cover this database and data».[29]

Attualmente la licenza PDDL 1.0 è mantenuta dall'Open Knowledge Foundation, un'organizzazione no-profit che ha tra i suoi obiettivi quello di promuovere l'utilizzo e la condivisione della conoscenza in rete. Talora le istituzioni culturali, pur adottando il principio del pubblico dominio, ritengono necessario mantenere, per quanto possibile, un controllo sull'opera digitalizzata e optano per un riconoscimento formale del lavoro di digitalizzazione svolto. In questo caso è possibile adottare le licenze CC-BY (attribuzione) e CC-BY-SA (attribuzione-condividi allo stesso modo) che prevedono l'attribuzione e consentono, allo stesso tempo, un ampio riutilizzo, anche commerciale, dell'opera.[30] Quanto a quest'ultimo in linea di principio è da sconsigliare, soprattutto, se la digitalizzazione è stata finanziata con fondi pubblici. Diverso è il caso delle partnership pubblico-privato che, data la crescente carenza di fondi pubblici, stanno diventando strategicamente sempre più rilevanti nei progetti

[29] http://opendatacommons.org/licenses/pddl/1.0.

[30] Non così invece la licenza CC-BY-NC che vieta un riutilizzo commerciale dell'opera. Per scoraggiare un uso commerciale da parte di terzi della copia digitalizzata è, comunque, buona prassi quella di rendere disponibile in rete una copia a bassa risoluzione dell'opera. Questa prassi è anche funzionale all'esigenza dell'utente finale di scaricare velocemente il file dalla rete.

di digitalizzazione.[31] In base ad accordi specifici, che sarebbe buona prassi rendere pubblici, i progetti di digitalizzazione realizzati dalle istituzioni culturali in collaborazione con partner privati possono prevedere alcuni limiti al ri-uso dei contenuti in pubblico dominio digitalizzati, ad esempio imponendo un periodo di embargo o un limite geografico alla diffusione dell'opera. È il caso dell'accordo concluso da ProQuest con la Biblioteca Nazionale Centrale di Firenze (BNCF) nell'ambito del progetto "Early European Books" (EEB) per la digitalizzazione delle fonti stampate in Europa fino al 1700.[32] L'accordo prevede l'accesso gratuito dal territorio italiano a tutti i contenuti BNCF digitalizzati da ProQuest. L'accesso alle opere è, invece, ristretto al di fuori dei confini nazionali per 15 anni. Il rapporto del Comité des Sages considera ragionevole un embargo massimo di 7 anni per le opere digitalizzate dalle istituzioni culturali in partnership con i privati.

4 La digitalizzazione del patrimonio culturale : i problemi delle opere orfane

Se nel caso della digitalizzazione di opere in pubblico dominio esistono dubbi e perplessità che, di fatto, non condizionano l'attività di digitalizzazione ma i suoi risultati, i problemi legati al mancato riconoscimento della paternità intellettuale delle opere orfane sono, invece, decisamente più complessi e rappresentano un serio ostacolo alla digitalizzazione e condivisione in rete di questa tipologia di opere.

[31] Sulle partnership tra soggetti pubblici e privati si legga il rapporto finale dell'High Level Expert Group on Digital Libraries – Subgroup on Public Private Parternships («Final report on Public Private Partnerships for the Digitization and Online Accessibility of Europe's cultural heritage»).

[32] http://eeb.chadwyck.com/marketing/about.jsp

> «Le opere orfane sono quelle opere delle quali con il trascorrere del tempo [per molteplici motivi n.d.a.][33] si sono perse le tracce dei titolari dei diritti di autore patrimoniali e/o morali» (Fabiani).

Non vanno confuse con le opere anonime, la cui paternità è ignota. Le opere orfane possono essere considerate un sottoinsieme delle opere fuori commercio, in quanto per le opere in commercio è dato conoscere almeno l'editore che, nella maggior parte dei casi, è il detentore dei diritti di sfruttamento economico sull'opera. Tutte le opere orfane rientrano, quindi, nella categoria delle opere fuori commercio, ma non tutte le opere fuori commercio sono di fatto opere orfane. Un'opera può essere interamente o parzialmente orfana, quando solo alcuni titolari dei diritti sono noti o rintracciabili o quando solo una parte dell'opera è nella condizione di opera orfana. Uno studio commissionato dal JISC nel 2009 stimava la consistenza ed il numero di opere orfane esistenti in Gran Bretagna in una percentuale variabile tra il 5 e 10%. Lo stesso studio rivelava che nel 60% dei casi esaminati le opere orfane avevano avuto un impatto negativo sui progetti di digitalizzazione delle istituzioni culturali (JISC). Questa stima è puramente indicativa e di per sé, in realtà, anche poco significativa. Il numero delle opere orfane varia, infatti, enormemente a seconda dei contesti legislativi e territoriali, nonché in relazione alla tipologia e consistenza delle collezioni e, quindi, varia sostanzialmente da biblioteca a biblioteca. La British Library, ad esempio, ha calcolato che il 40% delle proprie collezioni appartiene alla categoria delle opere orfane. Nel 2005 un'indagine condotta sulle proprie collezioni dalle biblioteche della Carnegie Mellon University (USA) ha rivelato che il 22% degli aventi diritto delle opere

[33] Ad esempio i titolari dei diritti non sono noti oppure sono noti ma non più rintracciabili o, ancora, sono defunti e non ci sono altre informazioni relative ad eventuali eredi o successori.

da digitalizzare non era più rintracciabile. La percentuale di opere orfane tende a crescere per tipologie di materiale diverse dalle pubblicazioni a stampa, ad esempio: il materiale fotografico, le opere audiovisive attualmente conservate presso biblioteche, musei o archivi, il materiale sonoro e, naturalmente, la letteratura grigia.[34] Le opere orfane non possono essere liberamente riutilizzate in quanto:

- non si può ottenere l'autorizzazione preventiva dei titolari del diritto di autore così come previsto dalla Direttiva 2001/29/CE del 22 maggio 2001 sull'armonizzazione di taluni aspetti del diritto di autore e dei diritti connessi nella società dell'informazione;[35]

- la maggior parte degli Stati europei non ha ancora definito una normativa che disciplini la questione delle opere orfane. Fanno eccezione i Paesi del Nord Europa nei quali l'utilizzo delle opere orfane è regolato dal sistema delle licenze estese di tipo collettivo.[36]

La questione delle opere orfane è stata ripetutamente affrontata dalla Commissione europea in quanto nodo cruciale dei progetti di digitalizzazione e conservazione del patrimonio culturale europeo. Nella sopra citata raccomandazione del 24 agosto 2006 la Commissione aveva già invitato gli Stati membri ad istituire gli strumenti

[34]La complessità dei diritti legati alle diverse tipologie di documenti ha spinto la Commissione europea a produrre studi per i diversi settori: opere fotografiche, audiovisive, settore musicale/sonoro.

[35]La stessa direttiva pone, tuttavia, all'articolo 5 un'eccezione a favore degli atti di riproduzione effettuati dalle biblioteche quando l'utilizzo sia fatto a scopo di ricerca o di attività privata di studio su terminali dedicati situati nei locali di queste istituzioni.

