BIBLIO THECAE .it

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA
DIPARTIMENTO DI BENI CULTURALI

Roberto Raieli

Oltre i termini dell'information retrieval: information discovery e multimedia information retrieval

Prologo

Nell'idea del Mundaneum,[1] e nei progetti di Paul Otlet e Henri La Fontaine, la condivisione di informazioni e conoscenze, la possibilità di ricercare e di accedere largamente e democraticamente alle risorse culturali, la pace e la fratellanza, coincidono sotto vari punti di vista teorici e pratici.[2] Condividere veramente la conoscenza significa non mantenere alcuna riserva nei confronti degli 'altri', dato che il sapere è alla base dello sviluppo tanto culturale quanto economico, tanto sociale quanto politico, e di quasi ogni forma di progresso. La conoscenza condivisa può posizionare gli Stati sullo stesso piano, distribuendo tra tutti i progressi di ognuno, rendendo quantomeno inutili le lotte che ogni nazione sostiene per possedere gelosamente l'esclusiva di scoperte e brevetti, la proprietà di risultati della ricerca che supportano primati industriali e politici.

* Ultima consultazione siti web: 20 dicembre 2016.

1 Il sito dell'attuale organizzazione del Mundaneum e all'indirizzo: <http://www.mundaneum.org/>.

2 Per esempio: Otlet 1914; Otlet 1934; La Fontaine 1916.

Allo stesso modo, anche l'idea di un possibile criterio universale di classificazione e organizzazione, per favorire la gestione e la ricerca delle conoscenze, nonché l'accesso ad esse, è compatibile con la preparazione di un terreno di incontro culturale sulla base del quale costruire la pace nel mondo.[3] Nulla di nuovo, dunque, nell'idea odierna di un'organizzazione delle risorse della conoscenza che coinvolga, a livello mondiale, tutte le risorse presenti nel Web,[4] al fine sociale di condividere tutti i progressi del sapere[5] (almeno con tutte le persone del mondo che riescono, in qualche modo, ad avere un accesso efficace al Web stesso). Tale idea, del resto, non era nuova anche per Otlet, risalendo all'antica aspirazione dei 'bibliotecari' e dei 'bibliografi' di riepilogare e controllare l'universo delle conoscenze rappresentate nei diversi supporti tecnologici che ogni epoca mette a disposizione, e che nelle diverse epoche ha prodotto specifici tentativi di elencazioni o bibliografie generali e universali, pur iniziando a strutturarsi come una precisa tradizione di studi solo all'inzio del XVI secolo.[6]

La *Cité mondiale* concepita da Otlet, 'ampliamento' del Mundaneum, è ben rappresentata dall'attuale rete Internet, sviluppata (quasi) dappertutto e interconnessa.[7] Le critiche poste al tempo a quel progetto erano simili a quelle poste alla nascente teoria della documentazione: troppo idealismo e poca attuabilità, 'fondamenta' utopiche. Adesso la città mondiale può stare ovunque, non c'è un luogo del mondo preferibile, e in questo riconduce alla rete distribuita, a Internet,

---

[3] La prima classificazione 'universale', la *Classification décimale universelle* (CDU), fu pubblicata nel *Manuel du répertoire bibliographique universel*, Bruxelles, Institut international de bibliographie, 1905.

[4] Vedere: Berners-Lee - Hendler - Lassila 2001; Heath - Bizer 2011.

[5] Sullo sviluppo di questa 'utopia', vedere anche: Castellucci 2013; Petrucciani 2013.

[6] Vedere: Balsamo 1992, in particolare p. 24-48. Per un compendio sulle istanze contemporanee dell'*universal bibliographic control* (UBC): Solimine 1995.

[7] Già il Mundaneum è stato definito da *Le monde* «Google de papier»: «Le monde», 19 dicembre 2009.

al Web. Su questa linea, si può anche azzardare un paragone tra la 'missione' di Larry Page e Sergey Brin di «organizzare le informazioni a livello mondiale e renderle universalmente accessibili e utili»[8] e l'utopia sostanziale di Otlet e La Fontaine, così come si possono paragonare i due sistemi di gestione delle informazioni da loro messi a punto e le relative basi ideali.[9] Google, in ogni caso, ha voluto il Mundaneum tra i partner del Google Cultural Institute per i progetti di digitalizzazione del patrimonio universale, e riconosce a Otlet e La Fontaine di essere stati «pioneers of the World Wide Web and the Internet» grazie alla loro «visionarietà».[10]

Senza eccedere nel paragone tra imprese culturali di alto ingegno, come il Web e Google, che restano pur sempre imprese, e le reti organizzate della cultura 'senza scopo di lucro', che si possono però magnificare in Internet, si può sottolineare che la conoscenza è una realtà diffusa, e da diffondere, che si allarga a macchia d'olio nella comunità di riferimento – non più una città in senso urbanistico, ma propriamente una città universale, ideale – di ogni istituzione culturale. La conoscenza, del resto, non solo non si 'esaurisce' a opera dei pochi (privilegiati) che ne possono fare uso, ma addirittura si accresce con l'uso diffuso. Pare che questa cosa sia più chiara da quando è stata ribadita da Elinor Ostrom nel 2006[11] – per tacere di Robert Merton negli anni Settanta del Novecento[12] –, ma poteva essere chiara già da tempo, almeno dai tempi di Conrad Gesner.[13] Se muta l'idea di conoscenza, può ben mutare anche l'idea di 'templi' della conservazione e della ricerca della conoscenza, le biblioteche,

8 Celeberrima frase di Larry Page, riportata in un'intervista del 2001 (Page 2001), che campeggia nella pagina di presentazione dell'azienda Google: https://www.google.com/about/company/

9 Confronto tentato in: Wright 2014.

10 Vedere: <https://www.google.com/culturalinstitute/beta/partner/mundaneum>.

11 Hess - Ostrom 2006.

12 Merton 1973.

13 Serrai 2016.

che sono oramai orientate alla massima accessibilità (meno che mai 'sacerdotale') e a non considerarsi le uniche teche dell'informazione, affiancandosi ad archivi e musei – sicuramente –, ma anche a gallerie, orti botanici, zoo, siti culturali di varia natura eccetera.[14]

Se la conoscenza è opera 'di tutti', e la scienza è 'di tutti', le risorse dell'informazione e della conoscenza sono, dunque, di tutti e a portata di tutti devono essere anche i mezzi di ricerca e scoperta di tali risorse. I punti da sottolineare, quindi, sono due: rendere liberamente disponibili le risorse e fare che la loro scoperta sia a portata di ognuno.

La parte 'pratica' del processo, cioè la fattiva condivisione di informazioni e di risorse, può avvantaggiarsi sempre più dei progressi della *Information and communications technology* (ICT). Il valore delle attuali tecnologie per il processo democratico di diffusione della cultura e della conoscenza è stato riconfermato dalla *Lyon declaration*,[15] ed è sempre stato chiaro nei fondamenti delle '*BBB declarations*'[16] e di tutto il movimento dell'*open access* (OA).[17] La parte 'teorica', relativa alla creazione di conoscenze a partire dalle informazioni e dai dati raggiungibili da tutti, trae vero vantaggio dall'ampliamento

[14] Come definito nei vari progetti GLAM (Galleries, libraries, archives, museums), per esempio: Musei Archivi e Biblioteche (MAB), 2014, <http://www.mab-italia.org>; International Federation of Library Associations and institutions. Libraries, archives, museums, monuments & sites (LAMMS), 2014, <http://www.ifla.org/lamms>.

[15] *The Lyon declaration on access to information and development*, 2014, <http://www.lyondeclaration.org>.

[16] *Budapest open access initiative*: http://www.budapestopenaccessinitiative.org/; *Bethesda statement on open access publishing*: http://legacy.earlham.edu/~peters/fos/bethesda.htm; *Berlin declaration on open access to knowledge in the sciences and humanities*: https://openaccess.mpg.de/Berlin-Declaration.

[17] Riguardo la storia e i riferimenti all'insieme del grande movimento per l'*open access*, si rinvia a uno scritto acutamente riepilogativo delle tappe principali dell'OA e dotato di una ricca bibliografia (anche se non molto recente): Santoro 2011. Sempre a fini riepilogativi, inoltre: Cassella 2012 e 2015.

e ottimizzazione dei processi di ricerca, scoperta, selezione e raccolta delle risorse da cui partire. La 'democratizzazione' dei processi di ricerca e accesso, inoltre, non può essere secondaria, e coinvolge ancora una volta le tecnologie che rendono possibili in modo aperto e semplificato le ricerche, ma coinvolge anche la diffusione della competenza relativa all'impostazione e gestione delle ricerche e dei sistemi di ricerca (che rientra nell'ambito della *information literacy* e della *digital literacy*[18]).

Sottolineando le utopie sociali e democratiche che si innestano in un progredito, rinnovato e potenziato processo di ricerca e scoperta dell'informazione, si può notare che, oltre quello terminologico e quello semantico, un sistema di comunicazione realmente universale per raggiungere tale organizzazione complessiva è, da sempre, quello visivo,[19] con il cui linguaggio si può di fatto mettere tutto alla portata di tutti, indipendentemente dal tipo di risorsa cercata e soprattutto dalla persona che la cerca: indipendentemente dalla lingua, dalla cultura, dall''intelligenza' di ognuno – e ovviamente di là dalla 'classe' e dal 'censo'. Oltre al sistema di comunicazione visivo anche quello dei suoni, quello dei movimenti, o ogni altro sistema di tipo più largamente 'multimediale' ha ampie possibilità di comunicare universalmente e 'liberamente' contenuti, concetti, dati, informazioni, conoscenze…[20]

Oltre i termini dell'information retrieval (IR), dunque, del sistema di ricerca e accesso ai documenti maturato in circa sessant'anni di attività più o meno limitate a cerchie di 'specialisti', l'information discovery e il multimedia information retrieval (MIR) si propongono come metodi di scoperta delle risorse perfettamente compatibili con quelli specialistici, ma anche ampliati, democraticamente, a ogni altra

[18] Tra le tante cose in proposito: Gilster 1997; ACRL 2016. Per riassumere: Testoni 2014.

[19] Si possono citare in proposito: Panofsky 1955; Arnheim 1969; Gombrich 1982; Smith *et al*. 2005.

[20] Per ricapitolare: Svenonius 1994; Spink - Jansen 2006; Divakaran 2009.

categoria di persone che si avvicini alla cultura, aprendosi a spazi e a metodi di indagine nuovi, liberi e largamente accessibili.

Si delinea un ampio panorama, quindi, ricco di 'paesaggi' bibliografici, biblioteconomici, documentali, tecnici e applicativi, ma anche sociali, tecnologici, culturali e storici, composto da istanze appartenenti all'information retrieval, alla scienza dell'informazione, alla *communication technology*, agli studi sul Web, all'*information behavior*, alla *digital literacy*, all'*open access*… Entro i limiti di una trattazione solamente riepilogativa, allora, sarà possibile giusto nominare le diverse componenti dell'insieme, concentrandosi sui rapporti tra l'approccio di tipo descrittivo e semantico alla ricerca di informazione e l'approccio di tipo semiotico, o 'content-based'.

Evoluzione del sistema dell'information retrieval

Il modello del *semantic Web*, proposto già nel 1999 dall'inventore del Web 'tradizionale' Tim Berners-Lee,[21] con le sue ramificazioni di collegamenti navigabili semanticamente è sempre più spesso considerato un valido elemento di confronto per molti sistemi di organizzazione, condivisione, ricerca e scoperta dell'informazione e della conoscenza, nonché come esempio per la riorganizzazione dei sistemi stessi. Questo vale, quindi, anche per il 'sistema' dell'information retrieval in generale, per gli (ultimi?) OPAC, per i *web-scale discovery services* e i discovery tool, per le *digital libraries* di nuova generazione, per diversi database,[22] come per i portali culturali, e per tutti gli strumenti a questi legati, in biblioteche, archivi, musei, gallerie eccetera.[23]

---

21 Berners-Lee - Fischetti 1999.

22 Per tutto ciò si può vedere: Iacono 2014; Bianchini 2012; Raieli 2015; Biagetti 2014.

23 Vedere i riferimenti, qualche nota sopra, ai progetti GLAM. Vedere anche: Guerrini - Possemato 2015; Vivarelli 2016.

Si possono evidenziare alcuni parallelismi tra le tendenze del Web e dei sistemi per navigarlo e le nuove 'tendenze' di IR che si diffondono nella documentazione e nella biblioteconomia. Da diverse parti si indicano come obiettivi quelli di consentire maggiore partecipazione alle persone, di dare più visibilità, apertura e 'amichevolezza' agli istituti culturali, di connettere i sistemi e scambiare i dati, di collaborare nella costruzione di *dataset* sempre più ampi, di non porre specifiche differenze di trattamento per risorse diverse sia per tipologia e provenienza, sia per contenuto testuale, visivo, sonoro, audiovisivo o multimediale. Questi obiettivi sono avvertiti come urgenti tanto dagli operatori del Web quanto dagli operatori degli istituti culturali, così che non mancano studi e modelli per la loro realizzazione.[24]

Scopo del rinnovamento proposto tanto nei sistemi del Web quanto in quelli bibliotecari e documentali è rendere granulari, interoperabili e 'integrabili/integrati' i dati relativi alle risorse dell'informazione, come gli altri dati prodotti e diffusi nel Web semantico da altri organismi che trattano in genere la conoscenza. Questa nuova organizzazione dei dati punta principalmente alla loro 'apertura', alla creazione in formato 'atomico', cioè autosufficiente e riutilizzabile, di ogni singolo dato, oltre le regole e gli schemi classici dei record bibliografici e dei formati di metadati. Ogni dato deve essere definito attraverso modelli largamente condivisi, preparato per essere esposto nel Web, diffuso, riutilizzato, in modo da consentire l'integrazione e la riaggregazione in combinazioni dinamiche, secondo le necessità e il 'punto di vista' dei diversi utilizzatori finali, assemblando e sintetizzando i dati granulari riferiti alle diverse entità correlate a una risorsa.[25]

---

[24] Questo è lo spirito di RDA, di FRBR-LRM e di ICP 2016, per esempio, come del *Thesaurus* del *Nuovo soggettario*. In specifico, poi, si può vedere: LOC 2012; W3C 2011; Guerrini - Possemato 2015.

