Diacronie
Studi di Storia Contemporanea

N° 15, 3 | 2013
Spazi, percorsi e memorie

L'archiviazione delle pagine dei quotidiani online

Federico Nanni

Edizione digitale
URL: http://journals.openedition.org/diacronie/418
DOI: 10.4000/diacronie.418
ISSN: 2038-0925

Editore
Association culturelle Diacronie

Notizia bibliografica digitale
Federico Nanni, « L'archiviazione delle pagine dei quotidiani online », *Diacronie* [Online], N° 15, 3 | 2013, documento 2, online dal 01 octobre 2013, consultato il 10 décembre 2020. URL : http://journals.openedition.org/diacronie/418 ; DOI : https://doi.org/10.4000/diacronie.418

Diacronie
Studi di Storia Contemporanea
www.diacronie.it

N. 15 | 3|2013 Spazi, percorsi e memorie

2/

L'archiviazione delle pagine dei quotidiani online

Federico NANNI *

Questo articolo analizza i metodi utilizzati dai siti d'informazione per permettere ai nativi digitali la preservazione e l'accesso nel corso del tempo ai propri contenuti. Prima di tutto si è descritto quali tipologie di documenti sono presenti su tali siti. In secondo luogo, confrontando gli archivi dei principali quotidiani digitali, si ipotizzano due tipi di interventi possibili: un primo volto a migliorare l'interrogazione per "metadati descrittivi" e un secondo incentrato sull'interrogazione full text attraverso strumenti di ricerca semantica. Si è voluta inoltre sottolineare la necessità di preservare queste testimonianze digitali conservandone il più possibile l'integrità. In conclusione, si evidenzia il legame inscindibile tra la ricerca storica sulle fonti native digitali e gli studi di archivistica informatica.

Premessa

La necessità di condividere informazioni è intrinseca all'esistenza stessa di Internet. Quello che prima si faceva su Usenet[1], grazie a strumenti che hanno poi tracciato la nascita dei forum, è stato subito ripreso dal web. I grandi monopoli dell'informazione cartacea sono così approdati, nei primi

[1] Usenet (contrazione inglese di "user network", in italiano "rete utente") è una rete mondiale formata da migliaia di server tra loro interconnessi ognuno dei quali raccoglie gli articoli (o news, o messaggi, o post) che le persone aventi accesso a quel certo server si inviano, in un archivio pubblico e consultabile da tutti gli abbonati, organizzato in gerarchie tematiche e newsgroup flussi di articoli sullo stesso tema (*topic*, o *thread*) (definizione tratta da: JAMES, Vincent, ERIN, Jansen, *Netlingo: The Internet Dictionary*, Netlingo Inc., 2003, p. 391). Tale struttura è passata in secondo piano con l'avvento del World Wide Web ed è lentamente andata in disuso, anche se tuttora è possibile utilizzarne i server i quali custodiscono testimonianze fondamentali dei primi decenni di vita di Internet.

anni Novanta, in rete, presentando inizialmente piattaforme sulle quali era possibile consultare recensioni, elenchi tematici e partecipare a gruppi di discussione, come è evidente se si prendono in considerazione le primissime versioni dei siti web di «The New York Times» o «The Washington Post». Nel corso del decennio successivo l'informazione è stata offerta sulla rete da un insieme sempre più eterogeneo di produttori e nei formati più disparati; questo continua tuttora perché, grazie ai progressi tecnologici, primo fra tutti il continuo potenziamento della banda larga, si è potuta ottenere quella "convergenza al digitale" descritta più volte dal punto di vista teorico[2].

A causa dell'aumento delle tipologie e della quantità di informazioni prodotte, un quotidiano digitale si trova oggigiorno a competere su scala globale prima di tutto con imprese che nei decenni precedenti lavoravano in settori limitrofi (pensiamo nello specifico all'importanza attuale del sito web della CNN o della BBC), ma anche con figure che fino a poco tempo fa non erano neppure presenti nel mondo dell'informazione; faccio riferimento in primo luogo al fenomeno del *citizen-journalism* su piattaforme digitali, quali sono state in un primo tempo i blog e successivamente sono diventati i social network. La competizione che questo tipo genere di giornalismo implica, costringe il sito web di un quotidiano ad aggiornarsi, espandersi e integrare continuamente nuovi strumenti, sia per fare concorrenza preventiva ai propri rivali nel campo dell'informazione digitale, sia soprattutto per poter orientare sulla propria home page sempre più visitatori e ottenere, grazie alle conseguenti visualizzazioni dei *banner*, gli introiti pubblicitari, ovvero quella che continua a essere ancor oggi una delle sue principali fonti di guadagno.

Questi enormi e multiformi produttori di informazioni sono tra i principali punti di riferimento offerti agli utenti sulla rete per tenersi aggiornati sul presente. Che si stia parlando di «The Guardian», de «La Repubblica» o di «The Huffington Post», tali piattaforme sono nodi centrali, dominanti per quanto riguarda la diffusione di notizie. Al di là dell'informazione giornaliera, ognuna di queste piattaforme offre inoltre ai propri utenti strumenti per permettere la consultazione futura di tali notizie, assicurando un servizio utile sia ai propri lettori abituali, sia agli storici che nei prossimi decenni potranno così avere a disposizione utili fonti per le proprie ricerche.

Nelle prossime pagine ho deciso così di analizzare approfonditamente proprio questo tipo di strumenti, per comprendere se sono all'altezza degli scopi che loro stessi si sono preposti; successivamente ho evidenziato alcuni tipi di intervento che reputo

[2] KLINENBERG, Eric, «Convergence: news production in a digital age», in *Annals - AAPSS*, 597, 2005, pp. 48-64.

necessari per migliorarli adeguatamente. Dato che mi concentrerò principalmente sugli archivi dei siti Internet dei quotidiani cartacei, non intendo comunque escludere da questo discorso tutto quell'insieme di piattaforme native digitali, o che esistono parallelamente a un omonimo periodico, impegnate nell'offrire sul web informazione giornaliera, quali possono essere il sito della rivista italiana «Internazionale», o quello del «TIME». Le conclusioni a cui giungerò sono infatti generalmente valide anche per questo tipo di realtà digitali.

1. Studi già esistenti

Per quanto riguarda la preservazione e condivisione in rete del patrimonio giornalistico, diversi autori si sono soffermati ad analizzare la digitalizzazione e conservazione attuata dalle versioni online dei corrispettivi quotidiani cartacei. Tra questi è fondamentale ricordare principalmente i lavori di Nicola Cowen[3] e Brunella Longo[4], che sono stati tra i punti di riferimento per i professori Ernest Abadal e Javier Guallar dell'Università di Barcellona, i quali nel 2009 hanno presentato un modello di analisi metodica[5] di questo particolare tipo di piattaforme.