[36]Ad esempio in Danimarca e in Norvegia per il progetto Bokhylla http://www.nb.no/bokhylla. Sulle licenze collettive estese nei Paesi del Nord Europa si legga Riis e Schovsbo («Extended collective licenses and the Nordic experience: it's a hybrid but it is a Volvo or a lemon?»).

necessari al facilitare l'utilizzo in rete delle opere orfane, ad esempio, promuovendo la disponibilità di elenchi di opere orfane note. Nel *Final Report on Digital Preservation, Orphan Works, and Out-of-Print Works* dell'High Level Expert Group – Copyright Subgroup (Gruppo di lavoro di esperti sul copyright), costituito in seguito alla già citata Comunicazione della Commissione sulle biblioteche digitali allo scopo specifico di mettere in luce i problemi di diritti che le istituzioni culturali europee devono affrontare nei progetti di digitalizzazione, questi strumenti venivano individuati nella creazione di un database di opere orfane e di un centro (europeo) – o più centri (nazionali) – di Rights Clearance (Rights Clearance Centres). Di fatto è soprattutto la mancanza di informazioni sui diritti delle opere orfane a determinare questa *empasse* gestionale e strategica con conseguenze rilevanti sulle attività di digitalizzazione, di conservazione e tutela del patrimonio culturale.[37] Il Gruppo di lavoro di esperti sul copyright elencava, quindi, nel suo rapporto alcuni principi chiave per la costruzione di un database di opere orfane (adozione di policy, criteri di interoperabilità, adozione di standard, struttura e contenuto del database, numero minimo di metadati) e dei Rights Clearance Centres (policy dei RCC, policy per le licenze, policy per la ricompensa dei titolari dei diritti, adozione di criteri di interoperabilità e di trasparenza) e suggeriva di stabilire criteri comuni per effettuare una "diligente ricerca" in merito alle opere orfane nei diversi Stati membri. L'interoperabilità ed il riconoscimento reciproco delle soluzioni adottate sono, infatti, due elementi chiave nell'economia della ricerca di informazioni sulle opere dell'intelletto e sui diritti connessi. Infine, il Gruppo di lavoro di esperti sul copyright sottolineava la necessità di minimizzare in futuro la possibilità di dare origine a nuove opere orfane. A tal fine propo-

[37] Nel caso dei film orfani, ad esempio, l'impossibilità di individuare il titolare dei diritti comporta, tra l'altro, l'impossibilità di procedere al restauro del film.

neva di arricchire i file digitali con i metadati dei diritti sull'opera e di associare alle risorse digitali un identificativo persistente. Nel Libro verde su "Il diritto di autore nell'economia della conoscenza" del 16 luglio 2008 la Commissione europea metteva nuovamente in primo piano il problema delle opere orfane considerandole un ostacolo alla diffusione in rete del patrimonio culturale e ponendo il problema come una liberatoria dei diritti "nel senso che occorre garantire che gli utenti che mettono a disposizione opere orfane non vengano poi chiamati a rispondere di una violazione del diritto d'autore quando il titolare ritorna sulla scena e fa valere i propri diritti" (Fabiani). Nello stesso anno veniva siglato dai principali *stakeholders* della filiera editoriale il protocollo di intesa sulle linee guida per una diligente ricerca sulle opere orfane (*Memorandum of Understanding on Diligent Search Guidelines for Orphan Works*). La natura essenzialmente transfrontaliera della questione relativa alle opere fuori commercio, in generale, e alle opere orfane, in particolare, ha richiesto un'iniziativa di armonizzazione da parte dell'Unione Europea che ha risposto alla duplice esigenza di un'infrastruttura di ricerca sui diritti delle opere orfane e di un quadro normativo di riferimento finanziando il progetto ARROW (Accessible Registries of Rights Information and Orphan Works) e approvando nel 2012 la Direttiva sulle opere orfane.[38]

[38]Rispetto alle opere orfane numerose iniziative legislative sono in atto in diversi Paesi. In Ungheria, in Canada e negli Stati Uniti dove nel 2006 è stato pubblicato dal U.S. Copyright Office il *Report on Orphan Works*, sulla base del quale sono stati presentati ben due disegni di legge sulle opere orfane. Per quanto dissimili possano essere gli ordinamenti legislativi tra i Paesi le proposte legislative si basano tutte su un principio comune: l'utente deve preventivamente effettuare una ricerca diligente per cercare di identificare e localizzare i titolari dei diritti.

5 Il progetto ARROW e ARROWPLus

Nel caso delle opere orfane "il superamento dell'*empasse* in cui si trova la gestione dei diritti nei programmi di digitalizzazione va cercata in una combinazione tra innovazione tecnologica e innovazione normativa." (Attanasio, «La gestione dei diritti di autore nelle biblioteche digitali: il caso ARROW») Il progetto ARROW ha sviluppato uno strumento tecnico di supporto nell'individuazione e gestione dei diritti delle opere pubblicate in Europa. Cofinanziato nel settembre 2008 dalla Commissione Europa nell'ambito del programma eContentPlus – sottoprogramma *ICT Policy Support* – in un momento di grande enfasi posta sulle opere orfane e sui progetti di digitalizzazione[39] ARROW viene spesso erroneamente identificato con un database di opere orfane. In realtà è una suite di servizi e di strumenti sviluppata da un consorzio di partner e associati di oltre 30 organizzazioni di 13 paesi europei tra biblioteche nazionali ed universitarie, associazioni di editori ed autori, società di gestione collettiva dei diritti, organismi internazionali.[40]

Il progetto nasce dall'esigenza più volte espressa dalla Commissione europea e dai suoi gruppi di lavoro di sviluppare strumenti per la gestione dei diritti di autore in ambiente digitale. Gli strumenti attualmente sviluppati da ARROW sono:

1. **Rights Information Infrastructure (RII)**: cioè un'infrastruttura distribuita basata su standard aperti per la gestione delle informazioni dei diritti d'autore connessi alle opere lettera-

[39] All'interno del programma eConterntPlus è stato finanziato anche il progetto triennale Metadata Image Library Exploitation (MILE), obiettivo del quale è arricchire di metadati i file di immagini digitalizzate in modo da favorirne la ricerca e la conservazione.

[40] Tra gli altri: l'European Digital Library Foundation (EDL), l'International Federation of Reproduction Rights Organizations (IFRRO), la Federation of European Publishers (FEP), l'European Visual Artists, The European Writers' Congress.

rie. Il RII non è un vero e proprio registro, ma è concepito come un'infrastruttura, un punto di raccolta di informazioni che provengono da fonti diverse e contribuiscono a creare un quadro di riferimento per la gestione dei diritti di un'opera. In quanto tale qualunque strumento che consenta la gestione delle informazioni sui diritti può essere considerato un RII. Attanasio («Rights Information infrastructures and voluntary stakeholders agreements in digital library programmes»; «La gestione dei diritti di autore nelle biblioteche digitali: il caso ARROW») elenca la natura di queste informazioni sui diritti:

- una precisa identificazione della manifestazione dell'opera che si vuole digitalizzare;
- una precisa identificazione dell'opera/opere inclusa/e nella manifestazione;
- una precisa identificazione dei soggetti che hanno diritti sull'opera, principalmente autori, autori di contributi secondari ed editori;
- una determinazione dello status commerciale dell'opera: in commercio/fuori commercio;
- l'individuazione dei soggetti che possono concedere l'autorizzazione alla digitalizzazione.