[25] Vedere, per esempio: Bianchini - Guerrini 2014. Cfr. anche: Bianchini 2010, p. 224-238.

Sembra possibile attuare tutto ciò condividendo le pratiche e i protocolli del World Wide Web Consortium (W3C),[26] e in particolare la tecnica (o la pratica) dei linked data, il sistema tecnologico che al momento si sta dimostrando il più funzionale per raggiungere tanto gli obiettivi del Web quanto quelli degli organismi più esplicitamente culturali.[27] I principi, i meccanismi e i protocolli dei linked data per la navigazione, e la 'ricerca/scoperta', nel *semantic Web* ispirano i *library linked data*,[28] trasformando l'informazione prodotta in formato «machine-readable» in un nuovo formato «machine-understandable».[29] Si può, così, mettere a punto una rete semantica per l'organizzazione delle risorse, 'interpretabile' anche dalle macchine, fatta di connessioni triple di entità e relazioni, tra svariate risorse descritte secondo il modello RDF[30] e identificate da URI,[31] con una sintassi XML composta di termini e marcatori.[32]

L'apertura e l'accessibilità dei dati, comunque, non si realizzano (ed esauriscono) solo con la creazione in formato granulare e interoperabile, ma è necessario anche abbracciare la filosofia sottostante le nuove metodologie aperte e 'democratiche' della metadatazione, e le ragioni

---

26 http://www.w3.org/

27 Per entrare nel mondo dei linked data: Berners-Lee - Hendler - Lassila 2001; Berners-Lee 2006; Heath - Bizer 2011; Guerrini - Possemato 2012; Iacono 2014.

28 Vedere: W3C 2011.

29 Definizione riportata in: Marchitelli - Frigimelica 2012, p. 23.

30 Resource description framework (RDF), 2014, <http://www.w3.org/RDF>.

31 Uniform resource identifier (URI). Per un'introduzione: *Uniform Resource Identifier*, in: *Wikipedia: the free encyclopedia.* <https://en.wikipedia.org/wiki/Uniform_Resource_Identifier>.

32 La tecnologia del Web semantico è fondata sull'utilizzo di URI, assegnati univocamente a un oggetto, che lo identificano esattamente distinguendolo da tutti gli altri, e che ne consentono la localizzazione. Il nuovo Web, tramite la rete dei linked data, si struttura organizzando in modo dinamico diverse relazioni di collegamento in forma di 'tripla' tra i dati individuati da tali URI. Le stringhe dei linked data sono scritte in linguaggio XML, in modo da essere pienamente interoperabili. Per una esemplificazione, vedere anche: Guerrini - Possemato 2012.

degli *open data* e dei linked open data (LOD), sviluppate in rapporto di interscambio con la società e il Web. Se vari generi di restrizioni alla circolazione e al riutilizzo dei dati limitano lo sviluppo della conoscenza e il progresso della società,[33] è necessario trovare un luogo di interconnessione tra *open science* e *open government* (una sorta di Mundaneum/Europeana/Google? una *Cité mondiale* della rete?), il quale può essere sostanziato dagli *open bibliographic data*, e realizzato da OPAC, database, *open archives*, archivi istituzionali, *knowledge organization systems* (KOS), e tutto un rinnovato apparato di information retrieval realmente aperto al mondo odierno.[34] Questa è la tendenza che attraversa l'intera società civile – la quale è sempre più spinta anche dalle leggi degli Stati a condividere i dati in massa e a renderli raggiungibili da tutti e interpretabili dalle macchine[35] – e non può non avere profondi effetti nel mondo delle biblioteche, come dei musei e degli archivi, che della società civile hanno sempre rappresentato un elemento costitutivo.

Il senso della necessaria evoluzione di quello che può essere definito il 'classico' sistema dell'information retrieval è già ben sviluppato:

---

[33] Contro tali limitazioni è da considerare l'impegno di iniziative e progetti quali: Open knowledge foundation: <https://okfn.org>; OpenAIRE: <https://www.openaire.eu>; Open data commons: <http://opendatacommons.org>; Open library project: <https://openlibrary.org>.

[34] Cfr. De Robbio 2012.

[35] La *Lyon declaration* è già stata citata. Da un punto di vita 'applicativo', il primo esempio è dato alla fine del 2009 dalla statunitense *Open government directive* dell'amministrazione Obama (anche se ora c'è da chiedersi che ne sarà di tutta la legislazione sull'*open government* sotto il despotismo qualunquista trumpiano). L'Unione Europea ripete subito tale iniziativa con la *Malmö Ministerial declaration on eGovernment*, e nel 2011, internazionalmente, viene lanciata la *Open government partnership* (OGP). L'Italia ha aderito all'iniziativa OGP nello stesso 2011 e, nel 2012 e nel 2014, ha presentato un primo e un secondo piano d'azione nazionale contenente le principali iniziative in materia di *open government* e di sfruttamento delle nuove tecnologie verso la trasparenza, la responsabilizzazione dei cittadini e la disponibilità delle risorse informative per il *welfare* e il progresso sociale.

sia i sistemi, sia gli operatori, sia le persone comuni mettono in atto già da tempo diverse 'nuove' metodologie di ricerca e di scoperta dell'informazione e delle risorse, utilizzando più di uno strumento, e nella sua forma tecnologicamente più evoluta. La tendenza è sovvertire le classiche gerarchie degli strumenti di ricerca, e il Web nella sua accezione più ampia può avere un ruolo primario, subito seguito dai discovery tool, e solo in seguito dalle rassicuranti interfacce degli OPAC e dei database proprietari.[36]

Non resta che attuare nel modo migliore il passaggio dalle strutture 'lineari' di ricerca alle strutture 'reticolari' di scoperta, ben oltre quanto è stato finora comunque innovato nell'ambito dell'IR. La filosofia del mondo dei dati aperti converge in varie iniziative per una consapevole rivalutazione dei principi del trattamento e della ricerca dell'informazione. La linea è quella di promuovere, al posto della tradizionale struttura lineare di descrizione di un insieme di dati, set reticolari di apertura dei dati singolarmente presi, che possano dare un'idea più ampia del contesto in cui una singola risorsa è inserita. Con FRBR-LRM,[37] RDA,[38] e i recenti sviluppi di BIBFRAME,[39] le biblioteche guardano sempre più alla possibilità di creare i dati, o di collegare dati già esistenti, secondo un nuovo metodo, e di gestirli con nuovi mezzi di trattamento e ricerca, che consentiranno di andare realmente incontro alle esigenze delle persone, fornendo strumenti più potenti e semanticamente organizzati.

---

36 Su questo, vedere: Luther - Kelly 2011; Richardson 2013.

37 World-wide review of the FRBR-Library Reference Model, a consolidation of the FRBR, FRAD and FRSAD conceptual models, 2016, <http://www.ifla.org/node/10280>.

38 Traduzione italiana di RDA (Resource Description and Access), 2016, <http://www.iccu.sbn.it/opencms/opencms/it/archivionovita/2016/novita_0007.html>.

39 Kroeger 2013; Guerrini 2014; Trombone 2015. Il sito del programma BIBFRAME è all'URL: <http://www.loc.gov/bibframe>.

Evoluzione dell'information discovery

Il sistema dell'information retrieval, dunque, ampliandosi verso la prospettiva e il senso di una più ampia scoperta di informazioni e risorse nel Web 'aperto', si spinge verso il più flessibile e 'vago' sistema dell'information discovery. L'evoluzione dei sistemi di ricerca, così, sembra voler seguire in profondità il senso di 'inquietudine', perenne indagine e utilità delle scoperte occasionali che caratterizza in ampia parte l'atteggiamento di ricerca contemporaneo, sia per i ricercatori 'di mestiere' sia per le persone comuni.[40]

I sistemi bibliografico-bibliotecari devono necessariamente evolversi per consentire una reale 'navigazione' nell'universo dei documenti e ben oltre, in un «docuverso» ampliato, non limitato ai 'documenti' ma comprensivo di molti altri generi di risorse.[41] Allo stesso modo, però, tale evoluzione deve impedire che l'universo di riferimento si perda all'infinito, essendo invece vagliato, e semanticamente orientato, con la competenza che le istituzioni bibliotecarie hanno da sempre sviluppato.[42]

I nuovi sistemi, in sostanza, devono consentire di trattare con eguale dignità tutte le tipologie di risorse che possono essere utili nella strada della scoperta, permettendo comunque di organizzarle e 'comprenderle' per quello che ognuna vale e può significare in rapporto al tema di ricerca, all'obiettivo conoscitivo, o allo scopo informativo di ogni persona. In molti programmi di diffusione dell'informazione in rete, come quello dell'*open access*, le metodologie

---

40 Su inquietudine, senso di incompiuto e *serendipity* si può vedere: Caproni 2007; Salarelli 2012, p. 41-43; Merton - Barber 2004.

41 Nel senso della ricerca contemporanea sembra essere limitativo anche l'universo di documenti, o «docuverso», descritto molto tempo fa da Ted Nelson (v. Nelson 1990). Vedere pure Castellucci 2009.

42 Il Web dei documenti, come già annunciato da Tim Berners-Lee (v. Berners-Lee - Fischetti 1999), si sviluppa nel Web dei dati (v. Berners-Lee - Hendler - Lassila 2001), e delle connessioni semantiche (v. LOC 2012).

di ricerca fondate sull'idea classica di documento appaiono un retaggio del passato, contrarie al progresso della conoscenza. Ad avere posizione sempre più centrale, infatti, è il concetto di dato, estraibile non solo dai documenti 'canonici', ma da tutta una serie di altre risorse che crescono di importanza, e non possono non essere oggetto della ricerca.[43] L'attenzione deve essere posta, per esempio, anche sui «nanodocumenti», o «nanopublications», quali annotazioni, commenti, *post*, *tweet*, che possono contenere informazioni preziose, molti dei quali da sempre in uso nella pratica della comunicazione scientifica, ma spesso marginalizzati. Le nuove tecnologie digitali e di rete possono ristabilire questi rapporti, diffondendo nuove risorse quali linee di ricerca, progetti, riflessioni, immagini, filmati e dati di ogni genere. Inoltre, un discorso speculare si può fare per i «megadocumenti», o «enhanced documents», risorse potenziate come «hub» in un sistema integrato di dati e servizi.[44]

Anche se i metodi di information discovery, basati sui *web-scale discovery services* o sui LOD, non hanno il 'rigore' dell'information retrieval cui hanno abituato le banche dati, né quello degli OPAC e altri strumenti di riconosciuta affidabilità, sono ormai indispensabili per condurre quantomeno una ricerca 'di base', di avvio, trasversale e multidisciplinare, in grado di scoprire e di essere continuamente rilanciata, accessibile a tutti e 'democratica'. Le banche dati, al momento, sembrano restare lo strumento migliore con cui fare ricerca specialistica, utilizzando metadati e campi dei record appositamente definiti, e soprattutto appositi linguaggi di indicizzazione,[45] ma se i nuovi strumenti perdono molte funzioni di interrogazione avanzate, rendendo il percorso di ricerca 'omogeneizzato' al livello più basso, la soluzione migliore non è negarli dinanzi alle metodologie specializzate

[43] Castellucci 2011.

[44] Vedere il progetto Open annotation di Herbert Van de Sompel: <http://www.openannotation.org>.

[45] In proposito, vedere: Ketterman - Inman 2014.

per i livelli più alti. A seconda delle necessità degli utenti, serve riflettere sull'effettivo valore degli strumenti in rapporto alle persone che li devono usare, ma anche rispetto ai contenuti cui si applicano.[46]

Se le persone necessitano di avere accesso a uno spazio di scoperta ampliato all'infinito, e le biblioteche non si possono più limitare a segnare i propri sicuri confini entro lo spazio identificato dalle raccolte propriamente dette, il sistema dell'information discovery può rappresentare il punto di convergenza tra le diverse esigenze, trovando uno spazio ampio ma sicuro, e seguendo le necessità di sviluppo in senso contemporaneo e democratico.[47]

La metodologia dell'information discovery, in ogni caso, non ha la strutturazione e l'affidabilità di quella dell'information retrieval. Per quanto la sua storia sia più breve, però, e gli studi in proposito meno specifici, i comportamenti e le attività tesi alla generica scoperta dell'informazione sono una realtà sempre esistita, e con gli sviluppi tecnologici attuali l'information discovery è in continua crescita come sistema 'preferibile' per le persone. Lo stesso concetto di information discovery resta, comunque, vago, non ben definito scientificamente, né approfondito teoricamente, anche se la ricerca intorno ai sistemi di scoperta *web-scale* è in fase di continuo sviluppo e si occupa di aspetti non solo applicativi.[48] A parte gli studi sulla *web search* – orientati a definire un altro ambito della ricerca e un'altra tipologia di utenti[49] –, è dalla solida tradizione degli studi sull'IR[50], nonché sull'*information*

---

46 Vedere, per esempio: Arnold 2009.

47 In proposito, vedere: Levine-Clark 2014; Frame 2004.

48 In proposito, vedere: Richardson 2013; Caplan 2012; Pagliero Popp - Dallis 2012. Cfr.: Thomsett-Scott - Reese 2012; Proper - Bruza 1999.

49 Non è possibile, qui, inquadrare l'argomento, ma per un orientamento si veda: Shroff 2013; Meghabghab - Kandel 2008; Spink - Zimmer 2008; nonché: Nielsen 2006.

50 Tra le varie trattazioni sull'ambito ampio dell'IR, vedere: Stock - Stock 2013; Manning - Raghavan - Schütze 2008; Lancaster 2003; Baeza-Yates - Ribeiro-Neto 2000; Doyle - Becker 1975.

seeking behavior[51], che possono venire una serie di riflessioni utili a meglio definire l'ambito dell'information discovery.

