SEMINARIO DI CULTURA DIGITALE A.A. 2015/2016

Il Processo di Digital Curation e il caso dell'Obvil

Il valore aggiunto delle collezioni digitali

Chiara Alzetta

13/09/2016

La seguente relazione tratterà delle questioni legate alla digital curation, cercando anche di darne una definizione che ne spieghi compiti e funzioni. Successivamente si osserverà un caso particolare di digital curation realizzata su un archivio di testi di letteratura francese: qui l'utilizzo di tecniche di NLP ha permesso di creare un archivio più facilmente navigabile e ricco di informazioni.

INTRODUZIONE

Molti sono gli aspetti da tenere in considerazione quando si parla di *digital curation*. Innanzitutto è proprio il termine *curation* quello su cui bisogna soffermarsi per primo: diretto discendente del termine latino *curator*, ovvero 'colui che si prende cura di qualcosa o qualcuno', identifica un concetto che difficilmente riusciamo a racchiudere in un solo vocabolo della nostra lingua, ma che chiaramente fa riferimento ai ruoli e ai compiti che un curatore deve assolvere nel suo mestiere. Come vengono modificati questi compiti con l'aggiunta dell'aggettivo *digital*? Curatori professionali ed enti culturali a lungo hanno dibattuto, e ancora spesso discutono, su questo tema nel tentativo di delineare una linea di comportamento standard per i professionisti e di classificare le attività svolte dagli utenti del Web, i quali, grazie a nuovi strumenti online aperti a tutti, si avvicinano a questa pratica sempre più numerosi ma non sempre consapevolmente[1]. Consapevolissimi invece altri gruppi di ricerca che sfruttano le nuove tecnologie al meglio per fornire agli utenti un'esperienza di navigazione più ricca e completa. Per ragioni di spazio, ma soprattutto per assecondare l'interesse di chi scrive, mi focalizzerò in particolare sull'esempio dell'Obvil, senza però dimenticare che si tratta solamente di un esempio virtuoso fra i molti a disposizione, anche in ambiti differenti.

[1] A dimostrazione del fatto che, al momento, si tratta di una pratica largamente diffusa, sono nati numerosi corsi di laurea in diverse università su questo tema di Stati Uniti e Regno Unito. Si possono citare ad esempio UCL, Johns Hopkins University e KCL.

Curatori e Collezioni Digitali

La figura del curatore era già presente sin dall'antica Roma, ma è solo con la nascita dei musei aperti al pubblico nel XIX secolo che ha assunto l'accezione che oggi siamo abituati a dargli. A partire dagli anni Settanta in poi il curatore è diventato una figura professionale che non solo si dedica alla conservazione e mantenimento, ma anche alla ricerca ed esposizione del valore aggiunto di una collezione di oggetti (Wolff and Mulholland, 2013).

La stessa definizione può essere applicata anche a chi svolge questa pratica in forma digitale? Il curatore delle collezioni di un museo svolge le stesse funzioni di un curatore di una collezione online? E cos'è una collezione online? Molte le domande che le *humanities* si sono poste negli ultimi anni e a cui hanno cercato di dare un'unica definizione che potesse essere valida sempre.

Innanzitutto pare essersi affermata l'equazione 'curare contenuti digitali = attività che si svolge sul Web', o che comunque nel Web ha il suo sbocco. In realtà *digitale* è un aggettivo che possiede un'accezione molto più ampia, dal momento che, come si legge sul Vocabolario Treccani, si definiscono digitali 'apparecchi e dispositivi che trattano grandezze sotto forma numerica', oltre che 'la qualifica delle grandezze trattate da tali dispositivi, e della loro rappresentazione'. Si può capire dunque che il Web teoricamente non è elemento imprescindibile nella pratica dei curatori digitali, tuttavia dalla letteratura in materia si percepisce fortemente che la diffusione di questa pratica è avvenuta prevalentemente attraverso di esso, e per questo vi si è rimasti concettualmente fortemente legati[2]. Una dimostrazione di questo fatto si può vedere benissimo in (Pange and Bonde, 2016) dove vengono passate in rassegna diverse definizioni di "biblioteca digitale" che sono state proposte negli anni a partire dal 1999: in ciascuna di esse si fa sempre riferimento ad un *network* per la diffusione dei contenuti ad una vasta comunità di utenti.

Collezione digitale è quindi un insieme di oggetti in formato digitale. Questa definizione è talmente vasta che può essere applicata a diverse tipologie di prodotti: infatti ricercando 'collezione digitale' su Google si ottengono moltissime pagine di musei, ma soprattutto

[2] In effetti, ancora prima della diffusione capillare di Internet, si ricorderanno i primi esempi di musei virtuali o collezioni digitalizzate che venivano distribuite su CD-ROM.

biblioteche, e dei loro cataloghi digitali, ma anche amatori appassionati ad un certo tema che attorno ad esso hanno creato un sito Web per raccogliere tutto il materiale da loro reperito.

Ad un livello più tecnico invece possiamo distinguere gli oggetti che fanno parte di queste collezioni in due macrocategorie: dati digitali e dati digitalizzati. Per dati digitali si intende quel materiale nato digitale, che non ha nessun equivalente analogico; viceversa i dati digitalizzati sono il risultato di una qualche conversione da un oggetto materiale concreto che è stato trasformato in digitale (Jones and Beagrie, 2001). Questa differenza è soprattutto tecnica, e riguarda in particolare il modo in cui questi prodotti devono essere trattati e preservati. Professionisti specializzati, come gli archeologi digitali, sono nati al solo scopo di garantire che gli oggetti digitali venissero catalogati, classificati e mantenuti allo stesso modo e con la stessa cura degli artefatti analogici: come quest'ultimi, infatti ciascuna categoria necessita che vengano applicate metodologie specifiche (Thibodeau, 2002)[3].