Le fonti che alimentano il RII sono: il database TEL - The European Library, il catalogo unico delle biblioteche nazionali europee; il Virtual International Authority File (VIAF), il file di authority internazionale mantenuto da OCLC che consente di identificare in modo univoco l'autore di un'opera, il catalogo dei libri in commercio (Books in Print) di ciascuna delle nazioni coinvolte, i repertori delle società di gestione collettiva dei diritti di autore. L'ordine di citazione delle fonti corrisponde

ai tre step che articolano il workflow del RII (Caroli et al.). Dal momento che, come scrive Attanasio («Rights Information infrastructures and voluntary stakeholders agreements in digital library programmes»), "the main value [of the ARROW project] is to provide interoperability among existing resources and to foster the collection of additional data or enrichment of existing data within a network" l'interoperabilità tra le fonti[41] è stata una degli aspetti chiave da risolvere per la creazione del RII. Nel 2012 il servizio è partito in fase sperimentale nei seguenti paesi: Regno Unito, Germania, Spagna e Francia. I test effettuati durante la fase pilota del servizio hanno permesso di calcolare un risparmio di tempo nell'individuazione dei diritti connessi ad un'opera che varia tra il 72% e il 97% (Caroli et al.) con una considerevole riduzione dei costi di transazione. ARROW è stato già utilizzato come strumento di verifica nel progetto di digitalizzazione dei testi scientifici della collezione privata del Wellcome Trust in Gran Bretagna e da Europeana per la collezione dedicata alla Prima guerra mondiale che sarà inaugurata a gennaio 2014;

2. **Arrow Work Registry (AWR)**: un repository delle opere orfane e di altre categorie specifiche di opere che viene popolato attraverso le sottomissioni fatte nel RII e popola, a sua volta, il **Registry of Orphan Work (ROW)** dedicato esclusivamente alle opere orfane. Il registro è lo strumento indispensabile che rende pubblici i risultati della ricerca del RII attestando se un'opera è orfana. Con ARROWPlus (1 aprile 2011- 30 settembre 2013) gli obiettivi del progetto ARROW sono stati ampliati. In particolare ARROWPlus prevede l'estensione della copertura del RII a nuovi Paesi, tra i quali l'Italia, l'allargamento a *stake-*

[41] Ad esempio tra il formato MARC che è il formato dei cataloghi di biblioteche e il formato ONIX dei cataloghi editoriali.

holders dei diversi domini, lo sviluppo di nuove funzionalità del sistema, la gestione dei diritti per le immagini contenute nei libri. L'Associazione Italiana Editori (AIE) è coordinatore generale del progetto, il CINECA è il partner tecnologico, mentre l'ICCU svolge la funzione di coordinatore nazionale e di *National contact point* per le biblioteche italiane (Martini). A tal fine l'ICCU dovrà coordinarsi con la Direzione Generale per le biblioteche, gli istituti culturali ed il diritto d'autore, che ha competenze in materia di diritto d'autore.

6 La Direttiva 2012/28/UE sulle opere orfane

Come risultato del cammino intrapreso nel 2006 con la Raccomandazione 2006/585/CE del 24 Agosto 2006 ad ottobre 2012 è stata approvata la direttiva europea su taluni usi consentiti delle opere orfane (Direttiva 2012/28/UE del 25 ottobre 2012). La Direttiva si inserisce nell'iter di semplificazione ed armonizzazione legislativa che l'Unione Europea ha intrapreso nel 2001 con la già citata Direttiva 2001/29/EC e rientra nella strategia "Europe 2020: a strategy for smart, sustainable and inclusive growth" che rappresenta uno dei punti cardine dell'Agenda Digitale Europea. Il diritto dell'Unione Europea ad agire rispetto al mercato interno emana direttamente dall'articolo 114, comma 1 del "Trattato di funzionamento dell'Unione Europea", versione consolidata, 2008/C 115/01 che sancisce il potere del Parlamento e del Consiglio europeo di regolamentare il funzionamento del mercato in terno: "Il Parlamento europeo e il Consiglio, deliberando secondo la procedura legislativa ordinaria e previa consultazione del Comitato economico e sociale, adottano le misure relative al ravvicinamento delle disposizioni legislative,

regolamentari ed amministrative degli Stati membri che hanno per oggetto l'instaurazione ed il funzionamento del mercato interno."[42] La Direttiva 2012/28 è stata concepita per facilitare la creazione di biblioteche digitali e consentire la digitalizzazione e diffusione di massa delle opere riconosciute orfane. Il suo ambito di applicazione è vasto ma "limitato" a: libri, riviste, quotidiani, rotocalchi o altre pubblicazioni conservate nelle collezioni delle biblioteche, istituti di istruzione o musei accessibili al pubblico, nonché nelle collezioni di archivi o di istituti per il patrimonio cinematografico o sonoro. La Direttiva si applica altresì a opere audiovisive, cinematografiche e fonogrammi. Non si applica, invece, alle immagini fotografiche. La Direttiva lascia impregiudicato lo sviluppo di soluzioni specifiche negli Stati membri per far fronte a questioni di più ampia scala sulla digitalizzazione di massa, come nel caso delle cosiddette opere «fuori commercio» (considerando 4).

L'art. 3, comma 1 della Direttiva stabilisce che per poter dichiarare un'opera "orfana" è necessario effettuare una "ricerca diligente" e "in buona fede" nello Stato membro di prima pubblicazione dell'opera. Occorre che tale ricerca diligente sia regolata da un approccio armonizzato nei diversi Stati membri. Le fonti appropriate per determinare lo status di opera orfana per le singole categorie di opere vengono individuate dai singoli Stati membri (art. 3 comma 2). Tra le fonti utili a completare la ricerca la direttiva cita : ARROW, il VIAF, le banche dati delle società di gestione collettiva, in particolare quelle delle organizzazioni che gestiscono i diritti di riproduzione, i cataloghi di biblioteche, le banche dati, i cataloghi dei libri in commercio. La Direttiva prevede, altresì, che un'opera riconosciuta orfana in uno Stato membro secondo i criteri stabiliti dalla normativa stessa sarà considerata tale anche negli altri (art. 4).

[42] Nella sua Revisione del mercato interno (COM(2007) 724 definitivo del 20.11.2007) la Commissione sottolineava la necessità di promuovere la libera circolazione del sapere e dell'innovazione nel mercato unico in quanto "quinta libertà."

Il principio di reciproco riconoscimento risponde ad una duplice esigenza:

- di evitare sforzi di duplicazione tra gli Stati membri. Si tratta di un punto di forza del documento che consentirà ad un'opera dichiarata orfana in uno Stato di essere riutilizzata in tutti gli Stati membri;
- di rendere disponibile l'opera orfana digitalizzata a tutti i cittadini dell'Unione Europea.