Sintetizzando, le maggiori differenze tra il flessibile ma vago metodo dell'information discovery e l'affidabile ma a volte rigido sistema dell'IR, si possono riferire a poche 'dispute':

- linguaggio naturale VS linguaggio di indicizzazione;
- *folksonomy* VS *ontology*;
- risorse disponibili VS risorse selezionate;
- spazio infinito del Web VS ambito database specifico.

Per ognuna di tali differenze, ovviamente, l'opposizione non è rigida, e su ognuna serve fare alcune specifiche. Il linguaggio naturale è largamente utilizzato anche nei sistemi di IR, e molti progressi sono stati fatti riguardo al *natural language processing*,[52] i sistemi di information discovery, dal canto loro, non rifiutano affatto l'uso dei termini di indicizzazione, se questi possono migliore o velocizzare la ricerca. I linguaggi di indicizzazione e i tesauri specialistici, però, restano lo strumento più adatto per svolgere ricerche precise e mirate in ambiti professionali e tra i professionisti di un dato settore – fermo restando che i sistemi di information discovery possono comunque aggiungere qualcosa nell'ampliare le ricerche che portano a pochi risultati.

Stessa cosa si può dire per il rapporto tra l'uso di ontologie e di folksonomie. Se i sistemi di information discovery fanno tesoro dei *tag* inseriti dagli utenti di un dato sistema, che possono presentare punti di vista particolari e aggiuntivi, o termini più semplici per segnalare determinati oggetti, non è per questo escluso l'uso dei termini definiti nelle ontologie.[53] Allo stesso modo, per i sistemi di IR si creano

---

[51] Argomento che necessiterebbe una trattazione separata, ma per un riferimento generale si vedano i tre volumi di scritti di Marcia Bates: Bates 2016; nonché: Bates 1989 e 1990.

[52] Kurdi 2016; Kapetanios - Tatar - Sacarea 2013.

[53] Per una panoramica, vedere: Bambini - Wakefield 2014.

sempre più spesso applicazioni in grado di rapportare alle risorse sia i termini di indicizzazione propri di un dato database sia i *tag* inseriti successivamente e progressivamente dalle persone che lo usano.[54]

Riguardo il rapporto tra l'ambito di applicazione dei due sistemi, relativo all'ampia gamma delle risorse disponibili sul Web per l'information discovery e all'insieme ristretto delle risorse appositamente selezionate per l'IR, forse la differenziazione è più netta. Non è vero, però, che i sistemi di information discovery effettuano la loro ricerca sull'intero Web, anche se ne hanno la potenzialità.[55] I motori di ricerca generalisti del Web hanno un ruolo anche nei metodi di information discovery, ma non è affatto tra i principali, e molte sono le indicazioni di fare attenzione alle lunghe liste non 'referenziate' che presentano. L'information discovery più 'strutturato' si attua principalmente tramite i *web-scale discovery services*, e non ha a che fare più di tanto con la *web search* realizzata dai motori di ricerca più diffusi, di essa riprende solo la capacità di operare tra oggetti molto diversi, che comunque sono inseriti dentro l'indice unico creato – più o meno a misura – per l'istituzione che lo usa. In ogni caso, i sistemi di IR operano solo entro uno specifico database, o pochi database in rapporto tra loro, con l'interfaccia specializzata dedicata a questi, e quindi il loro ambito di ricerca è predefinito e limitato secondo le politiche di costruzione del database. I sistemi di information discovery, invece, possono operare in un ambito ampliabile a qualunque tipo di risorsa, in questo simile al Web aperto.

Riguardo la differenza/non differenza tra lo spazio potenzialmente infinito su cui opera l'information discovery e l'ambito definito di un database specifico per l'IR, valgono alcune delle considerazioni già esposte. In un'ottica futuribile, si può aggiungere che se i linked data, la condivisione, l'interoperabilità e la riutilizzazione dei dati diventeranno la nuova realtà del mondo dell'informazione e della

[54] Vedere pure: Matthews *et al.* 2010; Yang 2012.
[55] Cfr. Fagan 2012.

conoscenza, allora il Web potrebbe rappresentare lo spazio in cui i motori di ricerca sono liberi di effettuare la loro indagine 'sconfinata', e gli spazi delle biblioteche rappresenterebbero quelli in cui specifici strumenti quali i discovery tool possono sviluppare la ricerca entro 'confini' disegnati nel disegnare una data biblioteca e la sua specifica *mission* – per quanto non facili da definire o teorizzare, per quanto ineffabili nell'essenza logica e tecnica. I sistemi di ricerca non generalisti, dunque, 'lavorerebbero' in un universo di risorse più controllato e affidabile, consentirebbero una ricerca e una scoperta ampie ma 'sicure', oltre che strutturate. Allargare i confini di questo universo, poi, sarebbe semplice, e nulla di nuovo: i sistemi dovrebbero applicarsi ai database di biblioteche 'alleate', e non solo tra loro, ma anche con altre istituzioni culturali, riguardo a un dato ambito disciplinare o appartenenti a una tipologia funzionale, a livello nazionale o internazionale. Da questo punto di vista, l'opposizione tra information retrieval e information discovery potrebbe risultare solo apparente: lo spazio infinito perderebbe le caratteristiche di 'rischio', e l'ambito specifico non sarebbe più caratterizzato dalla 'limitatezza'. Si tratta di bilanciare accuratamente, da parte dei 'professionisti dell'informazione', il tipo di strumenti di ricerca e di recupero dei dati e delle risorse – libere o non libere che siano – da fornire o da consigliare alle diverse persone. È necessario, poi, che si completi la rivoluzione tecnologico-culturale già in azione con i linked open data…

Evoluzione di principio dall'information retrieval/discovery al MIR

Nell'insieme dello sviluppo esposto, i sistemi di analisi, indicizzazione e ricerca dell'informazione, tanto per quanto riguarda l'IR quanto per l'information discovery, sono rimasti sempre legati alla propria originaria natura terminologica e semantica. Termini di descrizione/soggettazione/indicizzazione e linked data sono rivolti

maggiormente al significato, piuttosto che al 'contenuto concreto',[56] dell'informazione. Sia una formulazione di soggetto, sia un termine o un codice, sia una stringa RDF-XML, puntano comunque a catturare il significato dell'informazione, dei dati, che indicano. Tali metadati, al più, tentano di descrivere/surrogare l'*aboutness*,[57] o anche il 'senso', di una risorsa, ma non possono manifestare alcunché riguardo l'effettivo contenuto, la concretezza, dell'oggetto a cui si riferiscono. Eppure, tenendo opportunamente in cosiderazione ogni tipo di risorsa e non solo quelle principalmente testuali, nell'applicazione ai generali contesti di ricerca prima indicati, propri delle persone in genere, è più efficace e funzionale fornire un'indicazione diretta sulla forma visiva o sonora di una certa risorsa, piuttosto che una 'indiretta' riguardo quello che essa può significare per chi si preoccupa di descriverla o rappresentarla.

È necessario, allora, capire quali metodologie possono meglio assistere nella ricerca e nella scoperta, ma anche nella conoscenza, riguardo soprattutto le risorse di contenuto visivo, sonoro, audiovisivo o multimediale in senso lato – prodotto di diverse maniere e linguaggi, di differenti forme specifiche di comunicazione –, il cui significato è spesso un elemento molto relativo, mentre il contenuto concreto è lo stesso per chiunque ne abbia un'esperienza appunto sonora, visiva, o audiovisiva. Per il trattamento efficace delle risorse multimediali è necessario considerare uno sviluppo di principio che porti dalle 'normali' metodologie di trattamento basate sui termini verso nuove metodologie, alla cui base sono le (oramai non più nuove) tecnologie

[56] Riguardo l'argomento trattato, la definizione di 'contenuto', sulla base di quella del termine inglese *content* internazionalmente utilizzato, fa sempre riferimento alla concretezza e materialità di un oggetto, al contenuto di parole, forme, colori, suoni, empiricamente intesi – o meglio, alla trasposizione digitale di questi –, e non vuol dire concetto o significato.

[57] Su questo termine dal significato ampio – traducibile in italiano con 'circalità' – vedere: Fairthorne 1969; Hutchins 1978; Hjørland 2001.

di trattamento che si fondano sul contenuto effettivo delle risorse, e non più solo sulla loro descrizione terminologica o categorizzazione semantica.

Riguardo l'adeguatezza degli attuali metodi di gestione dei sempre più complessi insiemi di risorse multimediali, l'esigenza che oramai tutti i sistemi dovrebbero avvertire è quella di passare dal metodo dell'information retrieval/discovery – in quanto criterio di ricerca e scoperta per termini e semantica di risorse di tipo testuale (applicato anche a risorse visive, sonore, audiovisive) – a un metodo più ampio, quale quello del multimedia information retrieval – che ricomprende il primo criterio sviluppandosi come sistema di ricerca e scoperta tramite testi, immagini, suoni, semantica, per risorse di tipo testuale, visivo, sonoro, audiovisivo, multimediale…

La realtà è, invece, che ancora si rivela una contraddizione nella logica con cui i metodi e i sistemi di trattamento dell'informazione continuano a essere organizzati. Infatti, se nei casi in cui si tenta di interrogare fonti di documenti testuali con mezzi non testuali una certa confusione di 'linguaggi' di interrogazione si può sicuramente considerare paradossale, nei casi in cui s'interrogano tramite testo fonti di risorse visive o sonore non si può continuare a considerare opportuno lo scambio di linguaggi opposto, senza tenere in conto la diversa natura di ogni risorsa. Se non è possibile ricercare e recuperare un documento scritto tramite mezzi di linguaggio visivi o sonori, allo stesso modo non si deve continuare a ritenere (con pochi dubbi) un metodo efficace recuperare risorse consistenti in suoni o figure attraverso l'uso di soli testi descrittivi, o schemi semantici, i quali spesso non riescono nemmeno a elencare le molte e diverse particolarità contenutistiche di ogni risorsa.

Nella situazione culturale e tecnologica descritta, dunque, dovrebbero apparire evidenti i diversi limiti del continuare a operare nella logica e nei termini di un generico information retrieval. Nella pratica tradizionale dell'IR, infatti, ogni tipo di ricerca di informazione e conoscenza è riportato alle condizioni di un'indagine tramite

linguaggio testuale, è necessario, invece, definire più ampi criteri di MIR, dove ogni genere di risorsa digitale possa essere trattata e ricercata tramite gli elementi di linguaggio, o di 'metalinguaggio', più adatti alla sua natura propria. Tutto ciò vale anche in vista della messa a punto e diffusione di un metodo meno 'lambiccato' ed 'esclusivo' per la ricerca di risorse che comunicano 'sensibilmente': più attraverso le forme, i suoni, i movimenti, che con linguaggi descrittivi o semantici quale quello testuale – più o meno formalizzato. Piuttosto (per esempio), dovrebbe apparire dispersivo cercare una fotografia di paesaggio – un tramonto sul mare – tramite una complicata descrizione a parole delle linee e tonalità desiderate, anziché sottoporre a un apposito sistema di ricerca un campione delle linee e delle tinte stesse.

*Teorie dell'*information searching

Si possono discutere alcune linee che identificano le possibili teorie dell'*information searching*, premettendo che tali teorie non hanno uno 'statuto' chiaro che le differenzia l'una dall'altra e, se anche si susseguono cronologicamente, nella sostanza tendono a essere evoluzioni una dell'altra, con varie sovrapposizioni sincroniche, per le quali l'una è compresa nell'altra.

L'information retrieval, per esempio, sarebbe l'ambito teorico-applicativo più antico, al suo interno dovrebbe essersi sviluppato il multimedia information retrieval, e l'insieme delle teorie e metodologie relative dovrebbe ora aprirsi verso criteri ancora più ampi di information discovery. Le cose non sono così conseguenti, in realtà, e i passaggi non sono così netti. Il MIR, infatti, è spesso passato inosservato nelle rivoluzioni o 'riaggiustamenti' teorico-pratici delle metodologie della ricerca di informazione – che spesso si 'aggrovigliano' su se stesse per tentare di risolvere i problemi della ricerca di materiali non testuali senza intuire una prospettiva di 'rivoluzione'[58] –, e l'information

[58] Per un discussione più ampia di questa necessaria 'rivoluzione', vedere: Kovács 2014; Raieli 2010, p. 86-98.

discovery ha di fatto iniziato a 'convivere' con l'IR senza che prima questo si fosse aperto verso il MIR – e senza porsi problematicamente nei confronti dell'IR, sviluppando metodologie 'parallele' quando non contrastanti. Allo stesso modo, le forme di information discovery, per quanto allargate a tutti sistemi e i metodi di ricerca, non hanno mostrato quella specifica attenzione ai contenuti delle risorse multimediali che le avrebbero portate all'integrazione dei metodi del MIR.

Comunque, si può velocemente schematizzare lo sviluppo dei sistemi di ricerca, allo scopo di vederli a 'volo d'uccello' prima di discutere i loro livelli di coesistenza e la metodologia semiotica del multimedia information retrieval.

Information retrieval

La storia dell'information retrieval parte dagli studi sulla classificazione e l'indicizzazione, e ha anche per questo una lunga e solida tradizione.[59] La tradizione si può datare dal 1895 circa, anno della fondazione dell'Institut International de Bibliographie da parte di Paul Otlet e Henri La Fontaine,[60] e passa, intorno agli anni Quaranta del secolo scorso, attraverso le ricerche di Vannevar Bush, che con il suo Memex disegna uno dei primi modelli per il trattamento automatizzato dei documenti e dei dati informativi.[61]

In ogni caso, è solo nel corso degli anni Cinquanta e Sessanta del Novecento, quando gli studi sul trattamento dei dati, dell'informazione e dei documenti si uniscono con le tecnologie degli elaboratori elettronici (Uniterm system,[62] Cranfield experiments I e II[63]), che si

[59] Tra le molte opere appartenenti a questa tradizione di studi sono già state ricordate: Stock - Stock 2013; Manning - Raghavan - Schütze 2008; Lancaster 2003; Baeza-Yates - Ribeiro-Neto 2000; Doyle - Becker 1975.

[60] Vedere: Mundaneum 1995.

[61] Bush 1945.

[62] Taube 1955; Taube - Wooster 1958.

