www.ec-aiss.it

Testata registrata presso il
Tribunale di Palermo
n. 2 del 17 gennaio 2005
ISSN 1970-7452 (on-line)

Vita e morte dei Big Data. Il contributo semiotico dal data mining all'interpretazione

Rita Lisa Vella

Abstract

To focus on the entire life-cycle of Big Data implies a breaking down of aspects relating to texts, narrative mechanisms, risks linked to omission and manipulation. Behind the progressive emergence of sense there is a process of interpretation, translation, disambiguation of texts. The issue of Big Data is not only an issue of increasing complexity of available texts (of which Big Data constitutes just one aspect), but it is also the reflection of cultural change that requires the extension of the analytical approach to the context in which data are created and used. Who uses Big Data? The semiotic contribution, from data mining to the interpretation, measures its validity in the configuration of the systemic analysis: from the breaking down of the text in its basic recurrent elements to the relationship and comparison with other texts in the specific context in which they circulate and the situations in which they are used and which influence their signification.

"È il flusso continuo della folla, tessuto fitto come una stoffa senza strappi né rammendi,
composto da una moltitudine di eroi quantificati che perdono nome e volto
diventando il linguaggio mobile di calcoli e di razionalità che non appartengono a nessuno.
Fiumi di numeri lungo le strade."
Michel De Certeau (1990)

1. Big Data e Humanities. Che cosa significa analizzare i dati nel segno della cultura

Ogni volta che una persona usa uno smartphone, un bancomat, un navigatore GPS, è al contempo soggetto e oggetto di dati. In verità, usare attivamente il dispositivo non è più necessario, basta averlo in tasca mentre si cammina perché una compagnia telefonica locale possa "mappare" i tuoi percorsi[1].

[1] Il riferimento è al progetto "*Barcelona Cruise Passenger Behaviour*", Telefonicacatalunya.com, "*Big Bang Data; un mundo dadificado e hiperconectado*", 19 maggio 2014 http://telefonicacatalunya.com/big-bang-data-un-mundo-datificado-e-hiperconectado/

La storia di quello che sembra un fenomeno piuttosto recente inizia tra gli anni Trenta e Quaranta con il boom delle informazioni, collegato all'incremento della popolazione negli Stati Uniti[2], mentre il dibattito attuale investe questioni disparate che vanno dal management dei dati in real time alla visualizzazione interattiva, fino alla cultura e al comportamento umano in generale, non di rado in aperta critica all'abuso di metodologie quantitative (Lazer *et al.* 2009; Highfield, Leaver 2015; Manovich 2015).

> "Too often, Big Data enables the practice of apophenia: seeing patterns where none actually exist, simply because enormous quantities of data can offer connections that radiate in all directions" (Boyd, Crawford 2012, p. 668).

> "However, traditional 'small data' often offer information that is not contained (or containable) in big data, and the very factors that have enabled big data are enabling more traditional data collection." (Lazer *et al.* 2014, p. 3).

2. Il ciclo di vita dei Big Data. Alla scoperta della testualità

"Ciclo di vita" è un'espressione usata in biologia per riferirsi alle serie di cambiamenti che accompagnano un essere vivente dalla nascita alla morte. In ambito economico, tale modello concettuale trova particolare fortuna nell'ottica della possibilità di impostare strategicamente l'attività di impresa considerando le fasi attraversate da imprese, marchi e prodotti, senza però implicare "l'inevitabile decadenza e morte delle medesime, né una sequenza immutabile di stadi evolutivi assimilabili a quelli biosociali" ("Marketing", Enciclopedia Treccani.it, Enciclopedia delle Scienze Sociali 1996).
Questo modello è funzionale all'idea di un'analisi dei dati nel segno della cultura, poiché la scomposizione in fasi e la loro considerazione strategica consente di evidenziare il loro statuto di "prodotti" (culturali), in contrasto a una pretesa "purezza" (naturale) del dato stesso.