La già citata stretta connessione con Internet ha avuto anche una forte ricaduta sul modo in cui vengono concepite le informazioni, nonché il loro trattamento. Si è sempre detto che Internet è uno strumento democratico, dove si può trovare tutto e il contrario di tutto (Pitteri, 2013), e quindi mai come in questo "*ambiente*" può essere utile l'intervento di curatori che si occupano di organizzare il materiale e le diverse informazioni. In realtà questa esigenza si è avvertita sin dalla nascita Web, proprio a causa di questa sua natura democratica. Internet è sempre stato fortemente legato al concetto di *curation,* ovvero si è affermato come luogo in cui gli utenti creano e condividono contenuti; ma quello che è cambiato è la quantità di dati che vengono condivisi (o ri-condivisi) in ogni momento, oggi ingestibile senza l'ausilio di nuove tecnologie che si occupino di svolgere questa funzione in maniera automatica[4]. Senza contare che il controllo dei contenuti non è svolto da esperti, ma è affidato alla comunità di utenti stessi, e alcuni studiosi sostengono addirittura che essi spesso non capiscano come questo sistema funzioni realmente (Ovadia, 2013, p.60). La popolarità, diffusione e natura stessa di questo sistema ha fatto sì, ad avviso di una parte della comunità scientifica, che il pubblico del Web diventasse vittima di un abbassamento della

[3] Il problema della conservazione dei dati digitali è sentito come molto attuale ed è possibile trovare una vasta letteratura sul tema. Si veda per esempio la Digital Preservation Coalition: http://www.dpconline.org/ (visitato il 31/08/2016).

[4] La gestione dei grandi dati, i cosiddetti *big data*, è materia di studio e ricerca molto vivace, anche perché non si applica solo i dati del Web. Per un approfondimento si consiglia la lettura, fra gli altri, di (Manyika et al., 2011).

qualità dei contenuti (Keen, 2007) o comunque di una sostanziale rivoluzione nel modo in cui vengono concepiti, non necessariamente in peggio (Weinberger, 2014). Allo stesso tempo si sono affermate anche le posizioni opposte, ovvero quelle che sostengono che l'innato desiderio umano di esprimere se stessi, che su Internet e Social Media trova la sua espressione più naturale, abbia invece reso i contenuti disponibili attraverso le collezioni digitali più personalizzati ed espressivi, intendendo per collezioni digitali anche tutte quelle forme amatoriali che ritroviamo sui Social Network[5] (Macek, 2013) (Feinberg et al., 2012) .

Pare opportuno a questo punto provare a fare chiarezza su come si svolga la pratica di *digital curation*, sia essa svolta da amatori o professionisti, e quali nuovi ruoli e reazioni hanno avuto questi ultimi in risposta alla cosiddetta *social curation*.

DIGITAL CURTION/SOCIAL CURATION

Dopo aver chiarito cosa significa essere un curatore, si può passare alla descrizione dei suoi compiti.

L'Oxford Dictionary definisce il processo di *curation* come l'atto di "selezionare, organizzare e prendersi cura degli oggetti (in una collezione o una mostra)" e viene normalmente associato all'attività dei musei e biblioteche che collezionano oggetti e li arricchiscono di significato e valore aggiunto (Beagrie, 2008). Quando si parla però di *digital curation*, si tende invece ad accettare la definizione proposta dal *Digital Curation Centre*[6] (DCC):

> *"Digital curation involves maintaining, preserving and adding value to digital research data throughout its lifecycle".*

Sebbene le due definizioni siano in parte simili, entrambe infatti sottolineano l'importanza di non limitarsi alla semplice raccolta del materiale, le seconda mette in luce un aspetto nuovo e fondamentale, da tenere sempre a mente quando si ha a che fare con prodotti digitali: il Centro lo

---

[5] In un articolo pubblicato sul Time (Grossman, 2006), viene anche accennato ad un grandissimo aspetto completamente nuovo legato al Web 2.0: si tratta dell'enorme ambiente collaborativo che Internet ha permesso di creare permettendogli di raggiungere un'estensione che mai aveva raggiunto prima. Indipendentemente da vantaggi e svantaggi, il fatto di per sé è indubbiamente degno di nota.

[6] http://www.dcc.ac.uk/ (visitato il 30/08/2016)

definisce il **ciclo di vita dei dati**. A differenza delle collezioni analogiche tradizionali, per quelle costituite da dati digitali già si è detto che si deve prestare attenzione anche alle tecnologie con cui si accede ad esse e verificare che, se obsolete, vengano rinnovate. Così il compito del curatore non è più solo quello di raccogliere i metadati relativi all'oggetto e organizzare una mostra che li racconti e valorizzi al meglio, ma è anche di farsi garante dell'accessibilità dei contenuti e renderli disponibili al riuso da parte di altri in ogni momento (Higgins, 2008).