[63] Cleverdon - Mills - Keen 1966.

può parlare compiutamente di information retrieval. In questi anni sono messi a punto i primi sistemi computerizzati di trattamento e ricerca delle informazioni, basati su algoritmi di elaborazione statistica dei rapporti della distribuzione dei termini presenti nei documenti, nei loro descrittori e nel database che li contiene.[64]

Gli sviluppi della ICT hanno, da allora, sempre guidato gli sviluppi dell'information retrieval, tanto applicato ai database quanto ai sistemi che si rivolgono all'intero Web. Nel corso di questa evoluzione, comunque, i sistemi hanno sviluppato diverse tendenze teoriche e applicative, da quelle basate sull'elaborazione del linguaggio naturale a quelle *user-oriented* e cognitive, da quelle multilinguistiche a quelle multimediali, ampliando il proprio campo di ricerca anche verso la semantica o la psicologia.[65]

È importante sottolineare che, nel corso di tale evoluzione, i sistemi di IR si sono allontanati dalla prospettiva inizialmente definita dell'*exact-match*, principio della corrispondenza esatta da raggiungere tra termini di ricerca e termini descrittori per reperire un documento considerato utile. Lo sviluppo dei principi di *query* e *retrieval* si è orientato verso prospettive di *best-match* più dinamiche e aggiornate. I nuovi sistemi di IR, dunque, seguendo una linea di ricerca e scoperta più libera, data la grande quantità di informazioni disponibili anche a livello *web-scale*, tendono a non 'ripulire' in modo troppo radicale i risultati della ricerca, per presentare, ordinate per 'rilevanza',[66] una serie più ampia di risorse che può suggerire nuovi percorsi di ricerca, anche a discapito della 'precisione' della lista dei risultati.

---

64 Per una sintesi delle problematiche tecnologiche dell'IR, cfr.: Marchitelli - Frigimelica 2012, p. 5-12. Vedere anche: Salarelli 2012 (in particolare l'appendice sull'IR alle p. 101-117).

65 Cfr. Järvelin 2003.

66 Riguardo le criticità dei sistemi di ordinamento in base alla «presunta rilevanza» per l'utente, vedere: Biagetti 2010.

Indubbiamente, questa nuova logica della scoperta evidenzia diversi punti di connessione tra i sistemi di IR specialistici, applicati a database specifici, e i più ampi sistemi di information discovery che, oltre a essere applicati ad ambiti e risorse facilmente controllabili entro sistemi di descrizione o indicizzazione, possono essere applicati a ogni genere di risorsa più largamente disponibile, pronti a operare ricerche meglio definibili nello spazio in vario modo organizzato del Web semantico.

Multimedia information retrieval

Il MIR ha le sue radici nei primi esperimenti di *content-based image retrieval* (CBIR) e di *similarity search*,[67] collegati agli studi sulla *computer vision* e l'*image processing* degli anni Novanta del secolo scorso, anche se spesso aventi scopi diversi rispetto alla ricerca di informazioni. Dal punto di vista documentale, i sistemi sviluppati in quest'area sono presentati e discussi in un saggio di Peter Enser che, ampliandone i principi all'ambito delle risorse culturali in generale, fa il punto su una serie di questioni pratiche e teoriche associabili al «pictorial information retrieval».[68] Lo studioso, tra i primi, sottolinea che per la maggior parte i database di immagini, architettati secondo le strutture tipiche dell'IR, sono pensati «traducendo» in termini i contenuti visivi e le relative chiavi di accesso. Le stringhe di *query* per l'interrogazione di tali database, di conseguenza, devono essere espresse terminologicamente, e possono puntare esclusivamente al *match* con i «surrogati testuali» dei documenti visivi: keyword, termini di indicizzazione, soggetti, titoli, didascalie. In ogni caso, l'indicizzazione di tutti i termini utili per la descrizione di contenuti visivi non può mai essere esaustiva, e spesso le qualità visive degli oggetti rappresentati non rientrano in alcuna categoria linguistica, essendo inesprimibili e inclassificabili.

[67] Vedere: Kato 1992; Del Bimbo 1999.
[68] Enser 1995.

Alla fine degli anni Novanta, lo sviluppo dei documenti audiovisivi ha catalizzato molta attenzione e prodotto ulteriori progressi nel trattamento dei documenti visivi contenenti anche movimenti, suoni e parlato, orientando lo sviluppo degli studi sui sistemi di IR verso la considerazione delle più complesse tipologie dei documenti multimediali. Uno studio di Frederick Lancaster ricapitola il tema in modo complessivo, definiendo le possibilità e i limiti del sistema dell'IR, e rivedendo il suo sviluppo all'interno della struttura 'term-based' fino al raggiungimento del limite oltre il quale non si può che passare a nuove strutture 'content-based'.[69] Si ipotizza, quindi, la necessità di passare verso sistemi di ricerca «ibridi», nei quali l'utente potrà usare tutti i mezzi che gli sono utili per pianificare le *query* multimediali, per immagini fisse o documenti audio o video, anche senza conoscere alcun vocabolario di ricerca, pure rivolgendosi allo spazio del Web.

Agli inizi del Ventunesimo secolo sono state investigate questioni ancora più specifiche, dato lo sviluppo di potenti algoritmi capaci di calcolare un elevato numero di variabili, essenziali per il *processing* e l''interpretazione' di documenti sempre più complessi come quelli multimediali. Si tratta infatti di mettere a punto paradigmi di analisi e ricerca delle risorse in grado di rapportare le rappresentazioni automatiche e 'oggettive' fatte dalle macchine con le sofisticate analisi intellettuali e 'sensibili' realizzate dagli esseri umani.[70] A questo si collegano specifiche analisi sull'utilità e l'apprezzamento di tali sistemi dal punto di vista delle persone che li dovrebbero usare, per testare e migliorare il grado di rispondenza agli specifici bisogni del ricercatore di informazione e risorse.[71] Sempre in questa prospettiva si sviluppa

---

69 Lancaster 2003.

70 In proposito: Deb 2004; Adami *et al.* 2012; Gast *et al.* 2013.

71 Riguardo i test, si segnalano le attività di TREC Video Retrieval Evaluation: <http://www-nlpir.nist.gov/projects/trecvid>. Inoltre, vedere: Hanjalic 2012; Thomee - Lew 2012.

la problematica di quello che viene definito il *semantic gap* dei sistemi di MIR, cioè la limitata efficacia 'semantica' dei sistemi content-based, basati principalmente sull'analisi e la ricerca 'semiotica' delle risorse, che tende a far passare in secondo piano il livello del significato. L'approccio semantico, di conseguenza, non deve mai essere escluso nei sistemi di MIR, che devono possedere gli strumenti opportuni per la definizione contenutistica e concettuale della strategia di ricerca.[72]

Di là dalle sperimentazioni accademiche, o dagli usi specifici in ambiti ristretti, il successo di questi sistemi è affidato, al momento, ad alcune loro applicazioni commerciali. Rilevante è il sistema Google Goggles, applicazione per smartphone sviluppata dai Google labs da circa un decennio, che permette a chiunque di produrre una *query* content-based fotografando o filmando un oggetto o un luogo, ottenendo una pagina Google con i risultati di ricerca.[73] Anche la comune interfaccia di ricerca di Google ha sviluppato un'applicazione in grado di effettuare ricerche visive, migliorando un sistema sperimentato già nel 2000, che caricando dal PC o da Internet una figura modello effettua poi una *similarity search* nel Web.[74] Ma il vero successo commerciale appartiene a due applicazioni di 'audio retrieval', SoundHound[75] e Shazam,[76] che negli anni hanno perfezionato le proprie caratteristiche di ricerca content-based applicate ai cellulari e agli smartphone, catalizzando l'ampio interesse del mondo della musica per sistemi di accesso immediati e innovativi.

---

72 Enser 2008. Utile, inoltre, il sito del Semantic Media Network: <http://semanticmedia.org.uk>.

73 Google Goggles: <https://support.google.com/websearch/answer/166331?hl=it&ref_topic=25275>.

74 La pagina dell'*image search* è accessibile dalla normale pagina di ricerca di Google, o direttamnte da: <https://www.google.it/imghp?hl=it&tab=wi&ei=MM3-VKDlCYn4Uqmug7AP&ved=0CBYQqi4oAg>.

75 SoundHound: <http://www.soundhound.com>.

76 Shazam: <http://www.shazam.com>.

Information discovery

L'information retrieval gode di una lunga e ininterrotta tradizione di studi, e il multimedia information retrieval di una tradizione intesa (se non lunga), ma la letteratura più o meno recente non ha mancato di investigare una serie di problematiche relative all'information discovery, ai suoi possibili obiettivi, i limiti e l'efficacia. Le trattazioni disponibili hanno sviluppato un'ottica spesso applicativa, a volte programmatica, ma sono presenti alcune aperture di tipo teorico.[77]

Non si può comunque affermare che l'information discovery sia consapevolmente definito come un settore di studi, o che abbia un perimetro identificabile, e molto di ciò in cui consiste riguarda spesso gli studi sulla ricerca in Internet e il Web generalmente intesi. Alla fine degli anni Novanta ci sono, però, stati precisi tentativi di identificare e teorizzare l'information discovery, e non solo in confronto all'IR, ma definendo per esso principi e metodologia autonomi e conseguenti a un preciso paradigma logico.

Questi tentativi possono essere rappresentati dallo studio di Henderik Proper e Peter Bruza, che sono tra i primi a indicare le basi della problematica, coscienti che non era ancora presente alcuna teoria e proponendo di inquadrare l'information discovery quantomeno nei suoi rapporti con la tradizione dell'IR.[78] Gli studiosi partono dalla constatazione del grande aumento della quantità di risorse digitali presente in rete, della loro varia tipologia e dell'importanza assodata e crescente che esse hanno anche nella ricerca scientifica. Tali risorse sono, in modo ampio, definite come «information carriers» – supporti, vettori, veicoli di informazione – e ognuna di esse può avere un tale e specifico valore informativo che il valore complessivo della rete si fonda sempre più sulla quantità e diversità, nonché sull'accessibilità, di questo dinamico complesso di veicoli di informazione: pagine

[77] Cfr.: Thomsett-Scott - Reese 2012. In proposito, anche: Levine-Clark 2014; Frame 2004.

[78] Proper - Bruza 1999.

web, *newsgroup*, *mailing-lists*, database, archivi, portali eccetera, senza dimenticare, ovviamente, i documenti 'tradizionali'. Al di fuori dei parametri dell'IR, è dunque necessario mettere a punto validi strumenti e validi paradigmi metodologici per consentire al ricercatore, più o meno 'specialista', di trovare e identificare tali risorse, senza disperdersi nel profluvio dell'offerta o imbattersi in oggetti di scarsa qualità e scarso valore informativo e conoscitivo.

Sono, dunque, l'ambito di applicazione e il metodo per operare in questo a rappresentare la differenza fondamentale tra i due sistemi. L'IR è incentrato sulla ricerca di documenti rilevanti all'interno di collezioni stabilite e principalmente testuali, inoltre gli utenti devono avere una consapevolezza delle proprie necessità informative e una certa conoscenza dei linguaggi e degli stumenti di indicizzazione e ricerca. L'information discovery, invece, si apre a uno spazio ampio e condiviso, l'insieme delle risorse è in continuo sviluppo e non è rappresentato solo da fonti testuali, si può così effettuare la semplice ricerca di un oggetto digitale già noto come sviluppare un'intera strategia di scoperta tesa all'identificazione e all'organizzazione di una partizione dell'universo delle risorse potenzialmente utili.[79] La novità più importante – destinata ad avere ampi sviluppi tutt'oggi – è che l'intero modello del processo di information discovery si differenzia da quello dell'IR, in genere più lineare, per essere pienamente «user centered» e articolato in modo reticolare, nonché basato sulle casualità imprevedibili del processo di scoperta dei dati informativi.[80]

Le più recenti variazioni d'uso nella terminologia della Library and information science (LIS) rivelano come l'idea di information discovery sia più chiara solo all'inizio degli anni Dieci del nostro secolo, e solo quando correlata alle applicazioni dei *web-scale discovery services* e dei discovery tool. L'uso della stessa espressione 'discovery tool'

[79] Lynch 1995.
[80] Cfr.: Proper - Bruza 1999, p. 740-749.

diventa un dato significato solo intorno al 2009, come si può appurare intuitivamente verificando il suo utilizzo tanto nei siti web e nei blog tematici quanto negli articoli e nelle pubblicazioni specifiche.[81] Negli anni precedenti l'espressione non aveva un significato condiviso, e i termini 'discovery' e 'tool' non apparivano insieme, quanto piuttosto erano usate espressioni come 'discovery environment' o 'discovery service'.[82] 'Web-scale discovery service', addirittura, è una stringa usata in modo completo e significativo in anni ancora più recenti, e solo in alcuni specifici saggi.[83] L'espressione 'information discovery', infine, è usata in trattazioni di tipo diverso, per intendere in modo ampio sistemi e metodologie che vanno da quelli dell'information retrieval fino alla generica navigazione in Internet. Tale espressione più o meno generale, comunque, è adesso sempre più usata per indicare l'idea delle nuove possibilità di scoperta dell'informazione e delle conoscenze collegate ai nuovi servizi dei discovery tool, anche se la terminologia relativa ai nuovi strumenti e alle trasformazioni da essi portate è in continua fase di definizione.

Forme dell'information discovery

Se si può dare un'illustrazione diacronica delle linee che identificano le teorie dell'*information searching*, quando si tratta delle reali, attuali e cangianti forme dell'information discovery si devono considerare sincronicamente tutti i raggiungimenti delle diverse metodologie coinvolte, e il termine 'information discovery' può anche diventare il termine generale nel quale convergono le definizioni delle altre forme di ricerca, IR e MIR, oltre a essere il nome che dovrebbe rappresentare il sistema più evoluto diacronicamente.