La Direttiva consente, inoltre, di digitalizzare e comunicare al pubblico gli inediti posseduti dalle biblioteche e dagli istituti culturali quando è ragionevole supporre che i titolari dei diritti non si opporrebbero alla comunicazione al pubblico delle loro opere. Tra gli aspetti positivi della Direttiva 2012/28 a parte il già citato principio di reciproco riconoscimento va ricordato anche l'art. 1, comma 5 che esplicita che la Direttiva non interferisce con le modalità di gestione dei diritti a livello nazionale. In tal modo il legislatore europeo ha inteso conciliare i dettami della Direttiva con le normative di quei Paesi che abbiano adottato un sistema di licenze collettive estese o abbiano una legislazione che regoli l'utilizzo di opere orfane. Così, ad esempio, la Francia che, recentemente, ha approvato una legge sull'utilizzo in rete delle opere fuori commercio, che norma anche l'accesso alle opere orfane.[43] Nonostante gli elementi positivi sopra evidenziati la Direttiva sulle opere orfane ha dato origine a non poche critiche e perplessità da parte dei diversi attori della filiera editoriale. In modo particolare tre sono i punti critici che sembrerebbero avere un impatto negativo sui progetti di digitalizzazione:

1. la natura ambigua della definizione di "ricerca diligente" cui si aggiunge la mancata specificazione dei criteri per effettuare

[43] Loi n. 2012-287 du 1er mars 2012 relative à l'exploitation numérique des livres indisponibles de XXe siècle.

tale diligente ricerca. La Direttiva lascia agli Stati membri il compito di definire tali criteri;

2. la mancanza di soluzioni convincenti per superare il problema della ricerca dei titolari dei diritti sulle opere incorporate;

3. il riconoscimento di un "equo compenso" per il titolare dei diritti (articolo 6, comma 5) per le utilizzazioni effettuate sulla sua opera durante il periodo in cui non era stato individuato o rintracciato, compenso che non era previsto nella versione iniziale della proposta di direttiva e che rischia di vanificarne in gran parte l'utilità.[44]

Tra le critiche più ragionate anche quella espressa da Information sans Frontière (organizzazione che unisce EBLIDA, LIBER, JISC ed Europeana)[45] che sottolinea come il riconoscimento di tale equo compenso tenda a scoraggiare quelle istituzioni pubbliche, e ancor più i partner privati, che vogliano utilizzare l'opera orfana in progetti di digitalizzazione, quand'anche, così come previsto dal considerando 18, il livello dell'equo compenso tenga in conto gli obiettivi di promozione culturale degli Stati membri e la natura non commerciale dell'utilizzo fatto dalle organizzazioni:

> «this doubtful situation will discourage public institutions wishing to digitise their collections, even if the assessment of the level of compensation 'takes into account' the cultural purpose of the use».

Quanto al riutilizzo commerciale, tenendo in debito conto gli obiettivi di promozione culturale degli Stati membri, la Direttiva 2012/28 esclude che si possano acquisire dei diritti commerciali di alcun genere sull'opera riconosciuta orfana.

[44] Per questo commento sono debitrice verso l'Osservatorio diritto di autore e Open Access dell'AIB.

[45] http://informationsansfrontieres.eu.

7 Conclusioni

I temi sopra discussi non esauriscono le tematiche connesse con il riutilizzo in rete e la conservazione a lungo termine del patrimonio culturale. Non si è voluto affrontare in questo contributo il nodo delle opere fuori commercio che rappresentano un altro ambito estremamente problematico nei progetti di digitalizzazione. Quanto a questa tipologia di opere il *Memorandum of Understanding on key principles on the digitization and making available of out-of-commerce works* firmato nel settembre 2011 da associazioni di editori, autori, biblioteche e società di gestione rappresenta un timido passo in avanti verso l'utilizzo delle opere fuori commercio in progetti di digitalizzazione di massa, rimandando la soluzione definitiva ad accordi collettivi su base volontaria che negozino l'acquisizione delle licenze necessarie affinché le biblioteche e analoghe istituzioni culturali possano digitalizzare e pubblicare online questa tipologia di opere.[46] L'apertura e la disponibilità al dialogo tra editori, biblioteche e società di gestione collettiva resta la chiave per risolvere gran parte dei nodi relativi al riutilizzo ed alla conservazione a lungo termine delle opere orfane e delle opere fuori commercio. Il modello delle licenze collettive estese, adottato dagli Stati del Nord Europa, appare una soluzione estremamente pragmatica ed efficace per far fronte alle necessità dei progetti di digitalizzazione di massa. Si tratta, purtroppo, di un modello difficilmente esportabile in altri Stati membri.

Volendo, in conclusione, proporre una riflessione più ampia non si può non sostenere la visione di quanti invocano una nuova era del copyright nel mondo digitale, in primo luogo riducendo la sua durata[47] e semplificando il quadro normativo, ma, soprattutto, ripen-

[46] Il testo del memorandum di intesa contiene un insieme di principi fondamentali. È scaricabile alla URL http://ec.europa.eu/internal_market/copyright/out-of-commerce/index_en.htm.

[47] A favore della riduzione del copyright si sono espressi più volte eminenti giuristi

sando alle caratteristiche stesse del copyright in quanto strumento che tutela la creatività intellettuale ovvero affiancando al copyright tradizionale una tipologia di copyright più agile e flessibile, un copyright che di default lasci in capo all'autore alcuni diritti sull'opera: il Copyright 2.0:

> «creators should opt-in for Copyright 1.0 at the time of the original release of their work; otherwise the new and more flexible Copyright 2.0 would operate as a default set of provisions. This is why in the past I characterized this approach as "Lessig by default" or, in a less personalized way, "Creative Commons by default". The idea behind the approach is that the very successful uptake of Creative Commons licenses and other copyleft licenses by creators operating along the short route shows that out there, in the digital prairies and wilderness, there is a very large number indeed of creators who prefer to reserve only some rights rather than all rights» (Ricolfi).

come Lawrence Lessig e, per rimanere in ambito nazionale, Marco Ricolfi. La stessa raccomandazione è contenuta nel Manifesto per il Pubblico Dominio di Communia.

Riferimenti bibliografici

Attanasio, Piero. «La gestione dei diritti di autore nelle biblioteche digitali: il caso ARROW». *DigItalia* 6.2. (2011): 93–105. <http://digitalia.sbn.it/riviste/index.php/digitalia/article/view/477>. (Cit. alle pp. 225, 243, 244).

——. «Rights Information infrastructures and voluntary stakeholders agreements in digital library programmes». *JLIS.it* 1.2. (2010): 237–261. <http://leo.cilea.it/index.php/jlis/article/view/4539>. (Cit. alle pp. 244, 245).

Cardarelli, Maria Cecilia. «Il diritto sui generis: la durata». *AIDA. Annali italiani del diritto d'autore, della cultura e dello spettacolo* 6.1. (1997): 64–85. (Cit. a p. 233).

Caroli, Cinzia, et al. «ARROW: accessible registries of rights information and orphan works towards Europeana». *D-Lib magazine* 18.1-2. (2012). <http://www.dlib.org/dlib/january12/caroli/01caroli.html>. (Cit. a p. 245).

Caso, Roberto. «Open Access to legal scholarship and copyright rules: a law and technology perspective». *Proceedings law via the internet: free access, quality of information, effectiveness of rights*. A cura di Ginevra Peruginelli e Mario Ragona. Florence: European Press Academic Publishing, 2009. 97–110. (Cit. a p. 224).

De Robbio, Antonella. «La gestione dei diritti nelle digitalizzazioni di massa: un'analisi alla luce del caso Google Book Search». *Bibliotime* 12.2. (2009). <http://eprints.rclis.org/13506/>. (Cit. a p. 231).

Di Cataldo, Vincenzo. «Banche dati e diritto sui generis: la fattispecie costitutiva». *AIDA. Annali italiani del diritto d'autore, della cultura e dello spettacolo* 6.1. (1997): 20–28. (Cit. a p. 233).

Fabiani, Mario. «Opere orfane, diritti orfani». *Il diritto di autore* 74.2. (2009): 225. (Cit. alle pp. 239, 242).