Per quanto riguarda invece la generica preservazione di materiali digitali, fondamentali sono i contributi sull'argomento di Stefano Vitali[6], Niels Brügger[7] e Jinfang Niu[8]. È inoltre da tenere a mente l'articolata opera di riflessione di Serge Noiret[9], tesa a sottolineare l'importanza dei recenti studi di *digital history,* orientati proprio a venire incontro a queste esigenze.

[3] COWEN, Nicola, «The Future of Broadsheet Newspaper on the World Aslib Proceedings», in *Aslib Proceedings*, 53, 5/2001, pp. 189-200.
[4] LONGO, Brunella, «Gli archivi dei giornali on line», in *Biblioteche oggi*, 1/2006, pp. 9-21.
[5] ABADAL, Ernest, GUALLAR, Javier, «The digital press archives of the leading Spanish online newspapers», in *Information Research*, 15, 1/2009, URL: < http://informationr.net/ir/15-1/paper424.html > [consultato il 9 settembre 2013].
[6] VITALI, Stefano, *Passato digitale: le fonti dello storico nell'era del computer*, Milano, Mondadori, 2004.
[7] BRÜGGER, Niels, *Archiving Websites: General Considerations and Strategies*, Aarhus, The Centre for Internet Research, 2005.
[8] NIU, Jinfang, «An Overview of Web Archiving», in *D-Lib Magazine*, 18, 3/ 2012, URL: < http://dlib.org/dlib/march12/niu/03niu1.html > [consultato il 9 settembre 2013].
[9] NOIRET, Serge, *Storia Digitale: sulle risorse di rete per gli storici*, in *La Macchina del Tempo. Studi di informatica umanistica in onore di Tito Orlandi*, Firenze, Le Lettere, 2011, pp. 201-258.

Infine, per quanto riguarda nello specifico le tecnologie che suggerirò di integrare a tali piattaforme, un intervento molto interessante è quello offerto dal testo *Neptuno: Semantic Web Technologies for a Digital Newspaper Archive*[10].

Prima di affrontare nello specifico il mio caso di studio, ritengo sia necessario tracciare una iniziale e marcata divisione: il progetto dei professori Abadal e Guallar, mio principale punto di riferimento nello sviluppo di un modello valutativo, è stato incentrato su un'analisi di quelle piattaforme che sono solito definire gli "archivi storici" dei quotidiani, ovvero siti web, o sezioni di un sito web, rivolti a offrire l'accesso al giornale cartaceo precedentemente digitalizzato. Il mio studio sarà invece focalizzato sulla disamina degli "archivi digitali"; con questo termine voglio infatti identificare quegli strumenti che permettono all'utente di individuare gli articoli, le foto, i video e tutti i materiali multimediali che sono stati pubblicati, primariamente o soltanto, sulla versione digitale del quotidiano.

2. L'interrogazione del database

Tipicamente ogni sito d'informazione offre due diversi strumenti per accedere al proprio patrimonio archivistico nativo digitale. In un primo caso abbiamo la cosiddetta "ricerca semplice", ovvero l'interrogazione per parole chiave o stringhe di testo; in un secondo caso abbiamo invece la "ricerca avanzata", ovvero l'indagine attraverso descrittori specifici del documento, come possono essere il nome dell'autore, la data di pubblicazione, la sezione, etc. Se la ricerca semplice si ricollega necessariamente a problematiche più complesse, di cui mi occuperò nella seconda parte di questo testo, quella avanzata è invece da subito facilmente studiabile e valutabile.

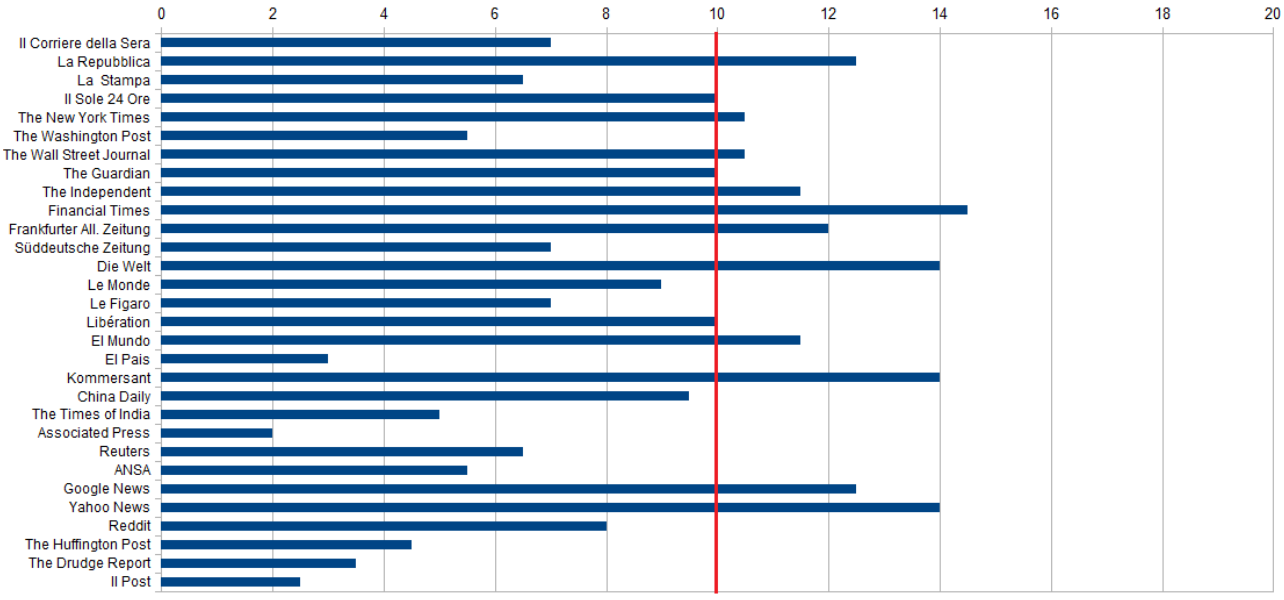
2.1 La ricerca avanzata

Nella mia tesi di laurea[11] ho preso in considerazione trenta tra i più importanti siti d'informazione online a livello mondiale; li ho selezionati principalmente in base all'Alexa Traffic Rank[12], integrandoli ad alcuni casi di studio particolari, utili per

[10] CASTELLS, Pablo, et al., «Neptuno: Semantic Web Technologies for a Digital Newspaper Archive», in *The Semantic Web: Research and Applications,* 3053, 2004, pp. 445-458.
[11] NANNI, Federico, *Metodi per la preservazione e l'accesso ai documenti digitali di quotidiani online*, Tesi di laurea magistrale in Scienze storiche, Università degli Studi di Bologna, Bologna a.a. 2011/2012.
[12] Tale indicatore è sviluppato da Alexa Internet Inc., un'azienda statunitense, sussidiaria di Amazon, che si occupa di statistiche sul traffico in Internet.