---

81 In particolare: Breeding 2010; Weare - Toms - Breeding 2011.

82 Per una compiuta ricognizione sull'uso di questa terminologia, vedere: Caplan 2012.

83 Vedere, per esempio: Hoeppner 2012; Ellero 2013; Richardson 2013.

Nella realtà attuale dell'information discovery, dunque, coesistono:

- una tipologia di ricerca terminologica, sviluppata tramite un linguaggio term-based: tale ricerca è fondata su un'analisi di tipo linguistico delle risorse, è astratta, punta a una loro comprensione e descrizione tramite le attitudini del linguaggio, ed è in grado di svilupparsi intorno al significato, al contenuto e al senso di ogni risorsa e del relativo contesto;
- una tipologia di ricerca semantica, sviluppata tramite il linguaggio dei linked data: tale ricerca è fondata su un'analisi semantica delle risorse e del loro contesto, presume di essere oggettiva nel ricavare categorie e relazioni, si sviluppa essenzialmente intorno al significato delle risorse e del contesto;
- una tipologia di ricerca semiotica, sviluppata tramite un linguaggio content-based: tale ricerca è fondata su un'analisi di genere semiotico delle risorse, è in parte oggettiva, fondata su evidenze concrete delle risorse digitali, si sviluppa sul contenuto delle risorse, sui dati contenutistici della loro costituzione.

Quindi, per quanto riguarda la ricerca terminologica, sviluppata tramite linguaggio term-based, nulla di nuovo da sottolineare. Dalle origini l'IR si fonda su tale tipologia di analisi e ricerca, e i termini sono gli strumenti di incontro tra l'analisi delle risorse e la loro indicizzazione con la definizione del bisogno informativo e la ricerca. Le risorse sono fondalmentalmente analizzate in base ai concetti che contengono o rappresentano, l'*aboutness*, e tali concetti sono espressi 'linguisticamente', secondo il linguaggio 'naturale', sia nella formulazione 'ideale' dell'operatore, sia nella resa in stringa alfanumerica di soggettazione, indicizzazione o classificazione. L'analisi non può, quindi, che essere astratta, proprio perchè punta a una comprensione e descrizione intellettuale e linguistica, ma ciò le consente di svilupparsi intorno sia al significato, sia al contenuto, sia al senso complessivo di ogni risorsa e del relativo contesto. Il linguaggio terminologico con cui si rappresenta la risorsa, allora, è quello che ne consente la ricerca 'a tutto tondo', pur con il limite dell'utilizzo del linguaggio terminologico stesso per l'indagine.

Per ciò che concerne la tipologia di ricerca semantica, anch'essa non è certo una novità, in quanto ben sviluppata tramite l'utilizzo di descrittori semantici all'interno della tipologia terminologica indicata sopra. La novità, dell'ultimo decennio, è che essa può essere condotta tramite il 'linguaggio' dei linked data. Scopo dei linked data è rendere «machine-understandable»[84] la ricerca intorno a concetti propri delle risorse, anche se essi sono definiti dall'operatore umano che riflette sui significati delle risorse stesse e li traduce in stringhe XML semantiche poi 'comprensibili' anche dagli elaboratori elettronici. Un'attività di 'riflessione' propria del sistema informatico, in realtà, è presente, e consiste nel meccanismo dell'inferenza, che permette di 'intuire' legami e 'conclusioni' magari nascoste allo stesso essere umano. Così, anche se il sistema è fondato su un'analisi semantica delle risorse più 'rudimentale' – che proprio perché non basata sulla concettualizzazione presume di essere oggettiva nel ricavare categorie e relazioni –, il processo si sviluppa intorno al significato delle risorse e del loro contesto. Questo criterio può premettere a una ricerca di tipo semantico 'padroneggiabile' dalle macchine, e premettere a un information discovery – in senso stretto – molto più 'umano' per quanto riguarda libertà e creatività, nonostante sia *computer-assisted*.

Infine, riguardo alla tipologia semiotica di analisi e ricerca sviluppata tramite il linguaggio content-based – approfondita nelle pagine seguenti –, il fatto che sia fondata su un'analisi di genere semiotico delle risorse, *computer-assisted* o interamente automatizzata, le consente di essere per buona parte oggettiva, fondata su evidenze 'concrete' delle risorse digitali, analizzate dal computer e quindi matematicamente obiettive. Quello a cui il sistema mira è il contenuto delle risorse (contenuto del contenitore formale-astratto), fatto concretamente/digitalmente di colori, linee, suoni, movimenti: cioè quello che di fatto una risorsa 'materialmente' è. Tale sistema perde i valori astratti del senso e del

[84] Marchitelli - Frigimelica 2012, p. 23.

significato delle risorse, ricompresi nel primo e nel secondo sistema – IR e linked data –, e si concentra sulla 'sensibilità', sul contenuto sensibile – come il sistema dei linked data si concentra sul significato –, ricorrendo all'integrazione con i primi due metodi per proporsi come sistema completo di ricerca.

Si potrebbe dedurre, allora, che il sistema veramente completo di ricerca sia il più tradizionale, quello terminologico, sviluppato tramite linguaggio terminologico, fondato sull'analisi linguistica, astratta, e in grado di puntare al significato, al contenuto e al senso di ogni risorsa e del contesto. Questo è anche il motivo della necessaria coesistenza dei diversi sistemi. Così come, però, è facile riconoscere le carenze dei sistemi dei linked data e del MIR, allo stesso modo si possono riconoscere le carenze del sistema tradizionale, almeno dal punto di vista della mancanza di capacità inferenziali e semiotiche. Tutto ciò non può che premettere – come *pars destruens* – al fatto che non ci sarebbe motivo, se non per chiarificare la questione, di distinguere nell'ambito di una generale metodologia di ricerca dell'informazione tra IR, information discovery e MIR. Su questa base, dunque, si può ulteriormente affrontare un percorso di discernimento e chiarificazione della metodologia del multimedia information retrieval.

Information retrieval e multimedia information retrieval

Nell'evoluzione tracciata finora, quindi, pare che il MIR non abbia ancora avuto il ruolo e la considerazione che avrebbe meritato, almeno non nell'ambito bibliotecario, archivistico, museale o documentale e della ricerca di informazione.

A dispetto di quanto avviene negli ambiti MAB, GLAM, LAMMS – che molti vantaggi potrebbero ottenere con l'applicazione di nuove metodologie per la ricerca di oggetti digitali riguardanti i beni culturali –, la sperimentazione e l'utilizzo delle tecnologie di MIR sono ben sviluppati nelle aree dell'ingegneria informatica, dell'intelligenza artificiale, della *computer vision* o dell'*audio processing*. La riflessione sullo sviluppo teorico e applicativo dei sistemi di analisi e ricerca content-

based, invece, è appena presente tra bibliotecari, documentalisti, archivisti, museologi e *information managers*. Probabilmente si tratta di introdurre meglio la problematica, ma spesso sembra che la questione sia nota, almeno a livello internazionale, e non interessi 'scioglierla', concedendo giusto la possibilità di preparare e condurre qualche esperimento.

In ogni caso, il contesto internazionale della LIS ha ancora l'occasione di accogliere per tempo la discussione, con la conseguente possibilità di indirizzare lo sviluppo di questi sistemi secondo necessità di ordine biblioteconomico-documentale. Questo dovrebbe essere un dovere per la LIS: interpretare e sviluppare tale rivoluzione culturale e tecnologica, incontrando e favorendo le necessità informative e conoscitive della società attuale. È compito della LIS interpretare queste necessità – di ordine anche democratico e sociale – risolvendo i problemi di descrizione, classificazione, indicizzazione e recupero che sono ancora ostacolo al libero e pieno accesso alle risorse dell'informazione e della conoscenza.[85] La ricerca ingegneristica e matematica è arrivata a capo di molti problemi relativi allo sviluppo in sé dei sistemi di MIR, o relativi ad applicazioni 'interne' a quegli stessi campi di sviluppo, ma questioni quali le necessità degli utenti, il loro approccio semantico, il rapporto con le interfacce di ricerca, rimangono una prerogativa della LIS.

È ancora necessario avviare, con il supporto degli studiosi delle tecnologie, un'ambiziosa, coraggiosa e anche utopica sperimentazione – anche a rischio di fallimenti. La sperimentazione deve contagiare biblioteche, archivi, musei, gallerie e altre istituzioni ed enti culturali che possiedono risorse multimediali digitali, raggiungendo anche televisoni, radio, laboratori e industrie interessati non solo a mettere a disposizone le proprie risorse a specialisti ed esperti, ma a chiunque possa e voglia, tramite la rete, accedere alle risorse culturali.

[85] In proposito, vedere: Pérez Álvarez 2006.

Obiettivo di sperimentazioni e ricerche, in vista dell'affermazione pratica del MIR, deve essere dimostrare e confermare la necessità di un 'cambiamento'. Se nella metodologia assodata dell'information retrieval ogni tipo di ricerca relativa alle risorse dell'informazione è riportato alle condizioni di una ricerca tramite linguaggio testuale, con i criteri del multimedia information retrieval ogni genere di risorsa digitale può essere trattato e ricercato tramite gli elementi di linguaggio più adatti alla sua natura oggettiva, coerenti con il contenuto concreto della risorsa stessa e il tipo di informazione ricercata.

Questa impostazione è quella che meglio si può adeguare al trattamento 'complessivo' e 'assoluto' di tutti i generi di risorse dell'informazione e della conoscenza. Si è spesso parlato del MIR come sistema 'organico', 'olistico', che raccoglie differenti aspetti specifici, ma adesso tale distinzione – in presenza di sistemi informatici che accolgono direttamente interrogazioni di tipo pienamente multimediale – è utile più che altro a fini 'storici' o esplicativi.[86] Esplicativamente, dunque, potremmo ancora distinguere quattro applicazioni specializzate, che comunque tra di loro si combinano continuamente interagendo: il 'text retrieval', nel quale il principio stesso del MIR legittima e garantisce l'utilizzo di dati testuali e terminologici per la ricerca di informazione testuale; il 'visual retrieval', secondo il quale le risorse visive sono analizzate e ricercate tramite dati visivi; il 'video retrieval', dove per il trattamento di risorse audiovisive si utilizza il linguaggio audiovisivo; l''audio retrieval', in cui l'informazione è ricercata tramite dati sonori[87].

[86] Queste differenziazioni, nell'ambito della teoria organica del MIR, sono trattate in: Raieli 2010, p. 177-223.

[87] L'*International journal of multimedia information retrieval* ha dedicato alcuni *special issues* a tali questioni specifiche: il vol. 2 (2013), n. 1, dedicato all'«hybrid music information retrieval»; il vol. 4 (2015), n. 1, sul «video retrieval»; il vol. 5 (2016), n. 1, sul «visual information retrieval». Come 'guida bibliografica' interna alla rivista, inoltre, si veda l'editoriale del direttore Michael Lew: Lew 2016.

Se, nei database dove il contenuto dei documenti è sostanzialmente un testo, appare ovvio e appropriato che le chiavi che ne consentono l'accesso siano termini e frasi estratti 'dall'interno' di quel contenuto stesso, nei database multimediali, invece, si rivela semplificativo e impreciso attribuire, 'dall'esterno', una descrizione testuale a contenuti che si fondano su un diverso 'regime di senso'. Inoltre, se per i testi è adeguato anche analizzare il concetto – pur con i limiti posti dalla soggettività dell'operatore – e attribuire a questo un descrittore terminologico, non è invece egualmente efficace fare lo stesso per le immagini o gli audiovisivi, dato che, anzitutto, i limiti soggettivi nel coglierne i concetti intimi sono maggiori e, subito dopo, questi sono alquanto 'indescrivibili' con i termini. Non sempre, inoltre, il concetto interessa maggiormente del contenuto concreto di una risorsa multimediale, delle rappresentazioni in se stesse, dei tratti, delle forme, dei colori e dei suoni, o anche delle parole in sé prese, di là da quelli che possono essere i significati impliciti di uno scritto, di un video, di una musica.[88]

I metodi di analisi e ricerca appartenenti all'ambito del multimedia information retrieval sono definiti content-based proprio perché utilizzano elementi di trattamento della stessa natura del contenuto concreto degli oggetti cui si applicano, in grado di 'mirare' con congruenza perfino al contenuto conoscitivo, agli aspetti di senso, di un dato oggetto digitale. L'information retrieval, 'di conseguenza', è definito come sistema di indicizzazione e ricerca term-based. La definizione term-based data al sistema tradizionale, infatti, nasce in relazione alle nuove concezioni del *content-based image retrieval* in quanto, nell'ambito dell'IR, la metodologia del trattamento terminologico è sempre apparsa come la naturale e unica via per la considerazione delle risorse dell'informazione.

88 Alle radici della questione, vedere: Svenonius 1994.

Dunque, come nel caso del trattamento di un insieme di documenti di natura principalmente testuale un sistema 'concettuale', semantico, di IR, basato sullo sviluppo di una cultura terminologica, può essere per buona parte idoneo ed efficace, nel caso dell'applicazione a risorse multimediali nel loro complesso è determinate, piuttosto, un sistema 'formale', semiotico, di ricerca e recupero, fondato anche sulle capacità percettive concrete e immediate che caratterizzano ogni tipo di persona.

Le interrogazioni multimediali tradizionali, espresse solo terminologicamente, sono inadeguate al complesso delle sempre più ricche esigenze delle persone, almeno quando queste propongono dati livelli o tipologie di ricerca – di là dalla conoscenza, o dell'utilità, di specifici ed 'esclusivi' sistemi di indicizzazione. Nella metodologia del MIR, invece, la formulazione di richiesta non deve essere necessariamente costretta entro i limiti della lingua, ma può essere inviata così come è spontaneamente prodotta – ben 'democraticamente' –, com'è nata nella persona, in caratteri immediatamente visivi, sonori, audiovisivi, e testuali giusto nei casi appropriati. Nella stessa maniera è 'afferrata' e soddisfatta dal sistema, attraverso colori, forme, strutture, suoni, movimenti, luci, ombre eccetera, non escludendo le parole quando il contenuto ricercato sono esse stesse.

La problematicità del metodo del multimedia information retrieval

Qualunque sia la metodologia di *information searching*, e in particolare per quella del MIR, i metodi content-based, e in genere quelli automatici, non sempre risultano i più adatti a soddisfare le esigenze più elevate di studiosi e specialisti, come delle persone comuni: il senso di un oggetto rappresentato da una risorsa deve essere colto nella 'totalità', nella considerazione simultanea delle molte qualità sensibili e intellettuali, di aspetto e di significato, concrete e astratte.