Frosio, Giancarlo. «Communia and the European Public Domain Project: a politics of the Public Domain». *The Digital Public Domain: foundations for an Open Culture*. A cura di Melanie Dulong de Rosnay e Juan Carlos De Martin. (Cit. a p. 230).

i2010 European Digital Libraries Initiative, High Level Expert Group on Digital Libraries, Sub-group on Public Private Partnerships. «Final report on Public Private Partnerships for the Digitization and Online Accessibility of Europe's cultural heritage». (2008). <http://ec.europa.eu/information_society/activities/digital_libraries/doc/hleg/reports/ppp/ppp_final.pdf>. (Cit. a p. 238).

JISC. «In from the cold: an assessment of the scope of Orphan Works and its impact on the delivery of service to the public: research report prepared for Strategic content Alliance and Collections Trust». (2009). <http://sca.jiscinvolve.org/wp/files/2009/06/sca_colltrust_orphan_works_v1-final.pdf>. (Cit. a p. 239).

Martini, Patrizia. «ARROW Plus National stakeholder meeting, Roma 16 dicembre 2011». *DigItalia* 7.1. (2012): 153–154. <http://digitalia.sbn.it/riviste/index.php/digitalia/article/view/547>. (Cit. a p. 246).

Pascuzzi, Giovanni. *Il diritto nell'era digitale*. Bologna: Il Mulino, 2006. (Cit. alle pp. 233, 235).

Ricolfi, Mario. «Making copyright fit for the digital agenda». *12th EIPIN Congress 2011 - Constructing European IP: Achievements and new Perspectives Strasbourg*. European Parliament. 2011. (Cit. a p. 251).

Riis, Thomas e Jens Schovsbo. «Extended collective licenses and the Nordic experience: it's a hybrid but it is a Volvo or a lemon?» *Columbia Journal of Law and the Arts*. (2010). <http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1535230>. (Cit. a p. 240).

Rodotà, Stefano. «Il sapere come bene comune. Il popolo di Internet». *La Repubblica*. (2007). 2007-09-15. (Cit. a p. 224).

Savenjie, Bas e Annemarie Beunen. «Cultural Heritage and the Public Domain». *LIBER Quarterly* 22.2. (2012): 80–97. <http://liber.library.uu.nl/index.php/lq/article/view/8089>. (Cit. a p. 232).

Smith, Kevin L. «Copyright risk management: principles and strategies for large-scale digitization projects in special collections». *Research Libraries Issues* 279. (2012). <http://publications.arl.org/rli279/>. (Cit. a p. 231).

MARIA CASSELLA, Università degli studi di Torino.
maria.cassella@unito.it

Cassella, M. "La gestione dei diritti nei progetti di digitalizzazione: il pubblico dominio e le opere orfane.". *JLIS.it*. Vol. 4, n. 2 (Luglio/July 2013): Art: #8797. DOI: 10.4403/jlis.it-8797. Web.

ABSTRACT: The essay is dedicated to the theme of digital rights management on digitisation projects with particular reference to works into the public domain and to orphan works. The aim of the essay is to explain the main issues and criticalities that libraries have to face proposing and carrying out digitisation projects. For some years, the digitisation of cultural heritage involves libraries, archives and museums and is one of the main goals by the Digital Agenda for Europe. Indeed the digitisation is one essential tool to broaden the accessibility to the cultural heritage of Europe and to promote the growth. As a part of this context, the essay highlights the legal issues related to the process of digitization of cultural heritage for works into the public domain and for orphan works.

KEYWORDS: Digitisation, Public domain, Orphan works, copyright, ARROW

Submission: 2013-04-08
Accettazione: 2013-04-23
Pubblicazione: 2013-07-01

JLIS.it

In search of Meaning: The Written Word in the Age of Google

Anita Paz

Every time he learned a new word [. . .], a beautiful word like "light" – my heart curdled around the edges, because I thought, Who knows what he is losing in this moment, how many infinite kinds of glamour he felt and saw, tasted and smelled, before he pressured them into this little box, "light", with a t at the end like a switch clicking off.
(D. Grossman, *Be My Knife*, 1998)

This work intends to be a brief and certainly not comprehensive appraise of the state of the written word and its meaning in the digitalized age, as a result of an ever growing utilization of online search engines, and its effects on the individual's acquaintance with and understanding of his or her world, and is to be considered as a reading of the ideas raised by Boris Groys in *Google: Words beyond Grammar* from a library and information science point of view. According to Groys, the questions one asks the world, the answers one is willing to receive, and the medium through which one chooses to conduct this dialog, depend on one's initial world perception (Groys 4). Today, claims Groys, the individual conducts his or her philosophical interrogation through the World Wide Web, and more specifically through search engines. In fact, proceeds Groys, Google can be described as "the first philosophical machine that regulates our dia-

JLIS.it. Vol. 4, n. 2 (Luglio/July 2013).
DOI: 10.4403/jlis.it-8798

log with the world, by substituting metaphysical presuppositions with strictly formalized and universally applicable rules of access" (5). This vision sits well with Manovich's theory of the database — an unorganized list of the world's phenomena — as the cultural form appropriate for the computerized age, which created a new cultural algorithm: reality > media > data > database (Manovich 194-9). A commonly discussed quality of this great portal to the understanding of the modern world is that it is highly subjective, while the paths it leads the user on — the results it returns — are partial, pre-selected and often inaccessible. Groys puts main parts of Google's "hidden subjectivity" (Groys 15) on the user (that fails to check the majority of the results) and third parties (which restrict access to their content) (14). Analyzing the user's interaction with Google Search, I would claim that it is Google itself that knowingly and intentionally manipulates the user's search and accessibility to results, in a way which hinders not only serendipity, but also free access to information.

1 Asking the question

The first part of an individual's dialog with the world consists of a question, or in this case, a query. The user can theoretically type in the search box whatever he or she pleases — from a single character to a sequence of sentences, yet Google Zeitgeist 2012 shows the most popular searches were those consisting of one, two or (less often) three words.[1] Manning, Raghavan and Schutze (432-3) identify three types of web queries:

1. informational, which is a search for general information on a certain topic;

[1] http://www.google.com/zeitgeist/2012/#the-world/searches.

2. navigational, which is a search for a specific website;
3. transactional, a prelude to a future transaction, such as an online purchase or a download of content.

Though the last two take up an undoubtedly significant percent of the user's interaction with Google Search, this work will focus mainly on the first type of web query — the informational query. All three types of queries, however, are formulated following the same specific type of logic — namely Boolean logic. In other words, web queries are words — strings of data — sequences of characters — which may be organized using operators (e.g., "and", "or" and "not"), jolly characters and quotation marks. Each query can be further specified putting certain limitation on the possible outcomes, in the form of linguistic, typological or file format preferences. It has been shown by Groys that these rules of dialogue, permitting a correctly formulated question to take the form of a single word (or a non-grammatical combination of words), do not correspond to the rules of the spoken language (Groys 5-6). Google's definition of a legitimate question, proceeds Groys, is one that concerns the meaning of a certain word, which is, according to him, the only possible form of question feasible for Google.

Three parts constitute the Google Search software: a spider, a BigTable database (DB) and an interface. The first scans the Web for word presence, the second indexes and stores the information, and the third allows users to access the information. The indexing is done per word, so that each word has a quantity of resources (e.g., web pages, images or audio files) related to it. When a user types a certain word (or a combination of words) in the search box, Google scans its DB and returns each resource connected to that word (or combination of words) in the form of a result — a link to the site where it appears. Groys views this as a disintegration of texts into a succession of freestanding words, which turns dis-

courses into word clouds, no longer expressing an idea, but simply comprising or not comprising a certain word (Groys 7). Thus, avers Groys, the liberation of individual words from their grammatical structure eradicates the difference between an affirmative and a critical position, inducing the commutation of a linguistic operation (of affirmation or negation) for an extra-linguistic one (of inclusion or exclusion of words in contexts) — i.e., word curatorship (11-12).