comprendere più nello specifico la realtà editoriale digitale italiana. Ho sottoposto ognuno di questi siti web a venti indicatori valutativi, ispirandomi principalmente al modello individuato dai precedentemente citati professori Abadal e Guallar. Ho integrato a tale schema opportune variabili atte a giudicare gli strumenti d'accesso al patrimonio nativo digitale (in particolare ai video, alle foto e agli interventi dei giornalisti sui propri blog). Attraverso questa analisi ho voluto comprendere prima di tutto che tipo di strumenti vengono attualmente offerti dai quotidiani ai lettori durante una ricerca sui loro siti web e se si possono riscontrare *pattern* che ci permettano di classificare questi archivi digitali per tipologie. In secondo luogo, a livello empirico, questo studio mi ha permesso di comprendere quali metadati descrittivi[13] i quotidiani digitali utilizzano generalmente per descrivere e permettere il ritrovamento dei propri documenti digitali.

Per avere una prima, generale, idea dei risultati del mio lavoro ho realizzato due rappresentazioni grafiche di rapida consultazione. Lungo la verticale sono elencati i quotidiani studiati e lungo l'orizzontale è visibile il numero di indicatori che questi hanno soddisfatto.

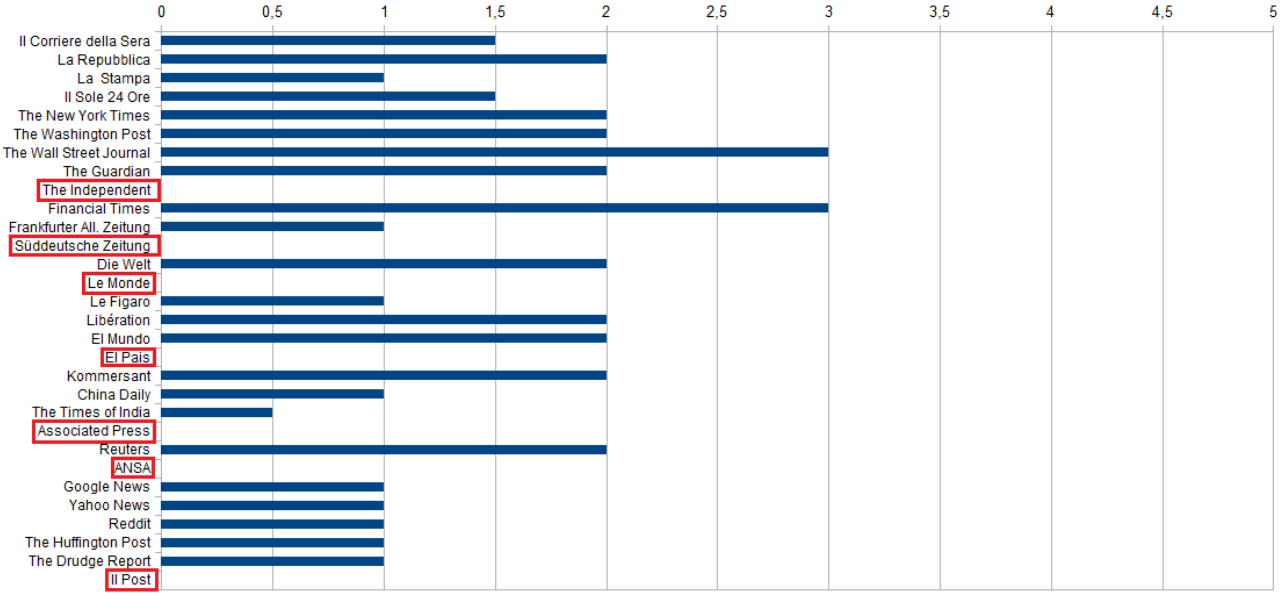


Nella prima figura è facilmente individuabile la conclusione di questa analisi. Come è chiaro dallo schema, vi sono alcuni quotidiani con archivi ben organizzati, in particolare «Financial Times», «Kommersant», «Die Welt» e «La Repubblica». Inoltre,

[13] Un metadato è un'informazione atta a descrivere un insieme di dati; in particolare, per quanto riguarda i metadati descrittivi, questi in archivistica informatica servono specificatamente a permettere il ritrovamento dei documenti.

tra i cosiddetti "aggregatori di notizie", spiccano Yahoo News e Google News, ma più che altro perché sfruttano molto bene i propri motori di ricerca generici, piuttosto che perché hanno realizzato piattaforme espressamente dedicate all'archiviazione delle notizie online. Tra gli archivi meno sviluppati invece sono da evidenziare in negativo il sito dell'«Associated Press», quello di «El Pais», «The Times of India» e soprattutto i siti d'informazione *all digital*, come l'italiano «Il Post», «The Drudge Report» e «The Huffington Post».

Se prestiamo attenzione alla seconda rappresentazione grafica, la situazione si manifesta ancora più complessa e preoccupante.



In questo caso ho voluto evidenziare i risultati riscontrati soltanto per quanto riguarda gli indicatori relativi alla preservazione dei materiali nativi digitali. Per la precisione, ho controllato la presenza di mezzi per la ricerca in maniera specifica di immagini, di video, etc. Come è evidente, praticamente nessun quotidiano offre strumenti efficaci, anzi esistono diversi siti web che non dispongono di alcuna tecnologia di questo tipo. Ciò significa che, ad esempio, se io voglio individuare tutti i video pubblicati da «The Independent» durante maggio 2011, o le gallerie fotografiche dedicate da «Le Monde» alle elezioni presidenziali francesi del 2012, attraverso la "ricerca avanzata" non posso farlo.

In conclusione, da questo tipo di studio appare evidente come nella maggior parte dei casi i quotidiani digitali non offrono assolutamente strumenti in grado di trovare con precisione i materiali presenti nei propri archivi. Osservando i risultati ottenuti non è inoltre possibile tracciare conclusioni significative per nazione, per ambito

(quotidiani generalisti, economici etc.) o per orientamento politico delle testate, poiché i casi più interessanti si trovano isolati. Ad esempio, prendendo in considerazione solo gli indicatori relativi al patrimonio *born digital*, l'attenzione rivolta a questo tipo di materiali per quanto riguarda i siti web dei quotidiani generalisti è particolarmente forte negli Stati Uniti e in Gran Bretagna, ma questo non implica che tali testate rivolgano un'attenzione maggiore alla completa impostazione dell'archivio, tant'è che siti come quello di «The Washington Post» assicurano un buon trattamento dei documenti nativi digitali, ma trascurano l'impostazione generale della piattaforma di ricerca. Infine, escludendo «The Associated Press», che offre un archivio privo di qualsiasi strumento d'interrogazione, gli altri siti d'informazione nativi digitali seguono tutti un modello piuttosto simile, molto poco sviluppato e rivolto essenzialmente a offrire articoli correlati genericamente alla *query*.