I sistemi di MIR mantengono validità nel caso di un approccio diretto e 'contenutistico-oggettivo' alle risorse, ma presentano limitatezze nel caso di un approccio teorico e 'intellettuale-interpretativo'. Un

buon livello di efficacia nel recupero delle risorse multimediali si può raggiungere solo utilizzando in combinazione metodologie di analisi e ricerca basate sia sulla rappresentazione del contenuto – attraverso elementi multimediali – sia sulla definizione dei significati – tramite termini e schemi semantici. Nell'insieme organico del MIR, dunque, è necessario integrare le tecniche content-based, semiotiche, con le modalità semantiche term-based – e, si potrebbe aggiungere, 'linked data-based'.

Detto ciò, trattandosi di sistemi che funzionano in base ad algoritmi sempre più complessi di *processing* dei dati informativi, è necessario capire bene quale efficacia possono avere le procedure freddamente matematiche dei sistemi content-based in rapporto agli obiettivi personali e pratici della ricerca di informazioni e di conoscenze delle persone. Proprio nel senso del superamento della 'distanza' tra essere umano e macchina si muove l'intera ricerca per la realizzazione di algoritmi di calcolo e procedure di elaborazione dei dati non solo matematicamente efficienti ma anche pragmaticamente efficaci.[89]

La questione più problematica è quella dell'interpretazione dell'oggetto multimediale, che ha un valore considerevole nel processo di ricerca quando l'esigenza di informazione va oltre le caratteristiche percettive dell'oggetto stesso, calcolabili automaticamente dalla macchina, e si spinge al livello dell'interpretazione semantica, definibile solo dall'essere umano. In alcuni casi, è necessario che la ricerca content-based sia *knowledge-assisted*, con il supporto di una descrizione soggettiva della propria esigenza di informazione da parte dell'utente di un sistema. In questi casi è utile che la descrizione possa incontrare i 'soggetti' dei documenti elaborati da un operatore umano, che indicano sia all'utente sia alla macchina quello che le analisi matematiche di una qualunque *query* non possono 'cogliere' direttamente.

[89] In proposito: Yoshitaka - Ichikawa 1999; Maybury 2012.

Se si può dare per certa l'efficienza meccanica e assoluta dei processi matematici, quindi, non si può fare lo stesso riguardo la loro utilità pratica in relazione alle esigenze di ogni utente finale. Serve capire, dunque, il valore delle operazioni matematicamente 'oggettive' compiute dai sistemi informatici, prive degli errori prodotti dalla valutazione umana delle risorse e dei contenuti, ma anche senza la peculiare flessibilità e intelligenza di questa nell'interpretarne gli aspetti non oggettivamente evidenti. La 'matematicità' delle operazioni del sistema, infatti, può contrastare non solo con l''umanità' delle richieste e delle attese delle persone, ma anche con i generali principi di libertà, sensibilità e democraticità dell'information discovery – di cui si è parlato – e dei sistemi content-based in specifico.

Resta, comunque, il problema più complesso da risolvere, teoricamente e praticamente: il 'senso' di un oggetto rappresentato da una risorsa deve essere colto nella reale totalità delle sue qualità sensibili e intellettuali – come, in fondo, fanno i sistemi tradizionali di IR. I sistemi di accesso orientati al contenuto concreto, invece, si dimostrano poco adeguati per indicare la molteplicità degli spunti interpretativi intellettuali, e l'inesistente 'sensibilità' della macchina non può essere prodotta pienamente da elaborazioni algoritmiche dei dati numerici rappresentativi delle qualità degli oggetti digitali o digitalmente rappresentati. Inoltre, se la questione del senso (e del significato) delle risorse si sviluppa nella problematica del cosiddetto *semantic gap* – di cui nel successivo paragrafo –, la questione della sensibilità delle macchine si inquadra in un ampio problema parallelo – che potrebbe essere oggetto di una parallela trattazione teoretica di natura estetica –, quello del «sensory gap», o *gap* semiotico.[90]

[90] Per queste argomentazioni, vedere: Smeulders *et al.* 2000.

La questione del semantic gap

La differenza e la distanza tra i due generi di approccio presentati sopra può essere definita *semantic gap*: la non coincidenza tra l'informazione oggettiva che si può estrarre direttamente da una risorsa e l'interpretazione diversa che gli stessi dati possono ricevere da ogni persona in ogni specifica situazione. Dato che il significato di una risorsa multimediale è raramente esplicito, scopo del sistema di trattamento deve essere fornire il supporto per superare questo vuoto tra la semplicità del trattamento semiotico offerto dalle macchine e la ricca aspettativa semantica delle persone.

Le caratteristiche di questo *gap* della rappresentazione, e le possibilitità di colmarlo, sono trattate in paticolar modo da Peter Enser.[91] I livelli rappresentativi di una risorsa variano da quello più basso composto dalla semplice estrazione dei suoi «raw data», elaborata immediatamente dalla macchina, fino al livello più alto costituito dalle «semantics» che esso convoglia, così come sono interpretate dagli utenti. In genere, le caratteristiche del *gap* semantico cambiano da un caso all'altro, secondo il livello di complessità della risorsa. Il problema del *gap*, comunque, nasce già nei primi livelli rappresentativi e cresce via via che si sale verso il livello semantico. Infatti, se in alcuni casi si può pure giungere – anche automaticamente – ad assegnare con facilità un nome o un'etichetta descrittiva a un oggetto, quando poi ci si interroga riguardo al suo significato si crea un vuoto tra l'etichetta e il livello più alto in cui lo si vuole nuovamente prendere in considerazione.[92]

Le persone possono giungere al più alto livello di esigenza formulando richieste di risorse con un valore intellettualmente raffinato, dotate di attributi di significato assegnati anche grazie a un 'contesto' culturale di riferimento, impossibili da identificare senza il supporto semantico-terminologico. Proprio di questo genere di

[91] Vedere: Enser 2008b; Hare *et al.* 2006.
[92] Hare *et al.* 2006, p. 75-76.

ricerche si occupano, in fondo, i sistemi tradizionali di IR, con tutti i limiti dell'astrazione concettuale, e proprio questo livello informativo è quello più difficilmente raggiungibile dai sistemi content-based, basati sulla considerazione semiotica, più che semantica, di una risorsa.

Soluzione di principio – proposta dallo stesso Enser – è integrare nei sistemi content-based l'apporto delle ontologie, strumenti per la concettualizzazione condivisa di un dominio, composti da classi di concetti e relazioni tra essi. La ricchezza semantica di una risorsa è vicina alla rappresentabilità se si posiziona l'oggetto nell'ambito di uno schema semantico, o KOS, tipico del *semantic Web*. Il ricorso alle ontologie nei sistemi di MIR consente, allora, di rendere esplicito parte del significato di una risorsa, e rende possibile formulare l'interrogazione anche tramite i concetti e le relazioni tra concetti. La *query* multimediale può essere, così, completata semanticamente, integrando continuamente gli strumenti di ricerca content-based che si concentrano sugli oggetti immediatamente presi. L'ontologia può essere, ancora, un ulteriore strumento per la navigazione in un insieme di risorse, nonché – ed è tra le cose più rilevanti – costituire una criterio di interoperabilità tra sistemi differenti e tra *dataset* diversi.[93]

Acquisendo l'apporto delle ontologie, per sfuggire al loro 'rigore' strutturale – che può riproporre una certa rigidezza e astrattezza in un sistema di ricerca che vuole essere molto più 'libero' dei tradizionali sistemi di IR – si può progettare la combinazione delle *ontologies* con le *folksonomies*, sistemi di libera categorizzazione collaborativa dei contenuti sulla base di *tag* assegnati direttamente dagli utenti finali.[94]

Tutto questo è in linea con i principi dei sistemi di MIR, dove la possibilità per le persone di cercare liberamente, anche tramite modelli o bozzetti personali, consente alla macchina di 'apprendere'

[93] Cfr. Hare *et al.* 2006, p. 83-84. Vedere anche: Mallik - Chaudhury 2012.

[94] Questione già da tempo presente tra gli stessi fondatori del Web semantico: Shadbolt *et al.* 2006; Guy - Tonkin 2006. Vedere inoltre: Yang 2012.

al momento una nuova informazione sulle risorse, che sarà registrata insieme alle informazioni già definite, integrando e ampliando le sue capacità 'interpretative'. L'integrazione tra gli strumenti semantici delle ontologie e delle folksonomie, a loro volta integrati agli strumenti content-based, consente di stabilire un approccio organico per tutti i tipi di risorse multimediali, che tiene in conto univocamente la loro rappresentabilità concreta e concettuale, semiotica e semantica, e può portare, dunque, alla conciliazione dell'opposizione tra i principi dei trattamenti semantico-interpretativo e contenutistico-oggettivo.

La vasta problematica sollevata dalle 'proposte' del MIR sembra possa confluire in una questione conclusiva, che rappresenta contemporaneamente la presa di coscienza dei limiti del sistema content-based e la proposta di superamento degli stessi.[95] Ricercare o scoprire l'informazione multimediale non potrà mai essere troppo facile, un processo realizzabile tramite uno strumento rigoroso ed 'esatto': come nella vita quotidiana, molti problemi possono essere risolti solo in modo occasionale, ricorrendo anche alle risorse della fantasia e della creatività.[96]

Non si può pretendere di eliminare l'apporto dell'interpretazione umana al processo della ricerca, e soprattutto della scoperta, in quanto fondato su un volume di conoscenze precedenti e su una raffinatezza di elaborazione dei concetti impossibili da raggiungere per le macchine. I più elaborati algoritmi, in grado di calcolare molte possibili interpretazioni, non sono in grado di colmare il *semantic gap*, che si continua a creare tra un esame meccanico di basso livello dell'aspetto di un oggetto e la valutazione di alto livello dell'idea umana di tale oggetto. Non resta, allora, che definire una prospettiva di collaborazione tra il ricercatore di informazione e gli strumenti di analisi e ricerca di essa. La possibilità, in tal modo, di strutturare

[95] Tale questione si trova sviluppata da tempo nelle considerazioni di Enser: Enser 2000.

[96] Ciò riporta a un atteggiamento di *serendipity*: Merton - Barber 2004.

organicamente la ricerca e la scoperta delle risorse permetterà alle persone, volta per volta, di usare in modo libero e flessibile, accanto agli strumenti semiotici per operare direttamente sui contenuti concreti, anche i mezzi semantici degli schemi concettuali.

Ai classici sistemi term-based dei database, o anche del Web, potranno quindi succedere sistemi che permettano di effettuare la ricerca e la scoperta in diverse dimensioni.[97] La ricerca avverrà attraverso dati che sono parole estratte dal pieno di uno scritto o dal parlato di un audiovisivo, immagini chiave di una sequenza, figure geometriche, melodie, nonché termini descrittivi e semantici, e stringhe di linked data…

Epilogo

Tutto quanto affermato sul MIR – come si capisce – non ha mai implicato il rifiuto delle forme di interpretazione e rappresentazione concettuale del contenuto di una risorsa e della risorsa stessa. Presi in considerazione i limiti semantici dei sistemi content-based, è sempre necessario un appropriato apporto intellettuale nell'organizzazione e nella ricerca delle risorse, per definire i significati oltre le sensazioni, per specificare le strategie di ricerca e per aumentare le possibilità di scoperta e recupero.

È necessario un approccio organico al sistema dell'information discovery in generale, che integri i mezzi contenutistici e semantici per la ricerca e scoperta delle risorse dell'informazione e della conoscenza. Tale approccio dovrà essere valido per tutte le tipologie di risorse multimediali, dovrà prendere univocamente in considerazione la loro rappresentabilità materiale e concettuale, e considerare le necessità informative contenutistico-oggettive e intellettual-interpretative.

I sistemi di ricerca multimediale più avanzati possono incarnare contemporaneamente le due distinte anime: quella di strumento avanzato

[97] Riguardo le interfaccie 'multidimensionali' di ricerca, vedere: Ah-Pine *et al.* 2015.

per la ricerca dei professionisti e degli studiosi, e quella di democratica guida per l'accesso alle risorse disposta per gli utenti generici.[98] In ogni caso ogni persona potrà attingere in modo diversificato alla propia intelligenza e alla propria sensibilità, alla preparazione culturale come all'intuito e all'immagnazione, interagendo con un sistema sempre in grado di accogliere inaspettate variazioni della linea di ricerca, e di 'comprendere' la strategia seguita, più o meno strutturata, apprendendo dal comportamento del ricercatore.

Le differenti procedure operano al meglio in una continua e organica interazione, in una singola e olistica strutturazione della strategia – più o meno consapevole – di ricerca in un database oppure sul Web.[99] I nuovi sistemi, non solo di IR, non solo di MIR o di information discovery, devono essere preparati per combinare insieme tutte le linee di ricerca e scoperta tradizionali e innovative, quelle dei database e quelle di Internet, quelle content-based e quelle del Web semantico, dagli approcci descrittivi e concettuali a quelli contenutistici e semiotici, fino a quelli 'onnicomprensivi' dei linked data. Consentire diverse strategie di ricerca – combinando termini, concetti, parole, figure, movimenti, suoni, classi e codici – è determinante per la scoperta di risorse molto complesse, il cui contenuto conoscitivo si estende attraverso tutti i livelli di senso e significato.

La prospettiva finale del MIR sarebbe il raggiungimento dell'automazione completa nell'interpretazione, la ricerca e il recupero contenutistici e semantici delle risorse multimediali, ma la soluzione del *semantic gap* resta il problema più arduo da superare.[100] Sarebbe necessario produrre macchine in grado di raggiungere i significati di 'alto livello' partendo dalle caratteristiche di 'livello base' degli oggetti, cioè riprodurre i processi conoscitivi umani tramite algoritmi capaci – davvero – di riprodurre le combinazioni e i 'montaggi' della

98 Beaudoin 2016.
99 Ah-Pine *et al.* 2015.
100 Vedere: Jiang *et al*. 2016; Tan - Ngo 2016; Jarrar - Belkhatir 2016.

mente, che portano dai dati materiali e scollegati alle forme astratte e complesse delle 'idee'.[101]

Ipotizzare che i sistemi automatici possano avvicinarsi al raffinato livello semantico-interpretativo del pensiero umano è, comunque, abbastanza difficile. Tale livello cognitivo è ampiamente 'logico', ma è rilanciato anche da forme di conoscenza ineffabili, o 'tacite', da intuizioni inesplicabili, da emozioni percettive. Il *gap* tra essere umano e macchina, nella sostanza, rimane, e può essere 'ponteggiato' solo in una prospettiva di collaborazione tra i due differenti approcci.