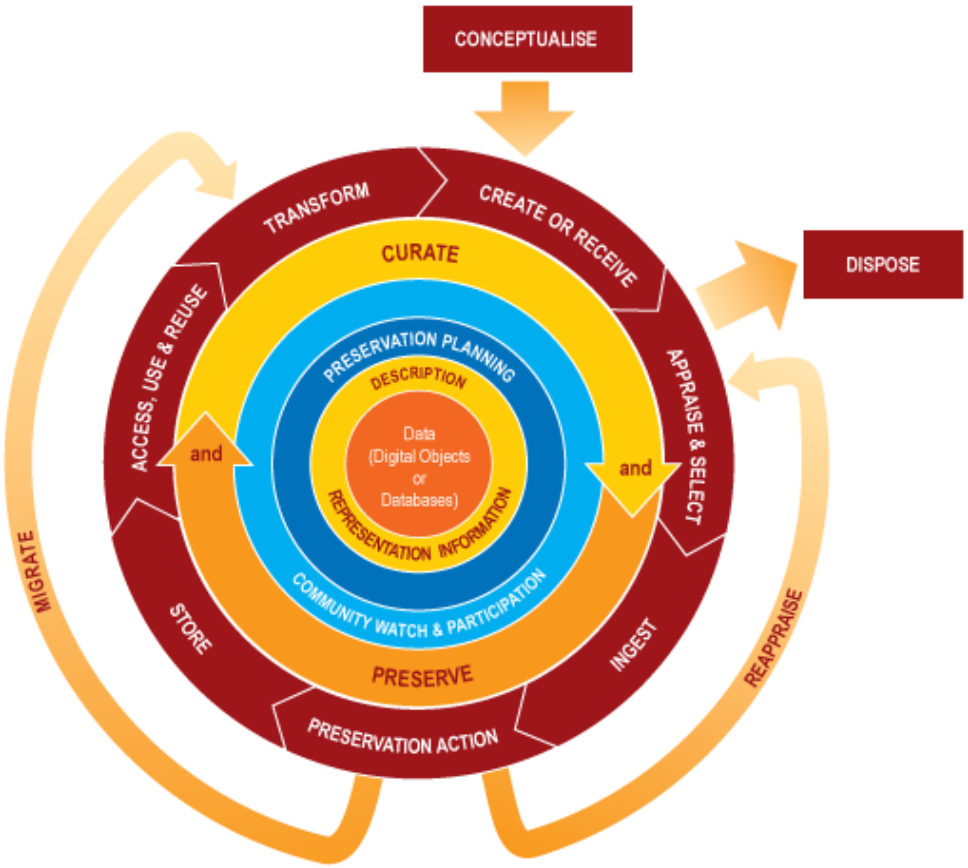

Figura 1 Curation Lifecycle Model del DDC. Spiegazione di tutto il processo si può trovare in (Higgins, 2008).

Prima di parlare di questo innovativo ultimo aspetto, legato allo *sharing* e alla condivisione, è utile fare chiarezza sulla natura del Centro di Digital Curation a cui è stato fatto cenno sopra. Il DCC si presenta come un centro di eccellenza a livello mondiale con lo scopo di fornire aiuto e supporto nella pratica di *digital curation* a individui o enti che desiderano avvicinarsi a questa attività. Il Centro nasce nel 2004, a dimostrazione della reale necessità che si avvertiva allora di avere a disposizione un punto di riferimento che desse delle indicazioni su come comportarsi di fronte alla grande quantità di dati che venivano, e tutt'ora vengono, prodotti su Internet ogni giorno. Nel

2013 si parlava di 571 nuovi siti Web al minuto, 175 milioni di tweet al giorno e 48 ore di nuovi video ogni minuto (Buck, 2013), e da allora il trend non ha fatto che aumentare.

Questo fatto sta avendo un'enorme ricaduta su moltissimi aspetti: sono nate nuove discipline che si occupano di gestire grandi quantità di dati e nuove tecniche per riuscire ad estrarre conoscenza da questa grande moltitudine disorganizzata, ma spesso il singolo utente si ritrova ancora come un piccolo pesce nell'oceano Web alla ricerca di informazioni. Liz Lyon, del Digital Curation Center UK, nei suoi discorsi utilizza spesso una metafora efficace, paragonando i dati di Internet ad "una zuppa: caotici e non sai mai cosa ci puoi trovare dentro". Sistemi di ricerca, come Google, sono stati pensati allo scopo di soddisfare la maggior parte delle nostre esigenze e risolvere questo problema (Pitteri, 2013), ma quando si tratta di ricerche specifiche, magari fatte da accademici e ricercatori, Google non è più sufficiente: difficilmente può gestire tutta la quantità di dati presenti su Internet e non fornisce nessuna garanzia sull'autorità della fonte, per non parlare poi dell'ormai noto problema del cosiddetto *filter bubble*[7].

Consapevoli di questo problema, i GLAMs si sono mobilitati sin dagli albori di Internet nel tentativo di guidare gli utenti e aiutarli a districarsi fra le molte informazioni che il Web offre. GLAM è un acronimo inglese per gallerie, biblioteche, archivi e musei (Smith-Yoshimura, 2011), ovvero quei cosiddetti *"agents of memory"* che si occupano da sempre di preservare il sapere, in particolare quello del passato, e costituiscono un'autorità in fatto di informazione (Robinson, 2012). Proprio in virtù di questa loro autorevolezza, alcuni di questi enti hanno cercato di sfruttare tutte le opportunità del digitale e hanno creato delle collezioni estratte dai loro cataloghi visitabili anche da remoto (Shahed, 2014), realizzando delle vere e proprie mostre virtuali parallele rispetto a quelle 'analogiche', come è possibile vedere sui siti dei maggiori musei del mondo[8].

Bisogna comunque far presente che ancora oggi non tutti gli enti sono pronti ad accogliere le nuove possibilità offerte dal digitale: a volte mancano le tecnologie, altre volte i fondi per finanziare le infrastrutture e il personale necessario a garantire la buona qualità e la possibilità di

---

[7] Questo aspetto ha indubbiamente una maggiore ricaduta sul diritto di ciascun utente di avere accesso alle informazioni piuttosto che sull'attività di *curation*, tuttavia ha intuitivamente delle conseguenze visibili anche per quest'ultima. Per maggiori informazioni riguardo questo argomento si consiglia di consultare (Pariser, 2011) e (Resnick et al., 2013).

[8] Possiamo citare, a titolo di esempio, il British Museum all'estero, e in Italia il catalogo online dei Musei Civici Veneziani.

durare nel tempo (Ray, 2009). Certi ambiti sono comunque stati più recettivi di altri, come l'editoria, biblioteche e archivi, per cui ormai è possibile trovare cataloghi e pubblicazioni in formato digitale con estrema facilità. Più complesso è stato l'avvicinamento dei musei a queste pratiche, ma ormai anche questi si dedicano alla creazione di contenuti digitali oppure, nel caso di musei famosi a livello internazionale, anche delle collezioni vere e proprie disponibili sui loro siti Web. In alternativa è anche possibile osservare le nuove tecnologie all'opera direttamente all'interno del museo, creando una vivace collaborazione fra reale e virtuale che pare riscuotere anche grande successo di pubblico[9].