2.2 Un primo tipo di intervento

Visto l'attuale stato degli strumenti d'accesso ai materiali digitali pubblicati dai siti d'informazione, quella che propongo è una prima forma d'intervento che prenda ispirazione dagli esempi esistenti meglio realizzati, nello specifico faccio riferimento alla "ricerca avanzata" offerta da «Financial Times» e da «La Repubblica». A questi integrerei alcuni descrittori che permettano l'interrogazione specifica per i materiali nativi digitali, in particolare appunto le foto, i video e i documenti testuali presenti solamente nelle edizioni online. Prestando continua attenzione nei confronti delle esigenze dei propri utenti, garantendo i cosiddetti servizi di "aiuto" e mantenendo un contatto diretto che permetta di offrire un sito web, e nello specifico degli strumenti di ricerca all'altezza, si potrà offrire un primo efficace strumento d'interrogazione del database, utile non soltanto per i propri lettori abituali, ma anche per gli storici che potranno così accedere più rapidamente a fonti digitali basilari per i propri studi.

2.3 Una preservazione completa

L'articolato processo di "conversione al digitale" di un quotidiano cartaceo è solitamente composto da alcune fasi. Prima di tutto abbiamo la digitalizzazione delle pagine attraverso uno scanner OCR, in seguito la descrizione del documento attraverso i metadati, i quali, come accennato in precedenza, permetteranno tra le altre cose l'individuazione dello stesso durante l'interrogazione dell'archivio, e infine la realizzazione di una piattaforma online che permetta parallelamente la lettura del

documento testuale e la sua visualizzazione per come era stato proposto sul cartaceo. Questo accade perché, come sappiamo, la conservazione di un documento non si limita a rendere sempre accessibile il suo contenuto, ma coinvolge anche la preservazione di quell'insieme variegato di informazioni che si possono ottenere attraverso l'osservazione dell'immagine digitalizzata, come possono essere le altre notizie presenti sulla stessa pagina, le pubblicità a cui il testo è affiancato, le scelte di impaginazione, etc.

Prendendo spunto da questo processo, ritengo che tale tipo di approccio debba essere riproposto in maniera equivalente anche per le notizie presenti sui quotidiani digitali, e quindi penso sia necessario affiancare a una solida preservazione dei contenuti anche una serie di *snap-shot*[14] – sul modello di quelli offerti dall'Internet Archive – dell'home page del quotidiano per tutta la durata della presenza dell'articolo sulla stessa. Un primo progetto al quale ci si potrebbe ispirare è quello proposto da «The Drudge Report». Nello specifico questo sito web, una sorta di complesso aggregatore di notizie relative alla politica interna statunitense con un solido orientamento repubblicano, affianca a un archivio molto poco implementato e di difficoltosa interrogazione una completa raccolta di fermi immagine della propria home page dal novembre 2001 a oggi.

Prendendo spunto da queste piattaforme e analizzando i cosiddetti *today's paper* e *day in a page* offerti da «The Guardian», «The New York Times» e «The Independent», si potrebbe pensare a solide integrazioni dei due modelli per offrire ai lettori un'efficace visualizzazione grafica della produzione giornaliera di informazione digitale.

3. L'interrogazione full-text

Come sappiamo, l'informatica ha permesso all'archivistica di superare gli strumenti di catalogazione e ricerca tradizionali, ovvero le schede bibliografiche che da sempre hanno affiancato il documento, descrivendolo e facilitandone l'individuazione all'interno dell'archivio. Con l'avvento del World Wide Web si sono infatti affermati motori di ricerca in grado di operare interrogazioni sull'intero contenuto testuale delle pagine, e non soltanto sui metadati, ovvero le informazioni associate[15]. Dai primi, Web

[14] Con *snap-shot*, o "fermo immagine" intendo un'immagine interattiva che permetta la preservazione completa di una pagina web nel corso del tempo, dei suoi contenuti multimediali e di tutti i link ipertestuali.

[15] Esistono diversi volumi che hanno ripercorso la nascita e l'affermazione dei motori di ricerca come principale strumento per la navigazione sul web. Uno dei più completi sull'argomento è la

Crawler e Altavista, passando per l'italiano Virgilio, fino ad arrivare a Google, Yahoo e il recente Bing, tali strumenti hanno permesso in questi vent'anni di interrogare una mole di dati gigantesca, da loro primariamente analizzata e gerarchizzata, offrendo all'utente una lista ordinata di risultati. Questo tipo di sistemi, come è noto, sono utilizzabili attraverso stringhe di testo o specifiche parole chiave e hanno il limite intrinseco di fornirci risultati senza comprendere il significato della nostra interrogazione e senza neppure capire il contesto nel quale un determinato termine è utilizzato all'interno di un documento.

Tale tipo di riflessioni riportano alla mente le problematiche individuate da Tim Berners-Lee in alcuni sui celebri interventi di inizio XXI secolo[16]; in questi testi lo studioso britannico sottolineava anche come, per rispondere alle difficoltà inerenti la ricerca digitale su moli infinitamente grandi di dati, sarebbe necessario poter offrire a agenti software "intelligenti" un dominio di sapere strutturato, dove i concetti sono disambiguati e i riferimenti precisi, in modo tale che sia infine presentato all'utente esattamente ciò che cerca.

Negli ultimi anni Google ha investito molto su questo settore, acquisendo in primo luogo la Applied Semantic nel 2003 (grazie alla quale ha potuto successivamente lanciare la tecnologia Ad Sense), Orion nel 2006 e infine Metaweb nel 2010.

Ed è proprio per merito di quest'ultima acquisizione che la stessa Google ha avuto la possibilità di integrare ai propri strumenti Freebase, un database di conoscenza strutturata che ha posto le fondamenta di Knowledge Graph, il servizio offerto dal motore di ricerca in tutto il mondo, a partire dalla seconda metà del 2012. Google, grazie a questo strumento, permette oggi agli utenti di raffinare e disambiguare le interrogazioni e affianca ai risultati informazioni aggiuntive. Anche se la stampa mondiale ha annunciato tale integrazione come un deciso passo verso il *semantic*

seconda edizione del testo di LEVENE, Mark, *An Introduction to Search Engines and Web Navigation*, Hoboken, Wiley, 2010. Un'altra analisi interessante è sicuramente il testo di IPPOLITA, *Luci e ombre di Google: futuro e passato dell'industria dei metadati*, Milano, Feltrinelli, 2007. Inoltre sul sito WordStream è consultabile un'infografica molto ben realizzata sulla storia dei motori di ricerca, raggiungibile all'URL:
< http://www.wordstream.com/articles/internet-search-engines-history > [consultato il 9 settembre 2013].
Infine, per un approfondimento su uno dei temi più importanti dell'attuale dibattito sull'evoluzione dei motori di ricerca è interessante la seconda edizione di: LIU, Bing, *Web Data Mining*, Berlin, Springer, 2011.