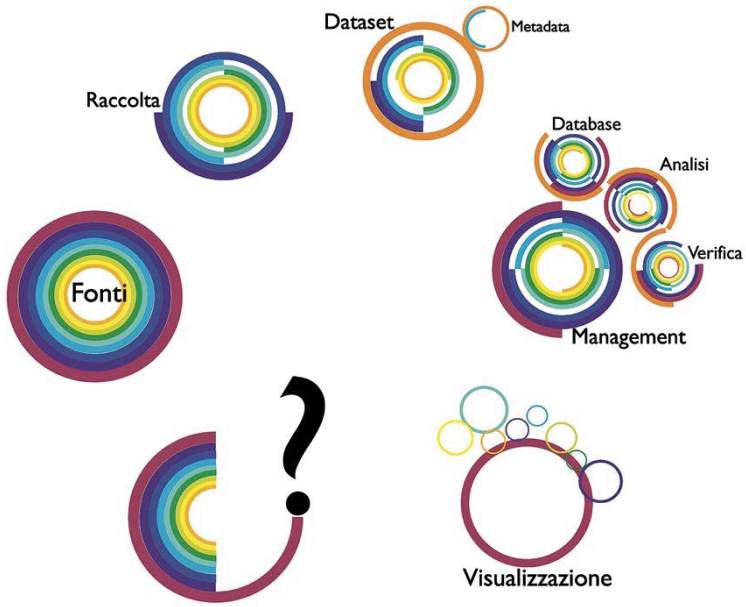

Fig. 1 – Ciclo di vita dei Big Data (Rappresentazione realizzata dall'autrice)

[2] Un riferimento utile alla cronologia dei Big Data: "*A very short history of Big Data*", Gil Press, Forbes, 9 marzo 2013 http://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/#4676afa255da

Come evidenziato dall'immagine in Fig. 1, il ciclo di vita dei big data può essere rappresentato in 5 fasi fondamentali e relative sottofasi:

- Fonti – è la fase dell'esistenza, i dati esistono in forma dispersa e disorganizzata.
- Raccolta – i dati vengono registrati e raccolti mediante dispositivi/software; l'analista seleziona le fonti e gli strumenti pertinenti.
- Dataset o Organizzazione – si manifesta un insieme di dati tra loro in relazione. In questa fase la gestione dei *metadata* è particolarmente delicata. Questi sono presenti fin dalla prima fase, tuttavia la loro organizzazione formale nel dataset è determinante per la completezza e profondità del database. In essi si annidano spesso dettagli cruciali anche se non direttamente applicabili.
- Management - eventuale integrazione del dataset in un *database*, un insieme di dati organizzati secondo un modello logico scelto dall'analista (gerarchico, reticolare, relazionale etc.). Sono possibili operazioni di interazione con i dati, non solo archivio, ma anche processi di elaborazione e gestione. In seguito, si procede con l'*analisi e la verifica* dei dati. La presenza del management è tanto più importante quanto più è resa invisibile dalla successiva fase di visualizzazione.
- Visualizzazione - è una delle fasi più dense, da un lato, ancora legata alla figura dell'analista che ne sceglie la forma; dall'altro, è l'unica fase di confronto e relazione con "il lettore".

Da dove vengono i dati? "I dati siamo noi" è una risposta diffusa che, però, è il risultato di un'arbitraria sostituzione di "dati" con "fonti". I dati possono riguardare azioni intenzionali e/o non intenzionali di diversi tipi di soggetti, che, per questo motivo, sono incessanti produttori di dati. Nel caso dei big data, la confusione si deve alla frequente coincidenza tra produttore e fonte di dati. Non è così, ad esempio, per gli hacker, che provano a non essere fonti di dati, ma agiscono, e quindi producono dati che, però, non si traducono in informazioni. Altro esempio, le indagini di polizia, dove la fonte (il soggetto che rivela le informazioni) può non coincidere del tutto con il produttore (il soggetto che ha commesso il reato). In questo testo, si intende con "fonte di dati" il soggetto o lo strumento che rende disponibili i dati, ovvero, ne facilita la reperibilità e la possibilità di trarne informazione e conoscenza. "Dato" è, quindi, ciò che è oggetto di interesse conoscitivo. Il fatto che un soggetto x sia passato in un luogo y in un'ora z, assume il valore di dato solo all'interno del contesto interpretativo "indagine di polizia", ovvero:

a) È un dato solo per i soggetti coinvolti nella ricerca e raccolta (l'investigatore, un indiziato etc.)
b) È un dato solo in relazione ad altri dati, in questo caso, luogo y e ora z