Se ci si riferisce alle possibilità generali del *semantic Web*, una cosa è costruire connessioni logiche tra stringhe di testo – in ultima analisi di significato 'linguistico' –, altra cosa è interpretare, non solo formalmente e logicamente, ma anche 'emozionalemente', le risorse multimediali. Così, anche l'uso di ontologie e linked data solo parzialmente copre il vuoto tra l'apparenza e l'essenza di un oggetto multimediale, posizionando l'oggetto in una categoria utile, ma mai scoprendo interamente l'enigma della sua 'reale' interpretazione.

Infine, gli sviluppi della società nella prospettiva della conoscenza come *commons*[102] hanno reso necessario che anche i sistemi informativi e le risorse diventino un bene comune, e hanno reso imperativa la 'democratizzazione' tecnologica dell'accesso. Accesso libero e aperto – si è detto – significa disponibilità delle risorse, ma anche disponibilità e semplicità dei sistemi di ricerca e scoperta delle risorse. A questo si collega lo spirito dell''assistenza' a coloro che non sono comunque in grado di usare nel modo migliore tali strumenti, e che necessitano di diventare *information literate*.[103]

---

101 In proposito: Xu - Wang 2015; Bakker 2016. Riguardo tale questione, vecchia di almeno dieci anni, vedere anche il vol. 4 (2015), n. 2, dell'*International Journal of Multimedia Information Retrieval*, dedicato, come *special issue*, alla «concept detection with big data».

102 Hess - Ostrom 2006.

103 Riguardo l'ampia questione dell'*information literacy* valga (per tutto il resto) il *framework* proposto non troppo tempo fa dall'ACRL: ACRL 2016.

In questo senso, molte nuove interfacce di ricerca hanno una configurazione 'estetica' e 'cognitiva', così come i *browser* semantici e simili,[104] rappresentando in forma pienamente visiva le informazioni, le loro classi, i legami, i contesti che le riguardano.[105] È, dunque, nelle prospettive della *data visualization* e dello *storytelling* dei dati che pare possibile individuare un'infrastruttura percettiva e concettuale comune, in grado di favorire l'adozione da parte degli utenti di comportamenti informativi adeguati ed efficaci.[106] Comportamenti di carattere visivo, audiovisivo… in vario modo simili a quelli degli utenti di un sistema di MIR nella loro ricerca essenzialmente contenutisitca delle informazioni e delle risorse.

Anche per questo, dunque, è già necessario pensare oltre il Web semantico, oltre gli stessi percorsi liberi dell'information discovery, dove anche lo spirito dell'accesso semiotico, la libera e occasionale *similarity search*, immediatamente intuitivi e sensibili, hanno un ruolo e un'efficacia, per favorire l'approccio alla conoscenza di circoli sempre più ampi di persone, anche se hanno poche possibilità di studiare o sviluppare attitudini intellettuali.

104 Un esempio classico è il progetto Linked Jazz: <https://linkedjazz.org/>.

105 In tal senso, può essere anche lo schema base di BIBFRAME: <http://www.loc.gov/bibframe>.

106 Questo è il senso di un'argomentazione recentemente proposta, che contestualizza la tematica: Vivarelli 2016.

BIBLIOGRAFIA

ACRL 2016 = Association of College and Research Libraries, *Framework for information literacy for higher education*, 2016, <http://www.ala.org/acrl/standards/ilframework>.

Adami *et al.* 2012 = *Analysis, retrieval and delivery of multimedia content*, edited by Nicola Adami *et al.*, Berlin, Springer, 2012.

Ah-Pine *et al.* 2015 = Julien Ah-Pine *et al.*, *Unsupervised visual and textual information fusion in CBMIR using graph-based methods*, «ACM transactions on information systems», 33 (2015), n. 3, p. 1-31.

Arnheim 1969 = Rudolf Arnheim, *Visual thinking*, Berkley, University of California Press, 1969.

Arnold 2009 = Stephen E. Arnold, *Real-time search: where retrieval and discovery collide*, «Online», 33 (2009), n. 6, p. 40-41.

Baeza-Yates - Ribeiro-Neto 2000 = Ricardo Baeza-Yates, Berthier Ribeiro-Neto, *Modern Information Retrieval*, New York, Longman, 2000.

Bakker 2016 = Erwin M. Bakker, *Open and free datasets for multimedia retrieval*, «International journal of multimedia information retrieval», 5 (2016), n. 3, p. 135-136.

Balsamo 1992 = Luigi Balsamo, *La bibliografia: storia di una tradizione*, Firenze, Sansoni, 1992.

Bambini - Wakefield 2014 = Cristina Bambini, Tatiana Wakefield, *La biblioteca diventa social*, Milano, Bibliografica, 2014.

Bates 1989 = Marcia J. Bates, *The design of browsing and berrypicking techniques for the online search interface*, «Online review», 5 (1989), n. 13, p. 407-424.

Bates 1990 = Marcia J. Bates, *Where should the person stop and the information search interface start?*, «Information processing & management», 1990, n. 26, p. 575-591.

Bates 2016 = Marcia J. Bates, *Selected works of Marcia J. Bates*, Berkeley, Ketchikan, 2016.

Beaudoin 2016 = Joan E. Beaudoin, *Content-based image retrieval methods and professional image users*, «Journal of the Association for Information Science & Technology», 67 (2016), n. 2, p. 350-365.

Berners-Lee 2006 = Tim Berners-Lee, *Linked data*, 2006, <http://www.w3.org/DesignIssues/LinkedData.html>.

Berners-Lee - Fischetti 1999 = Tim Berners-Lee with Mark Fischetti, *Weaving the Web: the original design and ultimate destiny of the World Wide Web by its inventor*, San Francisco, Harper, 1999.

Berners-Lee - Hendler - Lassila 2001 = Tim Berners-Lee, James Hendler, Ora Lassila, *The semantic Web: a new form of web content that is meaningful to computers will unleash a revolution of new possibilities*, «Scientific American», 284 (2001), n. 5, p. 34-43.

Biagetti 2010 = Maria Teresa Biagetti, *Nuove funzionalità degli OPAC e relevance ranking*, «Bollettino AIB», 50 (2010), n. 4, p. 339-356.

Biagetti 2014 = Maria Teresa Biagetti, *Sviluppi e trasformazioni delle biblioteche digitali: dai repositories di testi alle semantic digital libraries*, «AIB studi», 54 (2014), n. 1, p. 11-34.

Bianchini 2010 = Carlo Bianchini, *Futuri scenari: RDA, REICAT e la granularità dei cataloghi*, «Bollettino AIB», 50 (2010), n. 3, p. 219-238.

Bianchini 2012 = Carlo Bianchini, *Dagli OPAC ai library linked data: come cambiano le risposte ai bisogni degli utenti*, «AIB studi», 52 (2012), n. 3, p. 303-323.

Bianchini - Guerrini 2014 = Carlo Bianchini, Mauro Guerrini, *Introduzione a RDA*, Milano, Bibliografica, 2014.

Breeding 2010 = Marshall Breeding, *The state of the art in library discovery 2010*, «Computers in libraries», 30 (2010), n. 1, p. 31-35.

Bush 1945 = Vannevar Bush, *As we may think*, «The Atlantic monthly», giugno 1945, p. 101-108.

Caplan 2012 = Priscilla Caplan, *On discovery tools, OPACs and the motion of library language*, «Library hi tech», 30 (2012), n. 1, p. 108-115.

Caproni 2007 = Attilio Mauro Caproni, *L'inquietudine del sapere: scritti sulla bibliografia*, Milano, Bonnard, 2007.

Cassella 2015 = Maria Cassella, *A metà del guado: l'open access tra passato, presente e futuro*, «Biblioteche oggi trends», 1 (2015), n. 1, p. 60-68.

Cassella 2012 = Maria Cassella, *Open Access e comunicazione scientifica*, Milano, Bibliografica, 2012.

Castellucci 2009 = Paola Castellucci, *Dall'ipertesto al Web: storia culturale dell'informatica*, Roma, Laterza, 2009.

Castellucci 2011 = Paola Castellucci, *Tempo e massa: una nuova energia nella comunicazione scientifica*, «Bollettino AIB», 51 (2011), n. 3, p. 237-244.

Castellucci 2013 = Paola Castellucci, *Mundaneum. Una prospettiva geopolitica per la Documentazione*, «Nuovi annali della Scuola speciale per archivisti e bibliotecari», 27 (2013), p. 105-119.

Cleverdon - Mills - Keen 1966 = Cyril W. Cleverdon, Jack Mills, Michael Keen, *Factors determining the performance of indexing systems*, Cranfield, Aslib Cranfield Research Project, 1966, <https://dspace.lib.cranfield.ac.uk/handle/1826/2/simple-search?query=Factors+determining+the+performance+of+indexing+systems>.

De Robbio 2012 = Antonella De Robbio, *Forme e gradi di apertura dei dati: i nuovi alfabeti dell'Open Biblio tra scienza e società*, «Biblioteche oggi», 30 (2012), n. 6, p. 11-24.

Deb 2004 = *Multimedia systems and Content-Based Image Retrieval*, edited by Sagarmay Deb,
Hershey, Idea Group 2004.

Del Bimbo 1999 = Alberto Del Bimbo, *Visual Information Retrieval*, San Francisco, Kaufmann, 1999.

Divakaran 2009 = *Multimedia content analysis: theory and applications*, edited by Ajay Divakaran, Berlin, Springer, 2009.

Doyle - Becker 1975 = Lauren Doyle, Joseph Becker, *Information Retrieval and processing*, Los Angeles, Melville, 1975.

Ellero 2013 = Nadine P. Ellero, *Integration or disintegration: where is discovery headed?*, «Journal of library metadata», 13 (2013), n. 4, p. 311-329.

Enser 1995 = Peter G. B. Enser, *Pictorial information retrieval: progress in documentation*, «Journal of Documentation» , 51 (1995), n. 2, p. 126-170.

Enser 2000 = Peter G. B. Enser, *Visual image retrieval: seeking the alliance of concept-based and content-based paradigms*, «Journal of Information Science», 4 (2000), n. 26, p. 199-210.

Enser 2008 = Peter G. B. Enser, *Visual image retrieval*, «Annual review of information science and technology», 42 (2008), n. 1, p. 1-42.

Enser 2008b = Peter G. B. Enser, *Visual image retrieval*, in *Annual review of information science and technology (ARIST)*, edited by Blaise Cronin, New York, AAIST, 2008, p. 3-42.

Fagan 2012 = Jody Condit Fagan, *Top 10 discovery tool myths*, «Journal of web librarianship», 6 (2012), n. 1, p. 1-4.

Fairthorne 1969 = Robert A. Fairthorne, *Content analysis, specification and control*, «Annual review of information science and technology», 1969, n. 4, p. 73-109.

Frame 2004 = Michael T. Frame, *Information discovery and retrieval tools*, «Information services & use», 24 (2004), n. 4, p. 187-193.

Gast *et al.* 2013 = Erik Gast *et al.*, *Very large scale nearest neighbor search: ideas, strategies and challenges*, «International journal of multimedia information retrieval», 2 (2013), n. 4, p. 229-241.

Gilster 1997 = Paul Gilster, *Digital literacy*, New York, Wiley, 1997.

Gombrich 1982 = Ernst Hans Gombrich, *The Image and the eye*, Oxford, Phaidon, 1982.

Guerrini 2014 = Mauro Guerrini, *BIBFRAME: un'ipotesi di ambiente bibliografico nell'era del web*, in *Il libro al centro. Percorsi fra le discipline del libro in onore di Marco Santoro*, Napoli, Liguori, 2014.

Guerrini - Possemato 2012 = Mauro Guerrini, Tiziana Possemato, *Linked data: un nuovo alfabeto del web semantico*, «Biblioteche oggi», 30 (2012), n. 3, p. 7-15.

Guerrini - Possemato 2015 = Mauro Guerrini, Tiziana Possemato, *Linked data per biblioteche, archivi e musei*, Milano, Bibliografica, 2015.

Guy - Tonkin 2006 = Marieke Guy, Emma Tonkin, *Folksonomies: tidying up tags?*, «D-Lib magazine», 12 (2006), n. 1, <www.dlib.org/dlib/january06/guy/01guy.html>.

Hare *et al.* 2006 = Jonathon S. Hare *et al.*, *Mind the gap: another look at the problem of the semantic gap in image retrieval*, in *Proceedings of multimedia content analysis, management and retrieval 2006*, edited by Ei Y. Chang *et al.*, San Jose, SPIE, 2006, p. 75-86.

Hanjalic 2012 = Alan Hanjalic, *New grand challenge for multimedia information retrieval: bridging the utility gap*, «International journal of multimedia information retrieval», 1 (2012), n. 3, p. 139-152.

Heath - Bizer 2011 = Tom Heath, Christian Bizer, *Linked data: evolving the Web into a global data space*, New York, Morgan & Claypool, 2011.

Hess - Ostrom 2006 = *Understanding knowledge as a commons: from theory to practice*, edited by Charlotte Hess, Elinor Ostrom, Cambridge, MIT, 2006.

Hjørland 2001 = Birger Hjørland, *Towards a theory of aboutness […]*, «Journal of the American Society for Information Science and Technology», 52 (2001), n. 9, p. 774-778.

Hoeppner 2012 = Athena Hoeppner, *The ins and outs of evaluating web-scale discovery services*, «Computers in Libraries», 32 (2012), n. 3, <http://www.infotoday.com/cilmag/apr12/Hoeppner-Web-Scale-Discovery-Services.shtml>.