Ma fare *digital curation* non significa solo digitalizzare del materiale e pubblicarlo online: il valore aggiunto di una collezione organizzata che ruota intorno ad un unico tema e di cui viene possibilmente fornita una lettura guidata non si è persa. Il curatore è ancora colui che accompagna l'utente nella navigazione fornendo delle informazioni attendibili e frutto di attento studio (ad esempio i metadati, magari seguendo anche il *Dublin Core*) e mettendo anche a disposizione tutti quegli strumenti che possono essere di ausilio alla comprensione dell'intera collezione e del singolo oggetto. Non bisogna poi dimenticare altri compiti a cui già si è fatto cenno, come il compito di preservare il materiale e le tecnologie, oltre ad assicurarsi che i contenuti siano accessibili da tutti ed eventualmente addirittura creare contenuti nuovi (oppure modificare quelli già esistenti per mostrarli sotto una nuova luce) per rendere evidente il percorso che si vuole che le persone seguano (Ovadia, 2013). Gli obiettivi sono sempre gli stessi, quello che è cambiato sono le possibilità e le tecnologie a diposizione.

Accanto però ai prodotti di alta qualità creati da professionisti allo scopo di migliorare l'esperienza di apprendimento e scoperta, negli ultimi dieci anni si è assistito al fenomeno parallelo per cui la pratica di *digital curation* è "uscita fuori" dai GLAMs ed è entrata nei pc o negli smart phone di tutti quanti (Feinberg et al., 2012). Si tratta di una pratica ormai talmente tanto affermata che è stata addirittura ribattezzata *social curation,* intesa come il processo di collezione di contenuti dei Social Network allo scopo di riuso da parte propria o di altri (Duh et al., 2012) (Gehl, 2009).

[9] Il sito ufficiale del Ministero dei Beni e delle Attività Culturali e del Turismo ha pubblicato dati del 2015 per quanto riguarda l'affluenza e risultano essere in fortissima crescita rispetto al 2014.

Letteralmente *social network* significa rete sociale, e descrive appunto quel complesso sistema di relazioni che le persone intrecciano fra di loro. Sebbene, ci si augura, la maggior parte dei nostri rapporti sociali si svolgano nel mondo esterno, oramai tendiamo a riprodurli anche sul Web ed è attraverso essi che l'informazione viaggia. Le strutture che valorizzano la possibilità di creare relazioni fra utenti e che riusciranno ad attirare un cospicuo numero di *users* saranno quelle in cui l'informazione viaggerà meglio. Anche il Social Network per antonomasia, *Facebook*, con i suoi utenti in tutto il mondo, può essere usato come uno strumento di *social curation*. Solitamente non siamo abituati a guardarlo sotto questa luce, anche perché spesso se ne osserva un utilizzo molto disorganizzato da parte degli utenti, ma lo sanno invece molto bene i personaggi che rivestono cariche pubbliche o le aziende, che appunto investono su figure professionali che si occupino di curare la loro immagine sui social. La scelta consapevole di condividere con altri solo certi contenuti, siano esse fotografie, pensieri o link, ha una ricaduta diretta sull'immagine che i *followers* si creano di un certo utente: possiamo quindi dire, senza tanta paura di sbagliare, che Facebook, Instagram, Twitter e altri sono dei veri e propri strumenti di *social curation* di se stessi.

Ci sono poi altri strumenti più dichiaratamente finalizzati al *content curation*: si possono inserire in questa lista *Pinterest*, *Storify*, *Tumblr* e *Google+*[10], senza scordare software più professionali (es. *Omeka*[11]), e molti altri ce ne sarebbero in tutto il mondo visto che il loro numero è in costante aumento (Beagrie, 2008). Gli utenti sono rimasti tanto soddisfatti da questi nuovi strumenti che ne hanno subito approfittato in maniera massiccia, tanto da convincere alcuni investitori ad orientare i loro capitali in questo campo (Morin, 2011). Il grande vantaggio di appoggiarsi a questi *tools* per creare la propria collezione è che rendono molto più semplice ed intuitiva la pratica di *sharing*. Questa consiste nel rendere un nostro contenuto possibile oggetto di condivisione da parte di tutti gli utenti che vi hanno accesso sin dal momento in cui lo pubblichiamo, e possibilmente anche di diventare parte integrante delle loro stesse collezioni (Higgins, 2008).

Questo circolo, più virtuoso che vizioso, fa sì che ci siano un numero sempre maggiore di collezioni create da utenti non professionisti che vi si dedicano solo per assecondare una loro passione. L'esperienza personale mi insegna, e molti concordano con questa mia impressione, che spesso il

---

[10] In ordine https://it.pinterest.com/, https://storify.com/, https://www.tumblr.com/, https://plus.google.com/ (visitato il 30/08/2016).

[11] https://omeka.org/ (visitato il 30/08/2016).

fatto che lo stimolo venga da un interesse personale sia una caratteristica che influisce positivamente sulla qualità del prodotto finale. Ci sono poi casi in cui queste passioni si trasformano in vere e proprie professioni a tempo pieno. Basti pensare a quella categoria di 'nuovi famosi' che vengono catalogati come 'star di *YouTube'*, oppure anche casi come Jack Monroe, la quale, trovatasi in una situazione economica poco favorevole, ha iniziato a collezionare consigli e suggerimenti che trovava su Internet per garantire una dieta sana a lei e a suo figlio con un budget limitato a disposizione. Ora la sua collezione è diventata un libro e un blog sul risparmio e lo stile di vita sano seguito da moltissimi utenti[12].