[16] BERNERS-LEE, Tim, HENDLER, James, LASSILA, Ora, «The Semantic Web», in *Scientific American Magazine,* 17 maggio 2001, URL:
< http://www.cs.umd.edu/~golbeck/LBSC690/SemanticWeb.html > [consultato il 9 settembre 2013].

web[17], quello che Google si limita a offrire oggigiorno, cioè disambiguare le *query* e affiancare i risultati ad altre informazioni statisticamente utili agli utenti, è molto poco rispetto a quello che, attraverso il suo enorme patrimonio di informazioni raccolte su base "entità-relazione", potrebbe realizzare.

Come è comunque evidente, le riflessioni inerenti il cosiddetto *semantic web*, sono strettamente collegate alla necessità di interventi per migliorare qualunque tipo di motore di ricerca, compresi ovviamente anche gli strumenti di interrogazione *full-text* presenti sui quotidiani online.

Dato che, in primo luogo, esistono diversi approcci funzionali per ottenere un dominio di sapere strutturato, e quindi più facilmente analizzabile dalla macchina, nei prossimi paragrafi mi concentrerò su due diversi tipi di tecnologie, sottolineandone le caratteristiche principali e le più evidenti differenze.

3.1 La realizzazione di una ontologia in OWL

L'approccio suggerito dal W3C[18], principale punto di riferimento per quanto riguarda le riflessioni migliorative nei confronti del World Wide Web, per la concettualizzazione di un dominio di conoscenza è la realizzazione di una architettura a livelli, strutturata sul linguaggio di *markup* XML e sull'individuazione di URI[19] univoci, al fine di evitare problemi di ambiguità.

Nello specifico, questo significa che sarà necessario descrivere la conoscenza all'interno dei quotidiani digitali e rendere chiari i riferimenti nei testi e le relazioni tra gli elementi. Ad esempio, se prendiamo in considerazione un articolo nel quale si cita "Enrico Rossi", il presidente della regione Toscana, questo deve essere reso comprensibile alla macchina. Per tale motivo, quello che viene suggerito dal W3C è appunto utilizzare stringhe identificative univoche per descrivere i soggetti, le quali permetteranno inoltre al computer di "mantenere vivi" i riferimenti ipertestuali, anche se le risorse alle quali puntavano cambieranno collocazione. In parole povere questo significa per l'appunto che l'URI sarà così svincolato dall'URL, l'indirizzo web.

17 Per un approfondimento sul web semantico un punto di partenza importante è sicuramente il sito <http://www.websemantico.org/>, che raccoglie un insieme molto variegato di contributi e riflessioni sul tema. Inoltre la professoressa Francesca Tomasi dell'Università di Bologna ha dedicato all'argomento un capitolo del suo manuale: *Metodologie informatiche e discipline umanistiche*, Roma, Carocci, 2008, capitolo VIII.

18 Il World Wide Web Consortium è una organizzazione non governativa internazionale che si occupa, tra le altre cose, di sfruttare e sviluppare tutte le potenzialità del Web.

19 Con URI si intendono stringhe identificative univoche per le risorse digitali.

Dato che l'enciclopedia Wikipedia è un punto di riferimento imprescindibile per identificare in maniera univoca i soggetti, ad esempio nel nostro caso Enrico Rossi è legato indissolubilmente alla pagina <http://it.wikipedia.org/wiki/Enrico_Rossi>, si sta lavorando da anni per estrarre conoscenza strutturata da Wikipedia stessa al fine di rilasciarla online come *linked open data*[20], rendendola così utilizzabile per qualunque tipo di riferimento univoco. Nello specifico, il progetto internazionale DbPedia[21] si pone proprio questo obiettivo; la questione rimane complessa se i soggetti non hanno una propria pagina Wikipedia, in tal caso saranno necessari URI identificati alternativi.

Per quanto riguarda invece l'architettura vera e propria, questa si focalizza principalmente su RDF, strumento base per la codifica, lo scambio e il riutilizzo di metadati strutturati[22]. RDF non descrive la semantica, fornisce piuttosto una base comune per poterla esprimere attraverso asserzioni che rendano esplicite le relazioni. Quello che si ottiene nello specifico è l'associazione tra una risorsa (individuata da un URI) e una proprietà. Tale legame è espresso attraverso una tripla <risorsa, proprietà, valore>. L'affermazione diventa così definitivamente non ambigua e sarà utilizzabile anche da una applicazione autonoma. Un esempio può essere un testo che parla dell'elezione di Barack Obama a presidente degli Stati Uniti: prima di tutto chiariremo risorsa e valore con un URI, quindi espliciteremo la relazione.
Nello specifico possiamo così dire che la proprietà "presidente degli Stati Uniti":
«https://it.wikipedia.org/wiki/Presidente_degli_Stati_Uniti_d'America»
vale
«https://it.wikipedia.org/wiki/Barack_Obama»
ovvero che:
<rdf:Description rdf:about=
https://it.wikipedia.org/wiki/Presidente_degli_Stati_Uniti_d'America
<ns:PresidenteUSA>https://it.wikipedia.org/wiki/Barack_Obama</ns:PresidenteUSA>
</rdf:Description>

Infine, attraverso il linguaggio OWL, potremo definire un vocabolario caratterizzato da classi, proprietà e relazioni, ma sarà inoltre offerta la possibilità di aumentare la tipologia d'inferenze che il software è in grado di compiere, come ad esempio

[20] Con questo termine faccio riferimento al concetto presentato da Tim Berners-Lee nel 2006, URL: < http://www.w3.org/DesignIssues/LinkedData.html > [consultato il 9 settembre 2013].
[21] A questo indirizzo è consultabile la versione italiana,URL: < http://it.dbpedia.org/ > [consultato il 9 settembre 2013].
[22]Per un approfondimento si veda all'URL:
< http://www.websemantico.org/articoli/approcciwebsemantico.php#quattro > [consultato il 9 settembre 2013].

riconoscere che due parti di un documento descrivono, anche se con termini diversi, la stessa realtà o ancora rendere due classi disgiunte. In questo modo avremo strutturato la concettualizzazione del dominio studiato, ovvero realizzato quella che in informatica viene chiamata una ontologia.