2 Receiving answers

The second part of one's dialog with the world consists of the answer he or she receives. If a Google legitimate question is one about the meaning of an individual word, a legitimate answer, as it is defined by Groys, is a set of contexts in which the search word was located by the spider (Groys 5-6). Thus, the sum of contexts returned to the user by Google, represent the true meaning of the word, and since Google is the contemporary individual's main tool of interrogation, it is also the only truth to him or her accessible.

Groys' observation regarding the word's meaning, depends to a great extent on Wittgenstein's reflection on words and their meaning. For the great Austro-British philosopher the meaning of the word is not the object for which the word stands (as St. Augustine would have wanted it) (Wittgenstein N.1, 2), even though a word has no meaning if nothing corresponds to it. Never the less, to identify the "meaning" of a word with the corresponding thing is to erroneously equate the meaning of a name with the bearer of that name (N. 20, 40) — The meaning of a word is determined by its use (N. 139, 54) — its context.

Groys recalls that for Derrida a normative meaning was impossible, for the number of contexts is theoretically infinite (Groys 8-9). In this sense, Google can be viewed as a twofold response to decon-

struction: on the one hand it is based upon the same understanding of the language not having fixed normative contexts for meaning; on the other, it is also based on the believe that these contexts are finite, calculable and displayable (9-10). And so, according to Groys, by replacing what was thought to be infinite, with a finite search algorithm which looks for existent contexts, Google search has turned deconstruction upside down. Yet it does even more than that — Google returns not only the verbal context in which the word was located, but also images, maps, videos and audio files correlated to it. In fact, the answer Google is trying to provide for any given question is becoming more and more tridimensional, providing the user a dynamic multimedia Web 2.0 experience. In doing so, Google creates what seems to be a round a-posteriori understanding of the meaning of a certain word in its user's mind — in theory, this understanding should be based on the amalgamation of all contexts available; in practice, it is highly restricted, controlled and manipulated.

For Groys, Google is unable to display all contexts because some require special access, and the rest are prioritized (Groys 14). The mentioned prioritization takes place on two levels: per webpage and per user. This means that in addition to the Google algorithm assigning a PageRank to each webpage — determined on the basis of approximately 200 factors, among which the number of times the search word appeared on the page, longevity of the page, and number of external sites linking to it[2] — Google also actively profiles its users based on their IP, previously completed searches and general web behavior (Guerrini, Bianchini, and Capaccioni 91-92). Thus, a search performed by a user situated in Sweden using Google.com

[2]In regards of external links and reviews, in 2010 the New York Times has revealed that Google often does not differentiate between positive and negative reviews, high placing sites against which numerous complaints were shared — assigning a whole new meaning to "there is no such thing as bad publicity", and see: http://www.nytimes.com/2010/11/28/business/28borker.html?pagewanted=1&_r=0.

will receive different results — both content and quantity-wise — than a search performed several seconds later by the same user using Google.se; similarly, a search for "Jaguar" will return some users a higher percentage of vehicles while others will see more felines — with straight correspondence to their interests and activity, as they are mapped by Google. Of course PageRank and profiling affect only that limited percentage of Surface Web Google is actually capable of reaching, while the rest — the so called Deep Web — remains unreachable for the Google user.

But there is more. It is widely known and discussed that the vast majority of users does not bother checking beyond the first two or three results they receive, and only a scarce number will proceed checking beyond the first page. But what if a certain user is particularly determined on discovering the meaning of a word, and will try to read all possible contexts? Surely then will these limitations become less pivotal — well, not quite. Google will only allow a user to view up to 100 results for a page over a maximum of 100 pages — i.e., the top 1000 rated URL's for his or her specific profile.[3] From this point of view the Google result count presented at the top of every page is somewhat of a deceit, since it is technically impossible for a user to access any result posted beyond the 1000 line. This potentially creates situations in which as many as nearly 100% of the word's contexts are de facto unavailable.[4] In order to access them, one must narrow his or her search using language/region settings, or adding a new search word — a constraint that presupposes the

[3]In their Search Protocol Reference, Google specifically mention the 1000 result limitation, both under Filtering and under Sorting. See https://developers.google.com/search-appliance/documentation/50/xml_reference#request_filtering and https://developers.google.com/search-appliance/documentation/50/xml_reference#request_sort.

[4] For example, if one searches for words such as "Obama" or "football", both of which return hundreds of millions of results.

user's acquaintance with the word, and undermines the idea of presenting the user the complete meaning of it — the "betrayal of [the] utopian dream of word liberation" mentioned by Groys (14), is thus extended beyond the negation of deconstruction, to the negation of the idea of new media democratization.

Over the last few years, and especially since the rise of the so-called Arab Spring in late 2010 — much of which success was assigned to the power of social media — the terms new media and democratization were used together to express a strive to change political regimes, yet, originally, new media brought along the hope for a democratization of information, mainly news.[5] While some still claim great success to this concept,[6] a CNN research from 2010 revealed that in terms of information monopoly, no great change has occurred — the main contributor to the majority of content online remains a minority of web users.[7]

Manovich considers an important feature of new media to be the fact that unlike the traditional creative work, in which the work and interface were identical and interchangeable notions, the database allows a single work to manifest throughout a plethora of interfaces (Manovich 199-201). For Manovich this is a crucial observation for artistic multimedia projects, which can be experienced by different users in different ways. In the Google case, this means that coming from the exact same set of data, every user receives his or her own custom-fit set of results. Mathematically speaking, due to the 1000 results limitation, the likelihood of two users having access to the

[5]Much was written on this topic, but see especially M. Raboy ("Media and Democratization in the Information Society").

[6]For example, in an interview from 2011, Nobel prize winner Steve Running discussed the divulgence of science news to the masses through new media tools, such as blogs and videos. See: http://www.pbs.org/mediashift/2011/12/nobel-prize-winner-on-how-new-media-is-democratizing-science-news340.html.

[7]http://www.cnnmediainfo.com/pdf/cnn_booklet_pownar.pdf.

exact same set of results, and especially on wide-range searches with hundreds of thousands, if not millions of theoretically attainable results, is close to zero. From this point of view, however, it is difficult to assign Google the power of undermining deconstructional freedom — if to every user a different set of answers — i.e., contexts, every user maturates his or her own understanding of the meaning of the word, which is inevitably slightly different than his or her fellow user's understanding of the same word.

Ostensibly, the Google idea of every user his or her meaning is a propagation or even an implementation of the second law of library science — every person his or her book — announced and discussed by Ranganathan (*The Five Laws of Library Science* 199-201); in point of fact, Google's execution is a great perfidy toward Ranganathan's idea of a personalized service. In *Reference Service* Ranganathan explains the implication of the second law on the service the reference librarian should provide the reader: the reference librarian, understanding the reader's personal interest, should help him or her find the adequate micro and macro documents (Ranganathan, *Reference Service* 54-55). For Ranganathan the interaction between the reference librarian and the reader may never be unilateral — the reference librarian is an attentive companion rather than an imposing guide. The Google Search service, on the other hand, is basing its proposal of consultable documents on nontransparent, uncontrollable and undiscussable parameters, which allow the user no room for intervention — while for Ranganathan the personalized choice of documents is to be conducted *in praesentia*, the Google effectuation of this process is done *in absentia*.