Non bisogna comunque dimenticare i potenziali rischi che questo comporta: garantire l'attendibilità delle fonti e la veridicità dei contenuti diventa spesso un problema secondario per i curatori amatoriali, tanto che spesso si diffondono informazioni false ma verosimili che vengono condivise fra gli utenti senza una verifica sulla loro veridicità. Per questa ragione quando si svolge una ricerca accademica, si fa sempre riferimento ai GLAMs come fonti di informazione autorevole.

Come hanno reagito invece coloro che già si occupavano del mantenimento e della conservazione dei beni culturali all'avvento del digitale? Come in ogni ambito, le grandi innovazioni portano con loro anche delle forti resistenze da parte dei gruppi più tradizionalisti. L'indiscusso capofila di questo gruppo è Mel Buchanan che critica appunto la recente attitudine degli utenti del Web a considerarsi dei curatori solo perché possiedono una ricca collezione tematica su uno dei *tool* online sopra citati, dimenticandosi che essere curatori di una collezione è un vero e proprio mestiere[13].

Fatti saldi questi aspetti di cui fin'ora si è parlato, vorrei osservare nel dettaglio un caso particolare di GLAM che ha saputo trarre vantaggi dall'avvento del digitale.

Sebbene ci sia chi ha deciso di approfittarne solo con un certo ritardo, esistono realtà in cui la massiccia diffusione di Internet ha portato grandi vantaggi, e che invece di subire un abbassamento della qualità dei contenuti ha potuto godere dell'effetto totalmente opposto. Si

---

[12] https://cookingonabootstrap.com/page/2/ (visitato il 30/08/2016).

[13] L'articolo è stato pubblicato il 4 Ottobre 2011 col titolo *"An Open Letter to Everyone Using the Word 'Curate' Incorrectly on the Internet".* Divenne subito famoso e largamente citato il pungente incipit: 'Stop it. Just stop. Do you have a business card? Read it. Does it say "Curator" under your name? No? You are not a curator.'

possono citare ad esempio alcuni archivi digitali che invece di limitarsi alla digitalizzazione dei loro dati, si sono anche impegnati a finanziare dei progetti di ricerca con lo scopo di applicare le nuove tecnologie al loro materiale e di estrarne contenuti che altrimenti sarebbero rimasti latenti, invisibili.

Un esempio su tutti può essere quello offerto dal laboratorio Obvile dell'Università Sorbonne e il lavoro fatto sulla loro collezione di testi letterari. Partendo da una collezione di testi digitalizzati, la fase di *digital curation* ha riguardato il tentativo di enfatizzare quei contenuti e quelle caratteristiche che mettono in relazione fra loro i testi e che per uno studioso sono i più interessanti.

L'OBVIL

L' Observatoire de la vie littéraire (OBVIL)[14] è un laboratorio inserito all'interno dell'Università Sorbonne di Parigi che si occupa di osservare e analizzare l'evoluzione della letteratura e della critica letteraria degli ultimi cinque secoli. Scopo del laboratorio, così come da sito ufficiale, è servirsi di tutte le risorse messe a disposizione dall'informatica per esaminare sia la letteratura del passato che quella contemporanea nel tentativo di comprendere la maniera in cui si definiscono i canoni della letteratura.

Per svolgere questo compito il laboratorio si serve dell'archivio di testi digitalizzati della *Bibliothèque Nationale de France*[15], un archivio aperto nato con l'obiettivo di raccogliere tutto il materiale scritto in lingua francese.

Caratteristica degli archivi è di raccogliere tutto il materiale e presentarne i contenuti come se avessero tutti ugual peso, valore e rilevanza; in questo modo diventa necessariamente compito dei curatori delineare strategie per presentare i materiali in maniera appropriata al pubblico (Gehl, 2009), e questo è quello che l'Obvil ha cercato di fare. Appoggiarsi all'Archivio Nazionale di Francia ha oltretutto l'indubbio vantaggio di delegare ad altri il problema dell'autenticità e della conservazione del materiale digitale, essendo questi compiti specifici degli archivi e processi

[14] http://obvil.paris-sorbonne.fr/ (visitato il 31/08/2016).

[15] http://catalogue.bnf.fr/index.do Catalogo online (visitato il 31/08/2016).

dispendiosi in termini di tempo e denaro, concentrandosi esclusivamente sulle metodologie per l'analisi dei contenuti (Guercio, 2009). Quello che si sta per andare a vedere dunque va oltre il semplice processo di conservazione e preservazione che sì, si è visto essere importante, ma che rappresenta solo una parte del processo di *digital curation*. Digital Curation è anche e soprattutto lavoro sui contenuti, e questo è quello su cui questo laboratorio ha focalizzato le proprie risorse.

I progetti sviluppati all'interno del laboratorio sono molti, e per lo più si servono di strumenti di analisi di dati e di elaborazione del linguaggio naturale (NLP). Sebbene lo studio tradizionale della letteratura non si sia quasi mai servito di sistemi per l'analisi dei dati, magari cercando di rappresentarli sotto forma di grafi[16], oggi se ne vedono gli indubbi vantaggi e molti sono i gruppi di ricerca che si occupano di applicarli nelle maniere più creative a materiali anche molto diversi fra di loro per cercare di estrarre informazioni che fino ad ora erano rimaste latenti, o comunque poco visibili, nei testi.

Le tecniche di elaborazione del linguaggio naturale invece nascono proprio con l'intento di trattare il linguaggio, una abilità in cui si esprime tutta la creatività umana e quindi difficile da catturare in tutte le sue sfaccettature e ambiguità, attraverso strumenti dell'informatica che permettono di analizzarne le produzioni in maniera automatica[17]. Essendo così fortemente legato al linguaggio, ben si adatta allo studio dei testi letterari. Inoltre, sebbene sia stato applicato inizialmente all'inglese, oggi anche lingue come spagnolo, tedesco, francese e italiano sono ben rappresentate da corpus di dimensioni significative, quindi si realizzano algoritmi anche su diverse tipologie testuali con risultati soddisfacenti (Nadeau and Sekine, 2007).