Ciò che prima di tutto permettono le ontologie è recuperare documenti attraverso l'elaborazione di *query* complesse e di tipo semantico, cioè fondate sul significato. Partendo da concetti semplici sarà quindi possibile raffinare la ricerca attraverso asserzioni composte da soggetto, predicato e oggetto. Si potrà così arrivare finalmente a risolvere uno dei principali problemi legati all'utilizzo dei motori di ricerca. Per quanto riguarda la progettazione pratica di ontologie, online esistono diversi strumenti che ne permettono, con semplicità, la realizzazione, tra i quali il più noto è sicuramente Protegé[23].

3.2 La rappresentazione della conoscenza attraverso Topic Map

Prendendo ispirazione dalle linee guida tracciate dal W3C, nell'ultimo decennio si sono consolidati diversi tipi di tecnologie atti a permettere all'utente di realizzare rappresentazioni complesse e articolate di un dominio di conoscenza. Tra i tanti esempi disponibili Topic Map[24], standard ISO/IEC 13250:2003 per la rappresentazione del sapere, si è rivelato essere nel corso dell'ultimo decennio uno dei più interessanti. Originato da riflessioni dei primi anni Novanta, la sua implementazione ha portato a conclusioni simili a quelle di RDF, ma con maggiori potenzialità di sviluppo per quello che concerne questo caso di studio e inoltre una più marcata astrazione semantica.

La nascita delle Topic Maps è complessa. Una data di riferimento potrebbe essere il 1991, quando il Davenport Group decise di creare uno standard SGML DTD[25] per la documentazione dei software. La flessibilità di questa tecnologia, che si è quindi evoluta da sistemi per la creazione e fusione di indici analitici di manuali tecnici, permette oggi all'utente di superare alcuni problemi nella realizzazione delle ontologie. Le Topic Maps si sono infatti rivelate essere molto più che un mero strumento di navigazione del

[23] L'editor open source di ontologie Protegé è scaricabile a questa pagina, URL: < http://protege.stanford.edu/ > [consultato il 9 settembre 2013].
[24] Per un approfondimento: MESCHINI, Federico, «Le mappe topiche. Come imparai a non preoccuparmi e ad amare i metadati», in *Bollettino AIB*, 1/2005, pp. 59-72 .
[25] Lo Standard Generalized Markup Language (SGML), è un metalinguaggio, standard ISO 8879:1986, che ha come scopo definire linguaggi da utilizzare per la stesura di testi destinati a essere trasmessi e archiviati rendendoli "comprensibili" al computer. Principale funzione di SGML è appunto la stesura di testi chiamati Document Type Definition (DTD), ciascuno dei quali definisce in modo rigoroso la struttura logica che devono avere i documenti di un determinato tipo.

dominio di sapere descritto, anche se rimangono allo stesso tempo uno dei diversi possibili modelli per rappresentare la conoscenza, alternativi a OWL.

La caratteristica principale di questa tecnologia, rispetto a un'ontologia costruita come descritto nei paragrafi precedenti, è avvicinare il tipo di associazioni costruite al ragionamento umano. Le Topic Maps si pongono infatti a un livello superiore di astrazione semantica rispetto a RDF, presentando *topics*, associazioni e occorrenze quando RDF propone soltanto due argomenti collegati tra loro, e permettendo relazioni complesse tra grandi numeri di nodi quando RDF si limita a permettere triplette. Allo stesso tempo, come spiegato da Lars Marius Garshol[26] sul sito Ontopia.net, convertire i dati da una tecnologia all'altra non è un procedimento complesso, sono necessarie soprattutto mappature dichiarative e un vocabolario specifico condiviso; per questo lo studioso ritiene giusto affermare che sia possibile fare collaborare tra loro RDF e le Topic Maps.

Nello specifico, realizzare una Topic Map significa identificare i *topics*, e con questo termine si fa riferimento a qualunque tipo di concetto immaginabile come un "nodo" di una rete, classificarli all'interno di categorie (*topic type*), esprimere associazioni tra questi nodi e le relative occorrenze, ovvero le relazioni tra la mappa e i materiali digitali veri e propri.

Quello che concede questa tecnologia è una gestione più chiara (nodi-associazioni) rispetto alla logica di un database relazionale integrato a una strutturazione RDF/OWL. Per quanto riguarda il nostro esempio precedente, si potranno esprimere facilmente relazioni complesse, come possono essere i ruoli politici dei soggetti nel corso del tempo (Barack Obama è stato senatore dell'Illinois prima di diventare presidente degli Stati Uniti) e parallelamente il legame esistente tra soggetti che hanno ricoperto la stessa carica in periodi differenti.

Come per le ontologie in OWL, anche per Topic Map esistono dei software che permettono all'utente di realizzare facilmente delle rappresentazioni di un determinato dominio, tra questi l'applicazione di riferimento è sicuramente Ontopia[27], uno strumento *open source* per sviluppare applicazioni basate su mappe topiche.

[26]GARSHOL, Lars Marius, *Living with topic maps and RDF*, Ontopia, 2003, URL: < http://www.ontopia.net/topicmaps/materials/tmrdf.html > [consultato il 9 settembre 2013].
[27] Ontopia è scaricabile gratuitamente a questo link, URL: < http://www.ontopia.net/ > [consultato il 9 settembre 2013].

3.4 Mappare il web

Come è evidente, il web, e nello specifico il sito di un quotidiano d'informazione, è costituito da un insieme enorme e estremamente variegato di documenti non strutturati. Gli strumenti per la rappresentazione della conoscenza, sia quelli proposti dal W3C che quelli sviluppatisi parallelamente, come Topic Map, permettono di rendere tali materiali processabili da agenti software, dopo averli inseriti all'interno di un dominio di sapere. Tutto ciò è già oggi dimostrato da diversi progetti specifici, tra i quali uno dei più noti è l'Italian Opera Topic Map[28]. Il sito in questione genera, una volta interrogato il suo database, una serie di pagine web per rendere non ambigua la nostra ricerca e offre link e riferimenti per ottenere maggiori informazioni sugli argomenti cercati.