Claims in favor of the search system can, of course, be made. Firstly, Google *must* commit preselection to avoid overload (Guerrini, Bianchini, and Capaccioni 93). Secondly, Google *should* commit preselection in order to facilitate its user's work by providing him

or her with the content he or she was supposedly looking for. This second justification is based on Larry Page's — co-founder and CEO of Google — description of the "perfect search engine" as something that "understands exactly what you mean and gives you back exactly what you want."[8] Understanding what the other side means is a notion discussed by Wittgenstein as follows: One cannot explain to the other side what he himself understands; one can give examples — explanations, but the other side would always have to guess his or her drift. Out of the various interpretations that would seem plausible to the other side, he or she will then choose one — in that case he or she could ask: did you mean... (Wittgenstein N. 83, 2.10). If this phrase appears familiar to the reader, it is because up until not so long ago it has been the exact same wording Google Search was using to clarify its user's query.[9]

Let's recap. A user sits down in front of his or her computer and decides to look for a word on Google. He or she opens his or her browser, and navigates to one of Google's many interface pages. This first choice of interface will determine the number and type of results he or she will receive. The user then types his or her word in the search box, runs the search and receives a number of results — say 2,571. These results arrive in a certain order — the Google algorithm decides which ones are more or less pertinent to the user's interest — the user has no control over this part, and no way to offer his or her feedback. The user, unfamiliar with the object of search, now decides to study it carefully, going through the vast number of contexts his or her search returned, but alas, only 1000 of them are available. The user could try and change

[8] As quoted on the Google company products and services webpage: http://www.google.com/about/company/products.

[9] Google is gradually replacing this clarification feature with "Showing results for...", accompanied by a small print link to the originally searched set of characters, thus modifying its strategy from enquiring to assuming.

his search so that to see the other 60% of the results, yet the only way to do so is by knowing which other words could be found on those web pages — the user is unable to arrive at the full meaning of the word — of any word. This process, duplicated by millions of users, would result in each user having his or her own personal unique understanding of the meaning of the search word; some meanings may never come up in anyone's search. By showing the number of contexts to each word is finite, Google has turned deconstruction upside down, but by allowing the user to access only a limited and personalized set of results, Google has engendered a new type of deconstruction — normative meaning is stymied not by the unboundedness of possibilities, but by the impossibility of discovering the integral meaning of a word.

References

Groys, Boris. *Google: Words beyond Grammar*. Ostfildern: Hatje Cantz Verlag GmbH, 2012. (Cit. on pp. 255–259, 261). Print.

Guerrini, Mauro, Carlo Bianchini, and Andrea Capaccioni. *La biblioteca spiegata agli studenti universitari*. Milano: Bibliografica, 2012. (Cit. on pp. 259, 262). Print.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schutze. *Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2008. (Cit. on p. 256). Print.

Manovich, Lev. *The Language of New Media*. Cambridge: MIT Press, 2001. (Cit. on pp. 256, 261). Web. <http://www.manovich.net/LNM/index.html>.

Raboy, Marc. "Media and Democratization in the Information Society". *Communicating in the Information Society*. Ed. Bruce Girard and Seán Ó Siochrú. UNRISD, 2003. 101–120. (Cit. on p. 261). Web. <http://www.unrisd.org/unrisd/website/document.nsf/(httpAuxPages)/26BE21C65B15A339C1256E550056A85F?OpenDocument\&panel=additional>.

Ranganathan, Shiyali Ramamrita. *Reference Service*. Bombay: Asia Publishing House, 1961. (Cit. on p. 262). Print.

——.*The Five Laws of Library Science*. London: Edward Goldston, 1931. (Cit. on p. 262). Print.

Wittgenstein, Ludwig. *Philosophical Investigations*. Oxford: Basil Blackwell, 1958. (Cit. on pp. 258, 263). Print.

ANITA PAZ, student at the University of Florence.
paaz.anna@gmail.com

Paz, A. "In search of Meaning: The Written Word in the Age of Google". *JLIS.it*. Vol. 4, n. 2 (Luglio/July 2013): Art: #8798. DOI: 10.4403/jlis.it-8798. Web.

ABSTRACT: The article intends to be a short theoretical elaboration of the ideas raised by Boris Groys in *Google: Words beyond Grammar* from a LIS point of view. Through the use of critical tools such as Manovich's theory of the database and Wittgenstein's writings on meaning and context, the author delineates a double partial characteristic of the search conducted by Google in terms of quality (the Google algorithm is partial towards results claimed to appertain the user's interests) and in terms of quantity (the Google interface will only allow the user partial access to the search results). The author then re-reads Groy's claim of Google turning deconstruction upside down, suggesting the mere substitution of a classical Derridean deconstruction defined by the unboundedness of meaning possibilities, with a new deconstruction caused by the impossibility of discovering the integral meaning of a word.

KEYWORDS: Google; meaning; Groys; Manovich; Wittgenstein; deconstruction; partial; Google search

Submitted: 2013-02-27
Accepted: 2013-03-12
Published: 2013-07-01

JLIS.it

Collection development in the digital age

Klaus Kempf

At the beginning of modern librarianship, in the early modern period, at the time of the Kunst- and Wunderkammern, the so-called cabinets of curiosities, the common origin of libraries and museums, the library has still been equated with the term "collection" and vice versa. Back then it has already been an integral part of library collections that those were compiled with sustainability in mind and were aimed at the public, although this was restricted. The latter was the reason for the introduction of a rigid organization for the collection objects which turned the mere accumulation of books into an organized entity, thus facilitated their presentation and use, and foremost let them become an institution, the "library". It is important to be aware of those basic aspects of library collections, if one would like to understand, what changes have happened especially in the last two decades with the rise and triumph of digital information as well as with the invention of the Internet and with the libraries.

Today in the age of the so-called "hybrid library" that hosts in its collection printed resources as well as digital resources under on (real and virtual) roof, the libraries foremost have to develop a coherent service concept taking into account the media break. They have come under enormous pressure from the demand side as well

JLIS.it. Vol. 4, n. 2 (Luglio/July 2013).
DOI: 10.4403/jlis.it-8857

as from the supply side. On the one hand the users with their information behaviour have emancipated themselves from the library and its role as information intermediary. On the other hand the information and media market is, due to the Internet, predominated by a very strong competition. The newcomer on the market, the commercial information providers, the search engine operators such as Google and others, as well as a worldwide operating Internet bookstore such as Amazon are developing constantly new attractive offers. To fairly keep pace with this development, libraries have to make a change of course, better even a paradigm shift in their traditional service concept: They do not act foremost collection- or media-orientated anymore, but based on the concrete user needs they try to fulfil these at the best. In this framework the provision of the library holdings is no longer the priority, but only one of the possible service options. The own collection and the local collection development still do not have become obsolete, but the concept as well as the content of the collection have to be redefined. From now on the library collection also includes licensed resources, as well as resources free available on the Internet, so-called open access information resources. Cooperation plays an increasingly important role in the acquisition (keyword: consortia) as well as in collection development. Division of labour considering costs as well service factors is generally the guiding principle and today it is more important than ever before. The consequence however is, that the collections of libraries that are engaged in a consortium – and by now this is the majority of them – become more and more homogeneous, this means upon reversion that the collection as distinctive characteristic of a library, or even as the unique selling proposition will become less and less important. Next to the cooperation phenomenon the aspects of the presentation of holdings, of access to those holdings and the visualisation of the media offer – as for example the setup

of virtual subject libraries (ViFas) based on search engines, similar to Google etc. – have gained a greater significance than they had in a yesterdays world of only printed media.