Una delle applicazioni dell'NLP ai testi letterari proposta dall'Obvil serve appunto per svolgere l'analisi delle relazioni fra le entità nominate presenti nei testi di critica letteraria francese scritti fra il 1850 e il 1900 circa[18]. Pur trattandosi di un periodo storicamente ristretto, si tratta comunque di un'epoca molto florida e la collezione è già sufficientemente ricca da garantire dei risultati interessanti.

---

[16] Si vedano a tal proposito i progetti del LAB1100 http://lab1100.com/ (visitato il 31/08/2016).

[17] Per informazioni più dettagliate si faccia riferimento al seguente testo: C. Manning, H. Schutze. Foundations of Statistical Natural Language Processing. MIT Press, 2000.

[18] http://obvil.paris-sorbonne.fr/corpus/critique/ (visitato il 31/08/2016).

Per entità nominate si intendono tutte quelle espressioni linguistiche che operano da designatori rigidi[19] per un referente; questi referenti possono essere nomi di persone, luoghi, enti, date (Frontini et al., 2015), ma anche valori monetari e quantità (CoNLL, 2002).

Nel caso dell'Obvil sono state prese in considerazione quelle entità che rappresentano persone, luoghi e istituzioni (come ad esempio le case editrici o le università). Scelte di questo tipo vengono effettuate sulla base degli scopi che si vogliono ottenere dalla ricerca: in questo caso si vuole osservare le influenze e le connessioni fra i personaggio e i centri di influenza dell'epoca, ragion per cui la classificazione è stata ristretta a quegli elementi funzionali al raggiungimento di questo obiettivo.

Il task dell'NLP che si occupa di svolgere il compito appena descritto in maniera automatica è il *Named Entity Recognition and Classification* (NERC), parte fondamentale dei processi per l'estrazione di informazione. Si compone di due fasi, riconoscimento e classificazione, e viene realizzato allo scopo sia di fare *information filtering* che *information linking*: il primo permette di fare una previsione sul contenuto di un testo grazie all'osservazione delle entità (come definite sopra) che vengono citate all'interno di esso[20]; il secondo permette di collegare tra loro testi diversi, oppure entità stesse, per mezzo di *database* esterni, disponibili sul Web[21] e che rappresentano le informazioni in maniera strutturata, a cui tutti essi fanno riferimento. Quest'ultima tecnica in particolare è ormai molto diffusa e si stente molto spesso parlare di *linked data*; il funzionamento si basa sull'utilizzo di URI che rendono gli elementi universalmente rintracciabili (Frontini et al., 2015). Qui non si entrerà nel dettaglio delle tecniche adottate per realizzarli, ci si limiterà a dire quali grandi vantaggi offrano all'archivio dell'Obvil. In particolare si può far notare che il vantaggio è addirittura duplice, sia tecnico che informativo: fare riferimento a database esterni per quanto riguarda certe informazioni permette di non appesantire l'annotazione del testo in maniera eccessiva, e inoltre, dal momento che l'informazione in questi database è mantenuta aggiornata e possibilmente arricchita in continuazione, utilizzarli consente di servirsi di informazioni complete e la cui garanzia di autenticità e preservazione è a carico altrui.

---

[19] Si veda per 'designatori rigidi' la definizione del Stanford Encyclopedia of Philosophy (LaPorte, 2016).

[20] Tecnica utile quando si vogliono realizzare delle ricerche il più raffinate possibili su una vasta collezione di documenti: i testi che verranno restituiti saranno solo quelli realmente pertinenti alla nostra ricerca.

[21] DBpedia, GeoNames e YAGO sono solo alcuni di questi.

Sebbene il NERC sia un task largamente affrontato (ricordo che si occupa di assegnare ad ogni entità nominata la classe corretta corrispondente), permangono indubbiamente certe problematiche e difficoltà legate al task stesso. Fra queste indubbiamente il più importante è la disambiguazione del riferimento per una NE: per esempio uno stesso nome può essere associato a più elementi differenti[22], ed è quindi importante stabilire delle tecniche che il classificatore automatico possa adottare per realizzare la classificazione con il minor margine di errore possibile. Oltre alla ambiguità propria del linguaggio naturale, bisogna anche tenere conto della produttività degli scrittori: solitamente tendiamo a non ripeterci nella scrittura, per cui spesso utilizziamo perifrasi piuttosto che un nome proprio a cui il lettore riesce facilmente a fare riferimento servendosi della sua conoscenza enciclopedica, su cui tuttavia un calcolatore non può fare affidamento. Per rimanere in ambito francese, *l'autore dei Miserabili* è indiscutibilmente per tutti Victor Hugo, ma questo il calcolatore non può saperlo. Eppure è bene che riconosca la NE che questa formula rappresenta, "comprendendo" oltretutto che al suo interno essa ne racchiude un'altra, ovvero il titolo dell'opera *I Miserabili*, che deve essere opportunamente classificata[23].

L'obiettivo del progetto dell'Obvil è di riuscire a definire una strategia che permetta di riconoscere e classificare le NE attraverso un sistema automatico non supervisionato. Solitamente questi task vengono svolti o basandosi sulla similarità fra i testi e dei database di riferimento o su sistemi basati sui grafi che cercano di riprodurre il flusso di ragionamento umano per la gestione delle ambiguità (Frontini et al., 2015). L'Obvil ha scelto di utilizzare questo secondo metodo, sviluppando un algoritmo per la disambiguazione su grafo chiamato REDEN.

L'algoritmo è stato addestrato con una parte del corpus stesso che era già stata precedentemente trattata un maniera manuale: il testo puro era stato arricchito attraverso una annotazione in standard TEI.