Però, se si desidera prendere questo tipo di progetti come modello d'ispirazione e si vuole realizzare un intervento concreto di strutturazione del dominio di sapere custodito sul sito web di un quotidiano, la situazione diviene molto più complessa. Il problema principale, di fronte al quale necessariamente ci si verrà a trovare, è dato essenzialmente dall'enorme numero di documenti presenti su questo tipo di siti web, i quali sono da studiare, strutturare e inserire, una istanza dopo l'altra, all'interno della rappresentazione della conoscenza. Un lavoro di questo tipo, se prendiamo in considerazione un soggetto particolare caratterizzato da una redazione ridotta e una discreta produzione giornaliera di materiali digitali, come il quotidiano online «Il Post», si rivelerà già molto difficoltoso da mettere in pratica. Infatti questo sito d'informazione, nato nell'aprile 2010, offre abitualmente sulla propria home page 30 materiali ogni giorno, tra news, gallerie fotografiche e video; questo significa che nel suo archivio sono già presenti più di 30.000 documenti non strutturati. Se queste sono le cifre relative a un quotidiano digitale con una redazione formata principalmente da 4 giornalisti, quanti documenti saranno presenti sul sito web di «The New York Times», il quale è online dal 1994?

3.5 Come intervenire

Di fronte a questa complessa problematica ritengo che il punto di partenza migliore sia rivolgere la propria attenzione all'integrazione tra il *semantic web* e gli strumenti di *web mining*, originati da ricerche e studi consolidatisi negli ultimi anni. Tale tipo di

[28] Tale Topic Map è consultabile alla pagina, URL:
< http://www.ontopia.net/operamap/index.jsp > [consultato il 9 settembre 2013].

approccio è stato inizialmente proposto da Gerd Stumme, Andreas Hotho e Bettina Berendt in un paper[29] proprio dal titolo *Semantic Web Mining*. Il punto di contatto tra questi due campi di studio potrebbe così portare a ottenere le tecnologie necessarie a rispondere efficacemente alle problematiche precedentemente descritte.

Quello che infatti è necessario sono degli agenti software capaci di analizzare i testi, automaticamente individuare i temi principali e estrarre così i concetti; per certi versi quello che già oggi permettono tecnologie come la *latent semantic analysis*[30]. Le informazioni ottenute è fondamentale che vengano successivamente inserite, sempre automaticamente, e messe in relazione ad altre, all'interno di una rappresentazione strutturata della conoscenza. Soltanto in questo modo sarà in futuro possibile per un utente interrogare l'archivio di un quotidiano online e ricevere una risposta precisa e pertinente.

Ottenere un risultato di questo tipo necessità di avviare un percorso di ricerca lungo e articolato fondato su collaborazioni interdisciplinari sempre più intense, capaci di coinvolgere archivisti informatici, studiosi di linguistica computazionale e esperti di *data mining*. Tuttavia, ritengo sia l'approccio capace di dare finalmente i risultati sperati.

4. Conclusioni e prospettive future

L'obiettivo di un articolo di questo tipo era triplice: prima di tutto volevo evidenziare come gli attuali siti web dei quotidiani offrano un insieme molto variegato di materiali, disponibili istantaneamente all'utente per ottenere informazioni in *real-time*, ma molto più difficili da raggiungere se gli stessi non sono più presenti in home page. Archiviare e permettere l'accesso a questi documenti è, di conseguenza, un tipo di servizio importante da offrire non solo al lettore "abituale" di un quotidiano digitale, ma si rivela fondamentale soprattutto nei confronti dello storico, il quale potrà avere consultare testimonianze che esistono solamente in forma digitale.

[29] BERENDT, Bettina, HOTHO Andreas, STUMME, Gerd, «Semantic Web Mining: State of the art and future directions», in *Web Semantics: Science, Services and Agents on the World Wide Web*, 4/2006, pp. 124-143.

[30] La *latent semantic analysis* è una tecnica di *natural language processing* (l'elaborazione del linguaggio naturale) che, studiando le relazioni che esistono tra un insieme di documenti e i termini che questi contengono, mette in evidenza un elenco di concetti, collegati ai termini che compaiono più di frequente all'interno dei testi. Uno degli obiettivi della *latent semantic analysis* è quindi quello di provare a dedurre, con strumenti statistici e relazionali, il contesto generale basandosi sul confronto tra i termini.

In secondo luogo, era mia intenzione sottolineare l'importanza di realizzare un intervento migliorativo nei confronti degli strumenti di ricerca *full-text* presenti sul World Wide Web, e nello specifico proprio per quanto riguarda quelli presenti sui quotidiani digitali. Ho desiderato così evidenziare quanto siano stati importanti i passi avanti ottenuti dagli studiosi del *semantic web* in questo primo decennio e ho voluto indicare una possibile direzione futura. Ho inoltre deciso di analizzare i quotidiani online in quanto questi sono per loro natura piattaforme ricche di informazioni, ma il mio studio può ovviamente essere esteso a diversi tipi di fonti storiche native digitali presenti in rete (e-books, siti accademici, siti tematici, blog, etc.).

L'ultimo obiettivo di questo lavoro era enfatizzare l'importanza basilare dell'archivistica informatica per quegli studi che intendono concentrarsi sull'analisi del passato del World Wide Web[31], o anche soltanto per quegli lavori che vogliono utilizzare contenuti nativi digitali, consultabili in rete, come fonti per una propria ricerca. Quindi referenti di questo articolo non sono soltanto i produttori di documenti digitali, che dovranno comprendere l'importanza della preservazione delle informazioni native digitali o gli esperti dei diversi campi coinvolti nello sviluppo di nuovi strumenti di ricerca testuale, ma sono soprattutto gli "storici digitali" che dovranno trovarsi pronti ad affrontare questa nuova sfida.

4.1 Studiare il web del passato

Sull'importanza di considerare la *web history* come un campo di studi vero e proprio, nato dalla convergenza tra gli studi di *digital history* e gli *Internet studies*, ha pubblicato diverse riflessioni[32] il professor Niels Brügger, direttore del Centro di Internet Studies dell'Università di Aarhus. Prima di procedere evidenziando quali sono le caratteristiche principali di questo recentissimo settore e quanto la ricerca in questo ambito dipenda dall'evoluzione dell'archivistica informatica, è opportuno chiarire le principali caratteristiche dei campi di studio ai quali è più strettamente legato.

[31] Il testo di BETTINI, Andrea, *Giornali.it: La storia dei siti internet dei principali quotidiani italiani*, Catania, ED.IT., 2006, è stato un punto di riferimento fondamentale per il primo capitolo della mia tesi di laurea. Per l'argomento trattato, l'impostazione del lavoro e l'utilizzo di fonti native digitali, può essere considerato uno dei primi esempi italiani di "web history".

[32] In questi ultimi paragrafi farò in particolare riferimento a tre suoi lavori: la raccolta di saggi BRÜGGER, Niels, *Web History,* Aarhus, Niels Brügger ed., 2010; il saggio ID., «Web historiography and Internet Studies: Challenges and perspectives», in *New Media Society*, 2012; e soprattutto il saggio ID., «When the Present Web is Later the Past: Web Historiography, Digital History and Internet Studies», in *Historical Social Research,* 37, 4/2012, pp. 102-117.