Hutchins 1978 = William J. Hutchins, *The concept of 'aboutness' in subject indexing*, «Aslib proceedings», 30, 1978, p. 172-181.

Iacono 2014 = Antonella Iacono, *Linked data*, Roma, AIB, 2014.

Iacono 2014 = Antonella Iacono, *Dal record al dato: linked data e ricerca dell'informazione nell'OPAC*, «JLIS.it», 5 (2014), n. 1, p. 77-102.

Jarrar - Belkhatir 2016 = Radi Jarrar, Mohammed Belkhatir, *On the coupled use of signal and semantic concepts to bridge the semantic and user intention gaps for visual content retrieval*, «International journal of multimedia information retrieval», 5 (2016), n. 3, p. 165-172.

Järvelin 2003 = Kalervo Järvelin, *Information Retrieval (IR)*, in *International encyclopedia of information and library science*, edited by John Feather, Paul Sturges, London, Routledge, 2003, p. 293-295.

Jiang *et al.* 2016 = Lu Jiang *et al.*, *Text-to-video: a semantic search engine for internet videos*, «International journal of multimedia information retrieval», 5 (2016), n.1, p. 3-18.

Kapetanios - Tatar - Sacarea 2013 = Epaminondas Kapetanios, Doina Tatar, Christian Sacarea, *Natural language processing: semantic aspects*, Hoboken, CRC, 2013.

Kato 1992 = Toshikazu Kato, *Database architecture for content-based image retrieval*, in *Image storage and retrieval systems: SPIE proceedings vol. 1662*, San Jose, SPIE, 1992, p. 112-123.

Ketterman - Inman 2014 = Elizabeth Ketterman, Megan E. Inman, *Discovery tool vs. PubMed: a health sciences literature comparison analysis*, «Journal of electronic resources in medical libraries», 11 (2014), n. 3, p. 115-123.

Kovács 2014 = Béla Lóránt Kovács, Margit Takács, *New search method in digital library image collections: a theoretical inquiry*, «Journal of Librarianship and Information Science», 46 (2014), n. 3, p. 217-225.

Kroeger 2013 = Angela Kroeger, *The road to BIBFRAME: the evolution of the idea of bibliographic transition into a post-MARC future*, «Cataloging & classification quarterly», 51 (2013), n. 8, p. 873-890.

Kurdi 2016 = Mohamed Zakaria Kurdi, *Natural language processing and computational linguistics: speech, morphology and syntax*, Somerset, Wiley, 2016.

La Fontaine 1916 = Henri La Fontaine, *The great solution: magnissima charta*, Boston, World Peace Foundation, 1916.

Lancaster 2003 = Frederick W. Lancaster, *Indexing and abstracting in theory and practice*, Urbana Champaign, University of Illinois, 2003.

Levine-Clark 2014 = Michael Levine-Clark, *Access to everything: building the future academic library collection*, «portal: libraries and the academy», 14 (2014), n. 3, p. 425-437.

Lew 2016 = Michael S. Lew, *Top multimedia information retrieval papers*, «International journal of multimedia information retrieval», 5 (2016), n. 3, p. 133-134.

Lynch 1995 = Clifford A. Lynch, *Networked information resource discovery: an overview of current issues*, «IEEE journal on selected areas of communications», 13 (1995), n. 8, p. 1505-1522.

LOC 2012 = Library of Congress, *Bibliographic framework as a web of data: linked data model and supporting services*, Washington, LOC, 2012, <http://www.loc.gov/bibframe/pdf/marcld-report-11-21-2012.pdf>.

Luther - Kelly 2011 = Judy Luther, Maureen C. Kelly, *The next generation of discovery*, «Library journal», 136 (2011), n. 5, p. 66-71.

Mallik - Chaudhury 2012 = Anupama Mallik, Santanu Chaudhury, *Acquisition of multimedia ontology: an application in preservation of cultural heritage*, «International journal of multimedia information retrieval», 1 (2012), n. 4, p. 249-262.

Manning - Raghavan - Schütze 2008 = Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze, *Introduction to Information Retrieval*, Cambridge, Cambridge University Press, 2008.

Marchitelli - Frigimelica 2012 = Andrea Marchitelli, Giovanna Frigimelica, *OPAC*, Roma, AIB, 2012.

Matthews *et al.* 2010 = Brian Matthews *et al.*, *An evaluation of enhancing social tagging with a knowledge organization system*, «Aslib proceedings», 62 (2010), n. 4/5, p. 447-465.

Maybury 2012 = Maybury Mark T., *Multimedia information extraction*, New York, Wiley-IEEE, 2012.

Meghabghab - Kandel 2008 = George Meghabghab, Abraham Kandel, *Search engines, link analysis and user's web behavior*, Berlin, Springer, 2008.

Merton 1973 = Robert K. Merton, *The sociology of science: theoretical and empirical investigations*, Chicago, University of Chicago Press, 1973.

Merton - Barber 2004 = Robert K. Merton, Elinor G. Barber, *The travels and adventures of serendipity*, Princeton, Princeton University Press, 2004.

Mundaneum 1995 = *Cent ans de l'Office International de Bibliographie, 1895-1995: les prémisses du Mundaneum*, Mons, Mundaneum, 1995.

Nelson 1990 = Theodor Holm Nelson, *Literary machines*, Sausalito, Mindful, 1990.

Nielsen 2006 = Jakob Nielsen, *Designing web usability*, Berkeley, New Riders, 2006.

Otlet 1914 = Paul Otlet, *La fin de la guerre: traité de paix générale*, Bruxelles, Lamberty, 1914.

Otlet 1934 = Paul Otlet, *Traité de documentation*, Bruxelles, Mundaneum, 1934.

Page 2001 = *Google's Larry Page: good ideas still get funded*, «Business week magazine», 13 marzo 2001, <http://www.bloomberg.com/news/articles/2001-03-12/googles-larry-page-good-ideas-still-get-funded>.

Pagliero Popp - Dallis 2012 = *Planning and implementing resource discovery tools in academic libraries*, edited by Mary Pagliero Popp, Diane Dallis, Hershey, Information science reference, 2012.

Panofsky 1955 = Erwin Panofsky, *Meaning in the visual arts*, Garden City, Dobleday, 1955.

Pérez Álvarez 2006 = Sara Pérez Álvarez, *Aproximación al estudio de los sistemas de recuperación de imágenes 'CBIR' desde el ámbito de la Documentación*, «Documentación de las ciencias de la información», 29 (2006), p. 301-315.

Petrucciani 2013 = Alberto Petrucciani, *L'utopia della documentazione: a proposito di una lettera inedita di Paul Otlet a Luigi de Gregori (1937)*, «Nuovi annali della Scuola speciale per archivisti e bibliotecari», 27 (2013), p. 121-137.

Proper - Bruza 1999 = Henderik A. Proper, Peter D. Bruza, *What is information discovery about?* «Journal of the American society for information science», 50 (1999), n. 9, p. 737-750.

Raieli 2010 = Roberto Raieli, *Nuovi metodi di gestione dei documenti multimediali: principi e pratica del multimedia information retrieval*, Milano, Bibliografica, 2010.

Raieli 2015 = Roberto Raieli, *Vecchi paradigmi e nuove interfacce: la ricerca di un equilibrato sviluppo degli strumenti di mediazione dell'informazione*, «AIB studi», vol. 55 (2015), n. 1, p. 35-55, n. 2, p. 197-214.

Richardson 2013 = Hillary A. H. Richardson, *Revelations from the literature: how web-scale discovery has already changed us*, «Computers in libraries», 33 (2013), n. 4, p. 12-17.

Salarelli 2012 = Alberto Salarelli, *Introduzione alla scienza dell'informazione*, Milano, Bibliografica, 2012.

Santoro 2011 = Michele Santoro, *Dalla stampa all'open access: percorsi della comunicazione scientifica*, «Culture del testo e del documento», 12 (2011), n. 35, p. 27-73 (disponibile anche in: *Bibliografie, biblioteche e gestione dell'informazione: un omaggio a Francesco Dell'Orso*, «ESB forum», 2016-2017, <http://www.riccardoridi.it/esb/fdo2016-santoro.htm>).

Serrai 2016 = Alfredo Serrai, *In Conrad Gesner l'origine dell'era della informazione*, «Bibliothecae.it», 5 (2016), n. 2, p. 354-357, <https://bibliothecae.unibo.it/article/view/6397>.

Shadbolt *et al.* 2006 = Nigel Shadbolt *et al.*, *The semantic web revised*, «IEEE intelligent sistems», 21 (2006), n. 3, p. 96-101.

Shroff 2013 = Gautam Shroff, *The intelligent Web: search, smart algorithms, and big data*, Oxford, Oxford University Press, 2013.

Smeulders *et al.* 2000 = Arnold W. M. Smeulders *et al.*, *Content-based image retrieval at the end of the early years*, «IEEE Transactions on Pattern Analisys and Machines Intelligence», 22 (2000), n. 12, p. 1349-1380.

Smith *et al.* 2005 = *Handbook of visual communication*, edited by Kenneth L. Smith *et al.*, Mahwah, LEA, 2005.

Solimine 1995 = Giovanni Solimine, *Controllo bibliografico universale*, Roma, AIB, 1995.

Spink - Jansen 2006 = Amanda Spink, Bernard J. Jansen, *Searching multimedia federated content web collections*, «Online information review», 30 (2006), n. 5, p. 485-495.

Spink - Zimmer 2008 = *Web search: multidisciplinary perspectives*, edited by Amanda Spink, Michale Zimmer, Berlin, Springer, 2008.

Stock - Stock 2013 = Wolfgang G. Stock, Mechtild Stock, *Handbook of information science*, Berlin, De Gruyter Saur, 2013.

Svenonius 1994 = Elaine Svenonius, *Access to nonbook materials: the limits of subject indexing for visual and aural languages*, «Journal of the American Society for Information Science», 45 (1994), n. 8, p. 600-606.

Tan - Ngo 2016 = Chun-Chet Tan, Chong-Wah Ngo, *On the use of commonsense ontology for multimedia event recounting*, «International journal of multimedia information retrieval» 5 (2016), n. 2, p. 73-88.

Taube 1955 = Mortimer Taube, *Application of the Uniterm system and the association of ideas to a special library file (Documentation Inc Bethesda Md. Report no. TR-10)*, Belvoir, Defense Technical Information Center, 1955.

Taube - Wooster 1958 = *Information storage and retrieval: theory, systems and devices*, edited by Mortimer Taube, Harold Wooster, New York, Columbia University Press, 1958.

Testoni 2014 = Laura Testoni, *Quali literacy al tempo dei social network?*, «Biblioteche oggi», 32 (2014), n. 4, p. 28-36.

Thomee - Lew 2012 = Bart Thomee, Michael S. Lew, Inte*ractive search in image retrieval: a survey*, «International journal of multimedia information retrieval», 1 (2012), n. 2, p. 71-86.

Thomsett-Scott - Reese 2012 = Beth Thomsett-Scott, Patricia E. Reese, *Academic libraries and discovery tools: a survey of the literature*, «College & undergraduate libraries», 19 (2012), n. 2/4, p. 123-143.

Trombone 2015 = Antonella Trombone, *Il progetto BIBFRAME della Library of Congress: come stanno cambiando i modelli strutturali e comunicativi dei dati bibliografici*, «AIB studi», 55 (2015), n. 2, p. 215-226.

Vivarelli 2016 = Maurizio Vivarelli, *Vedere le informazioni: dati, persone, mediazione documentaria*, in "Convegno AIB CILW 2016. La rinascita delle risorse dell'informazione: granularità, interoperabilità e integrazione dei dati", (Roma, 21 ottobre 2016), <https://www.academia.edu/29341164/Vedere_le_informazioni._Dati_persone_mediazione_documentaria_Convegno_La_rinascita_delle_risorse_dell_informazione_Roma_21_ottobre_2016>.

W3C 2011 = World Wide Web consortium. Library linked data incubator group, *Library linked data incubator group final report*, W3C, 2011, <http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025>.

Weare - Toms - Breeding 2011 = William H. Weare Jr., Sue Toms, Marshall Breeding, *Moving forward: the next-gen catalog and the new discovery tools*, «Library Media Connection», 30 (2011), n. 3, p. 54-57.

Wright 2014 = Alex Wright, *Cataloging the world: Paul Otlet and the birth of the information age*, Oxford, Oxford University Press, 2014.

Xu - Wang 2015 = Lei Xu, Xiaoguang Wang, *Semantic description of cultural digital images: using a hierarchical model and controlled vocabulary*, «D-Lib magazine», 21 (2015), n. 5/6, <http://www.dlib.org/dlib/may15/xu/05xu.html>.

Yang 2012 = Sharon Q. Yang, *Tagging for subject access*, «Computers in libraries», 32 (2012), n. 9, p. 19-23.

Yoshitaka - Ichikawa 1999 = Atsuo Yoshitaka, Tadao Ichikawa, *A survey on content-based retrieval for multimedia databases*, «IEEE Transactions», 11 (1999), p. 81-93.

ABSTRACT

L'articolo discute a confronto le attuali metodologie per la ricerca e la scoperta dell'informazione e delle risorse informative: la ricerca di tipo terminologico e il linguaggio term-based, tipici dell'information retrieval (IR); la ricerca di tipo semantico e l'information discovery, in fase di sviluppo soprattutto tramite il linguaggio dei linked data; la ricerca di tipo semiotico e il linguaggio content-based, sperimentati dal multimedia information retrieval (MIR). È approfondita, quindi, la metodologia semiotica del MIR.

information retrieval; information discovery; multimedia information retrieval; content-based retrieval; discovery tool; linked data; semantic web

Beyond information retrieval: information discovery and multimedia information retrieval The paper compares the current methodologies for search and discovery of information and information resources: terminological search and term-based language, own of information retrieval (IR); semantic search and information discovery, being developed mainly through the language of linked data; semiotic search and content-based language, experienced by multimedia information retrieval (MIR).MIR semiotic methodology is, then, detailed.

information retrieval; *information discovery*; *multimedia information retrieval*; *content-based retrieval*; *discovery tool*; *linked data*; *semantic web*