---

[22] Goncourt, per esempio, può essere utilizzato per fare riferimento a uno qualsiasi dei due fratelli Goncourt, Edmond e Jules. Esempio tratto da (Brando et al., 2015), ma molti altri se ne potrebbero fare.

[23] Non si è ancora qui fatto riferimento alla questione dell'adattabilità del dominio per un classificatore. Sicuramente utilizzare dei classificatori il cui dominio è specifico per i testi di riferimento permette di essere più precisi, ma trattandosi in questo caso di testi molto eterogenei fra loro, un dominio eccessivamente specializzato potrebbe rivelarsi uno svantaggio. Per ogni progetto è sempre importante trovare il giusto equilibrio fra dominio specifico e generico sulla base delle proprie esigenze.

La TEI[24], acronimo per *Text Encoding Initiative*, è un consorzio di istituzioni linguistiche e letterarie che mette a disposizione uno standard per la rappresentazione dei testi in formato digitale. La sua missione è quella di sviluppare una serie di norme che fungano da linee-guida per la codifica di testi umanistici a livello internazionale. Il TEI permette di aggiungere informazione alla struttura o ai contenuti di un testo per mezzo di etichette, o *tag*. I vari *tag* sono suddivisi in moduli in base alla loro funzione, e uno di questi serve appunto per la marcatura delle entità nominate di persone, luoghi e organizzazioni[25]. Questo trattamento al testo ne rende possibile l'utilizzo per l'estrazione di informazione. L'annotazione in TEI rappresenta sicuramente un lavoro di qualità: innanzitutto poiché realizzata manualmente da annotatori specializzati, i quali non solo conoscono molto bene lo standard di annotazione ma che possono anche fare affidamento sulla loro conoscenza enciclopedica per gestire i casi più ambigui. Tuttavia si tratta di un processo molto dispendioso in termini di tempo ed energie, e proprio per questa ragione i ricercatori del laboratorio hanno cercato di utilizzare strumenti di NLP per individuare le entità nominate in maniera totalmente automatica sviluppando l'algoritmo REDEN che rappresenta le informazioni sotto forma di grafo e ne calcola le misure di vicinanza[26] per cercare di riprodurre il ragionamento umano.

L'annotazione che già precedentemente era stata realizzata è servita come *baseline* per l'addestramento degli algoritmi di apprendimento automatico non supervisionato che sono stati utilizzati, e forniscono anche quello che viene generalmente definito un *gold standard*, ovvero una annotazione data per corretta.

A seguito di questo processo quello che otteniamo sono delle entità "cliccabili" a cui l'algoritmo REDEN aggiunge in maniera automatica dei puntatori alle basi di dati esterne. Per quanto questo possa sembrare riduttivo, in realtà i pregi che porta con sé sono innumerevoli: il fatto è che una volta che si è in possesso delle informazioni custodite dai database, grazie ad esse si possono andare ad indagare aspetti che gli studiosi di letteratura non avrebbero mai potuto fare

---

[24] Sito ufficiale *http://www.tei-c.org/index.xml* (visitato il 31/08/2016).

[25] Si tratta dei tag <persName> <placeName> e <orgName>. Per le specifiche sul loro utilizzo si faccia riferimento al sito del Conorzio.

[26] Si tratta delle misure di centralità (Betweenness, Closeness e altre). Non è questa la sede per spiegarle nel dettaglio, ma in generale permettono di calcolare delle informazioni sulla struttura del grafo e permettono di estrarre delle informazioni in base a come sono legati fra di loro gli elementi.

manualmente, oltre ad avere garanzie di autorevolezza ed una maggiore visibilità dei progetti. Per esempio, la ricerca di informazioni molto specifiche (ad esempio l'influenza delle idee scientifiche su testi di letteratura in certe epoche storiche) può essere facilmente indagato a seguito di questo trattamento, e può anche fare affidamento su informazioni non contenute nei testi analizzati ma che sono riportare nelle basi di dati e che vengono visualizzati e aggregati ogni volta in maniera diversa.

Questa strategia, per quanto efficace, porta con sé delle questioni metodologiche importanti abbondantemente trattate nella letteratura relativa pubblicata, di cui qui però si preferisce non parlare. Ciò che mi interessa mettere in luce è invece il grande potenziale che questo trattamento del testo offre a tutti i ricercatori che si servono del corpus di testi di critica letteraria. Innanzitutto la rappresentazione a grafo ha l'indubbio vantaggio di essere una rappresentazione visiva, che mette subito in rilievo le informazioni più importanti. Inoltre l'utilizzo del supporto informatico permette di avere rapido accesso alle informazioni desiderate riducendo al massimo la quantità di *rumore*, ovvero quei risultati che otteniamo ma che non rappresentano perfettamente ciò che in origine cercavamo.

Conclusioni

Si è visto come il Web abbia avuto forti ricadute sul modo di fare informazione e cultura, sia da parte dei GLAMs che dei singoli utenti. Si è parlato anche di quali problematiche questa innovazione porti con sé, le quali devono necessariamente convincerci a servirsi di Internet con un atteggiamento critico e consapevole. Tuttavia i vantaggi sono di gran lunga maggiori: l'innovazione di cui si è parlato ha sicuramente aperto grandi possibilità alle Digital Humanities in tutte le loro declinazioni nel trovare e diffondere nuovi metodi per la diffusione della conoscenza a distanza.

Come si sarà potuto intuire, anche utilizzare l'informatica all'interno dell'archivio dell'Obvil ha permesso di ottenere enormi vantaggi, rendendo più ricca ma di sicuro anche piacevole la collezione di testi che presenta. Questi strumenti e i risultati ottenuti di cui si è parlato sono più utili per gli accademici che per gli utenti generici, i quali tuttavia possono comunque sicuramente godere di un lavoro di qualità. Per i ricercatori invece si tratta di strumenti preziosi che mettono in luce informazioni nuove e aprono a nuovi orizzonti di studio.