So what will the future bring for libraries and their collections of tomorrow? Facing the ongoing rapidly progressing development of information and communication technology and the just emerging virtualization of the research infrastructure that will in turn strongly influence the information and communication behaviour of the individual researcher, the look into even the nearest future is complicated by significant elements of uncertainty especially for such a specific question. An attempt of a prudent prognosis will nevertheless be made: The future library world will consist of contrasts on the one hand and division of labour on the other hand. We will be witnesses, especially in the field of collection characteristics, of an almost merciless differentiation, or even selection of libraries regarding the institutional-specialist aspects. In the all-digital-world of tomorrow the topic collection development in the classic sense, this means the establishment and expansion as well as the maintenance of collections as comprehensive as possible, more and more compromised of mostly genuine online information resources, some of them possibly free accessible over the Internet as open access publications, will play a significant role only for a few, selected, large-scale and high-performance libraries. Those will nevertheless cooperate worldwide and especially cross-divisional, this means together with other memory institutions, such as (also selected) archives and/or museums, and will work on for example specialist collections of relevant digital objects. The vast majority of scientific libraries in contrast will restrict themselves to more modest collections than they have today and will adjust this local "online core collection" strictly to the actual basic needs of their users. For unusual information needs the "comprehensive collections" described above will come

into play as a *lender of last resort*. Also smaller libraries however can play an active part in a completely networked world through locally produced (digital) document and object collections that possibly have unique character and are preserved in the respective institutional and/or local subject repositories, and thus contribute their specific addition to the forming, geographically and institutionally distributed "digital world memory".

A reorientation of the collection policies could last but not least be relevant in the framework of the establishment of so-called virtual research environments especially for data-intensive research projects. Here an extension or even a change of the concept of collection objects would take place. The main focus would not lie, as it usually has been, on documents of whatever kind, at best a publication or similar Internet resource, but their place is more and more taken by primary research data in previously unknown magnitude and complexity. Collecting data however is in this context only a small proportion of a conceivable, new librarian "service package", that, tailored to each particular case, will include the development and the use of metadata schemes, research documentation, (online-) publications published according to open access or open science concepts and last but not least the long-term preservation of the collection objects including the associated metadata and possible further secondary information. For libraries that define themselves as an integral part of the infrastructure of a modern scientific community or even better as functional partner of the researchers, this is a very demanding new challenge, but definitely worthwhile and strongly capable of development. This basic, almost revolutionary reorientation anyway is surely not a panacea for every library and certainly not a guarantee for the survival as institution.

We have thus (for now) come full circle. The view from the early modern period up to the information world of tomorrow shows,

how the understanding of objects collected in a library and with this the understanding of what a library is, can change, without abandoning certain fundamental principles or values. Concerning the librarian activity of collecting this means, that regardless of the collection object – yesterday the printed book, today the digital *content*, in any form whatsoever – the aspects of organization and visualization or presentation in the sense of "providing access" are always implied and remain the constants of the librarian work. This, incidentally, the libraries had in common with the cabinet of curiosities. Due to the Internet or, more precisely, the WWW as the worldwide information and communication platform, those aspects, which have literally been out of sight, this means disappeared from the library magazines, undergo a renaissance. The new digital collection objects – similar to the books in a baroque library hall – have to be made directly visible and should be accessible immediately. In this context once again, as already during the early days of modern librarianship, at the time of the cabinet of curiosities, a specific form of indexing, a kind of "spatial organization" of collection objects comes into play. This time though it is not the architecturally closed space, which is supposed to make the books, presented systematically and according to a specific, given organization, as visible as possible and thus accessible, but it is the infinite space of the Internet with its seemingly unlimited possibilities to "stage" information in a way, that the user can "climb through" to the needed or offered information as conveniently and as directly as possible. Google has set standards also in this context. The search engine giant has, in the opinion of some art and media experts, on the one hand created an icon of the 21. Century with its – commercial-free – front page only determined by the company name and the search field and one the other hand brought about a revival of the cabinet of curiosities in a new appearance:

> «Brilliant software programmers like the Google founders are the creators of todays digital wunderkammer. They create an endless number of possibilities for the global storage, networking and representation of knowledge» (Burda182).

Whether or not Google is recognised as the significant originator of the design of the new, the virtual world, it can be stated with Bredekamp, that the drawing on the idea of the cabinet of curiosities, this means the here included concepts of the organization and presentation of knowledge as well as their application to the flood of images and information caused by the digital media,

> «involves for the users or viewers a training of visual association and mental processes, that run ahead of the language systems, . . . (Bredekamp102)
> And (on the other hand) – as Anke Te Heesen states – doubtlessly . . . the adaptation of the principles of the cabinet of curiosities, also "tames" the chaos of objects, which our today's disciplinary thinking considers as separated, to an aesthetically appealing and calming order of things» (Heesen and Spary7-21).

Bredekamp once more:

> «The high-technology societies are going through a phase of a Copernican revolution from the dominance of language (and of text; K.K.) to the hegemony of the image» (Bredekamp102).

The libraries with their collections are an integral part of this newly forming virtual world. Their collections – from now on in digital form – continue to exist in the larger collections of the new "digital cabinet of curiosities". In this respect libraries are and will always remain also collections, no matter what institutional fate they will take.

References

Bredekamp, Horst. *Antikensehnsucht und Maschinenglauben. Die Geschichte der Kunstkammer und die Zukunft der Kunstgeschichte*. Berlin: Wagenbach, 1993. (Cit. on p. 272). Print.

Burda, Hubert. *The Digital Wunderkammer. 10 Chapters on the Iconic Turn*. Munchen: Wilhelm Fink Verlag, 2011. 182. (Cit. on p. 272). Print.

Heesen, Anke te and Emma C. Spary. *Sammeln als Wissen : das Sammeln und seine wissenschaftsgeschichtliche Bedeutung*. Gottingen: Wagenbach, 2001. 7–21. (Cit. on p. 272). Print.

KLAUS KEMPF, Bayerische Staatsbibliothek, München.
kempf@bsb-muenchen.de

Kempf, K. "Collection development in the digital age". *JLIS.it*. Vol. 4, n. 2 (Luglio/July 2013): Art: #8857. DOI: 10.4403/jlis.it-8857. Web.

ABSTRACT: The present essay, shows the collection development in the digital age and the changes occurred in recent decades with the rise of Internet and "hybrid library". Today, in digital age the libraries must be able to meet the real needs of users, working for a great collaboration and cooperation in building and managing online digitized collections, as well as taking care and resize their traditional concept of service.

KEYWORDS: Collections; Digital age; Libraries; Wunderkammern; Search engines.

Published: 2013-07-01

JLIS.it

Dipartimento SAGAS, Storia, Archeologia, Geografia, Arte e Spettacolo

con il supporto di: supported by:

Casalini libri

Le Lettere

La piattaforma ICT, lo sviluppo e la manutenzione dell'installazione di OJS che ospita JLIS.it sono forniti da:

ICT platform, developing and maintenance for the OJS installation hosting JLIS.it are provided by:

CINECA

Direttore Responsabile ai termini di legge: Nicola Cavalli

In attesa di iscrizione nel registro stampa del Tribunale di Milano.

Finito di stampare nel mese di febbraio 2014 da

Ledizioni
The Innovative LEDIpublishing Company

http://www.ledizioni.it