Innanzitutto con il termine *digital history*[33] si intende[34] lo sviluppo e l'applicazione nella ricerca storica di nuovi strumenti tecnologici per trovare, manipolare, analizzare testimonianze digitalizzate e condividerne i risultati. Risulta di conseguenza evidente come, in questo approccio, la rete sia considerata principalmente uno strumento al servizio dello storico, e non un soggetto vero e proprio su cui concentrare le proprie riflessioni.

Parallelamente, gli *Internet studies*[35] si sono affermati nel corso del decennio scorso come un campo di studi estremamente interdisciplinare, volto a porre l'attenzione su un grande numero di temi che caratterizzano la realtà digitale. Anche se le riflessioni esposte in tale campo si focalizzano essenzialmente su fonti native digitali, al contrario di quanto accade generalmente per la *digital history*, la tendenza comune in questo tipo di ricerche è di concentrare l'attenzione sul "presente" del web, piuttosto che sul suo "passato"[36].

Per questo motivo Brügger individua nella *web history* un punto di contatto tra questi due campi: ci si concentrerà infatti su fonti native digitali presenti in rete (come accade quindi negli *Internet studies*) e, parallelamente, si parteciperà alla realizzazione degli strumenti tecnologici necessari per preservarle, individuarle e analizzarle approfonditamente (tali approcci sono studiati in diversi settori delle *digital humanities*, tra i quali per l'appunto la *digital history* e l'archivistica informatica).

Di conseguenza gli studi di storia del web dei prossimi decenni potrebbero mutare radicalmente il rapporto esistente tra lo storico e le fonti su cui basa le proprie ricerche. In primo luogo infatti, come ho più volte sottolineato, lo "storico del web" sarà sempre più dipendente dai progressi in campo archivistico, sia per quanto riguarda la preservazione delle fonti che per il successivo reperimento delle stesse. In secondo luogo quegli stessi documenti sui quali lo storico concentrerà i propri studi saranno

[33] Il testo di riferimento per la digital history è sicuramente: COHEN, Daniel e ROSENZWEIG, Roy, *Digital History: A Guide to Gathering, Preserving and Presenting the Past on the Web*, University of Pennsylvania Press, 2005. Tuttavia esistono molteplici applicazioni delle riflessioni di digital history e svariate analisi. Tra queste è importante ricordare: NOIRET, Serge, «"Public History" e "Storia Pubblica" nella rete», in *Ricerche storiche*, 2-3/39, 2009; la lezione tenuta da Frédéric Kaplan alla Summer School in Digital Humanities di Berna a giugno 2013, focalizzata sulla creazione di ricostruzioni digitali delle realtà urbane del passato, consultabile a questo link: URL: < http://vimeo.com/70760197 > [consultato il 9 settembre 2013]; e, infine, il numero 10-2 della rivista di storia contemporanea Diacronie, URL:
< http://www.studistorici.com/dossier/n-10-giugno-2012/ > [consultato il 9 settembre 2013].

[34] Le definizioni a cui faccio riferimento in questi paragrafi sono tratte dai lavori di Niels Brügger citati in precedenza, in particolare da *When the Present Web is Later the Past*, cit.

[35] Per approfondire gli Internet Studies: WELLMAN, Barry, «The Three Ages of Internet Studies: Ten, Five and Zero Years Ago», in *New Media & Society*, 6, 1/2004, pp. 123-129.

[36] Questo tema ritorna in entrambi i saggi di Brügger citati in precedenza, in particolare le pagine 106-107 di *When the Present Web is Later the Past*, cit.

reborn-digital material[37], testimonianze "ricostruite", generalmente incomplete o in alcuni casi addirittura troppo complete[38]. Tutto questo dovrà necessariamente generare nuove riflessioni teoriche dedicate al cambiamento del rapporto tra lo storico e le fonti e, parallelamente, all'affidabilità di quest'ultime[39]: in una realtà nella quale la produzione e condivisione di informazioni cresce giorno dopo giorno in maniera esponenziale, queste testimonianze conservate sui *web archives* rimarranno infatti, nella maggior parte dei casi, l'unica traccia di risorse digitali non più consultabili.

Tuttavia, per concludere, oltre che nuove riflessioni teoriche su come ovviamente dovrà mutare la storiografia di fronte a testimonianze principalmente digitali, sarà opportuno pensare a come dovrà parallelamente mutare la figura dello storico per poter affrontare, concretamente, queste problematiche: ritengo allora che partecipare alla realizzazione degli strumenti archivistici che gli saranno sempre più necessari possa essere un primo, importantissimo passo.

[37]Il professor Brügger dedica diversi passaggi dei saggi precedentemente citati al tema, in particolare le pagine 108-109 di *When the Present Web is Later the Past*, cit.
[38]Pensiamo ad esempio a una singola pagina web conservata sull'Internet Archive nella quale i suoi link ipertestuali potrebbero non essere stati mantenuti attivi o le immagini preservate. Allo stesso tempo potremmo però avere molte altre sue copie archiviate, relative ad altri giorni nei quali è stato effettuato lo stesso identico *snapshot*.
[39] Questo sta già accadendo nel campo della digital history; per approfondire consiglio i due saggi di Noiret citati in precedenza.

* L'autore

Federico Nanni è dottorando (Ph.D.) in Science, Cognition and Technology presso l'Università di Bologna. Ha conseguito la Laurea Triennale in Lettere a Bologna nel 2010, con una tesi in Informatica Umanistica sulla digitalizzazione degli atlanti storici. Nel 2013 si è laureato in Scienze Storiche discutendo una tesi in Archivistica Informatica dal titolo *Metodi per la preservazione e l'accesso ai documenti digitali di quotidiani online*. Si occupa principalmente di web history, semantic web e archivistica informatica, con particolare attenzione all'utilizzo di nuove tecnologie nei processi di preservazione, analisi e individuazione delle risorse native digitali.
URL: < http://www.studistorici.com/progett/autori/#Nanni >

Per citare questo articolo:

NANNI, Federico, «L'archiviazione delle pagine dei quotidiani online», *Diacronie. Studi di Storia Contemporanea : Spazi, percorsi e memorie*, 29/10/2013,
URL:< http://www.studistorici.com/2013/10/29/nanni_numero_15/ >

Diacronie Studi di Storia Contemporanea www.diacronie.it

Risorsa digitale indipendente a carattere storiografico. Uscita trimestrale.
redazione.diacronie@hotmail.it

Comitato di redazione: Marco Abram – Jacopo Bassi – Luca Bufarale – Alessandro Cattunar – Elisa Grandi – Deborah Paci – Fausto Pietrancosta – Matteo Tomasoni – Luca Zuccolo