Anche guardare sotto una nuova luce dati vecchi è fare *digital curation*, non secondo la definizione più standard probabilmente, ma in un'ottica più creativa e forse ancora più utile per coloro che stanno migrando le loro attività dall'analogico al digitale.

Bibliografia

Beagrie, N. (2008) Digital curation for science, digital libraries, and individuals. *International Journal of Digital Curation*. 1 (1), 3–16.

Brando, C. et al. (2015) 'Disambiguation of Named Entities in Cultural Heritage Texts Using Linked Data Sets', in Tadeusz Morzy et al. (eds.) *New Trends in Databases and Information Systems*. Communications in Computer and Information Science. Springer International Publishing. pp. 505–514.

Buck, S. (2013) *If You Use the Web, You Are a 'Curator'* [online]. Disponibile al link: http://mashable.com/2013/05/09/curator/#8sYIYX_XPiqO (Visitato il 12 Agosto 2016).

CoNLL (2002) *Language-Independent Named Entity Recognition - Shared Task* [online]. Disponibile al link: http://www.cnts.ua.ac.be/conll2002/ner/ (Visitato il 19 Agosto 2016).

Duh, K. et al. (2012) 'Creating Stories: Social Curation of Twitter Messages.', in *ICWSM*. 2012

Feinberg, M. et al. (2012) 'Understanding personal digital collections: an interdisciplinary exploration', in *Proceedings of the Designing Interactive Systems Conference*. 2012 ACM. pp. 200–209.

Francesca Frontini, Carmen Brando, Jean-Gabriel Ganascia. (2015) Semantic Web Based Named Entity Linking for Digital Humanities and Heritage Texts. Arnaud Zucker; Isabelle Draelants; Catherine Faron Zucker; Alexandre Monnin. *First International Workshop Semantic Web for Scientific Heritage at the 12th ESWC 2015 Conference*, Jun 2015, Portorož, Slovenia. Proceedings of the First International Workshop Semantic Web for Scientific Heritage at the 12th ESWC 2015 Conference

Gehl, R. (2009) YouTube as archive Who will curate this digital Wunderkammer? *International Journal of Cultural Studies*. 12 (1), 43–60.

Grossman, L. (2006) You — Yes, You — Are TIME's Person of the Year. Time [online]. Disponibile al link: http://content.time.com/time/magazine/article/0,9171,1570810,00.html (Visitato il 5 September 2016).

Guercio, M. (2009) ARCHIVI DIGITALI in 'XXI Secolo'. Treccani [online]. Disponibile al link: http://www.treccani.it//enciclopedia/archivi-digitali_(XXI-Secolo) (Visitato il 31 Agosto 2016).

Higgins, S. (2008) The DCC Curation Lifecycle Model. The International Journal of Digital Curation 3 (1) p.134–140.

Jones, M. & Beagrie, N. (2001) *Preservation Management of Digital Materials: A Handbook*. London: British Library, Science Reference & Information Service.

Keen, A. (2007) *The Cult of the Amateur: How blogs, MySpace, YouTube, and the rest of today's user-generated media are destroying our economy, our culture, and our values: Andrew Keen: 9780385520812: Amazon.com: Books*. London: Currency.

LaPorte, J. (2016) 'Rigid Designators', in Edward N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy*. Spring 2016

Macek, J. (2013) More than a desire for text: Online participation and the social curation of content. *Convergence: The International Journal of Research into New Media Technologies*. [Online] 19 (3), 295–302.

Manyika, J. et al. (2011) *Big data: The next frontier for innovation, competition, and productivity | McKinsey & Company*

Morin, B. (2011) *The Curated Web* [online]. Disponibile al link: http://www.huffingtonpost.com/brittany-morin/the-curated-web_b_1096186.html (Visitato il 12 Agosto 2016).

Nadeau, D. & Sekine, S. (2007) A survey of named entity recognition and classification. *Lingvisticae Investigationes*. 30 (1), 3–26.

Ovadia, S. (2013) Digital Content Curation and Why It Matters to Librarians. *Publications and Research*.

Pange, B. M. & Bonde, H. (2016) Digital Library: Definitions and Its Interpretation. *International Journal of Innovative Knowledge Concepts*. 2 (3) .

Pariser, E. (2011) *The Filter Bubble: What the Internet is Hiding from You*. New York: Penguin Press.

Pitteri, D. (2013) *Internet garantisce davvero la democrazia?* [online]. Disponibile al link: http://www.ilfattoquotidiano.it/2013/03/29/internet-garantisce-davvero-la-democrazia/546005/ (Visitato il 29 Agosto 2016).

Ray, J. (2009) Sharks, digital curation, and the education of information professionals. *Museum Management and Curatorship*. 24 (4), 357–368.

Resnick, P. et al. (2013) 'Bursting Your (Filter) Bubble: Strategies for Promoting Diverse Exposure', in *Proceedings of the 2013 Conference on Computer Supported Cooperative Work Companion*. CSCW '13.

Robinson, H. (2012) Remembering things differently: museums, libraries and archives as memory institutions and the implications for convergence. *Museum Management and Curatorship*. 27 (4), 413–429.

Shahed, M. (2014) *History & Re-convergence of Galleries, Libraries, Archives, Museums (GLAM) - A systematic literature review*. Queensland University of Technology. Science and Engineering Faculty.

Smith-Yoshimura, K. (2011) *Social metadata for libraries, archives and museums Part 1: Site Reviews*.

Thibodeau, K. (2002) Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years. *The state of digital preservation: an international perspective*. 4–31.

Weinberger, D. (2014) *Too Big to Know: Rethinking Knowledge Now That the Facts Aren't the Facts, Experts Are Everywhere, and the Smartest Person in the Room Is the Room*. Reprint edition. New York: Basic Books.

Wolff, A. & Mulholland, P. (2013) 'Curation, curation, curation', in *Proceedings of the 3rd Narrative and Hypertext Workshop*. 2013 ACM.