La questione delle fonti, tutt'altro che scontata, è il punto di partenza di quell'operazione di ritaglio e giustapposizione che determina l'esistenza e la validità del dato. Supponiamo che ci sia un soggetto interessato ai percorsi dei turisti in città e che, per questo motivo, abbia necessità di accedere ai dati frammentati e dispersi in diverse tracce e tecnologie. Nel suo percorso da A a B, il turista ha incrociato nell'ordine: un ripetitore, la videocamera di una banca, quella di una gioielleria, poi è entrato in una stazione metro e uscito da una diversa stazione metro dove il suo cellulare ha agganciato un nuovo ripetitore. Non esiste ancora il dato "percorso da A a B", ne esistono solo le singole tracce che lo compongono che il soggetto interessato dovrà raccogliere. La completezza e l'esattezza di questo dato, pertanto, dipendono: dalla possibilità di accedere ai diversi dispositivi e dalla quantità/qualità della raccolta di queste tracce. I dispositivi che hanno catturato e conservato le tracce, diventano fonti per il soggetto che sta operando la raccolta.
Si rivelano quindi due livelli di fonti e di raccolta: il primo, che potremmo chiamare "livello zero", in cui un soggetto è produttore e, spesso, fonte di dati per i dispositivi o i software che li catturano e li conservano, ovvero li raccolgono; il secondo in cui tali dispostivi diventano la fonte dei dati che un soggetto intenzionalmente raccoglie per uno scopo. Complessità ulteriore si manifesta quando il

soggetto intenzionato alla raccolta, in aggiunta o in sostituzione dei dati già esistenti, dispone nuovi strumenti[3].
Uno dei motivi dell'attuale enfasi pluridisciplinare sui big data è proprio la cosiddetta "moltiplicazione delle fonti", legata all'avanzare della tecnologia che ha aumentato la disponibilità di dispositivi e software di raccolta. Tali dispositivi e software di raccolta, come vedremo, non sono esenti dal contenere (e quindi trasmettere ai dati) deviazioni, come evidenziato in particolare dalle più recenti ricerche relative ai social media big data (Highfield, Leaver 2015; Zeynep 2014).
In definitiva, senza forme di raccolta non esistono "dati" ma solo "fonti di dati", o meglio, esistono dati che però diventano informazioni solo una volta raccolti per uno specifico scopo e messi in relazione ad altri dati. Joanna Drucker (2011) sostiene l'inesistenza del dato scientificamente puro proprio perché l'atto di raccolta è necessariamente orientato a specifici obiettivi con specifici strumenti. Per questo, sarebbe utile parlare non di "data", passiva accettazione di ciò che è dato, ma di "capta", attiva costruzione dell'oggetto di interesse.
Già da una considerazione delle prime due fasi, fonti e raccolta, si intende che l'unico modo di conoscere o fare esperienza del dato è nella sua forma mediata.
Si propone un lavoro realizzato con Stefano Perna e Pierluigi Vitale, presso il laboratorio di Data Factory dell'Università degli Studi di Salerno.
Case study: #grexit
Fonte di dati: twitter
Obiettivo: Individuare i maggiori influencer delle conversazioni sull'eventuale uscita della Grecia dal sistema monetario europeo.

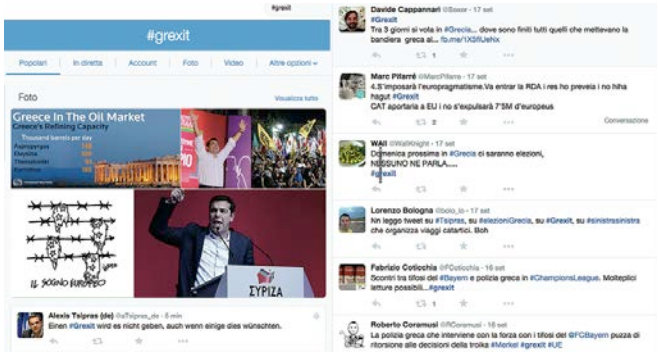

Fig. 2 – #grexit (Twitter, 19 settembre 2015)

Le immagini si riferiscono a tweet diversi rispetto al periodo esaminato dal 19 al 23 aprile, quando cioè la keyword *grexit* è diventata un *trending topic* arrivando a raccogliere circa 40.000 mentions senza contare gli hashtag collegati. La raccolta è stata fatta con un tool[4] di interrogazione delle API scelto dall'analista. I dati sono stati esportati in un dataset che spesso può arrivare a comprendere anche più di venti colonne. Nel caso in analisi, il dataset comprende tutti i tweet, gli utenti, gli eventuali destinatari, il numero di follower e il numero di following, ora del tweet, lingua con cui l'utente usa twitter, geolocation, tipologia: mention/replies/tweet, numero di retweet e di preferiti per ciascun tweet. Su queste basi, l'analista può avviare altri tipi di operazioni come estrarre hashtag e domini web.

[3] Es. Campagna cuoredinapoli in cui all'interno del tessuto urbano viene appositamente collocata una scultura antropologica-relazionale che stimola l'attività e la partecipazione degli utenti. L'utente è un produttore di dati consapevole, per quanto sia variabile il grado di consapevolezza e/o intenzionalità che possiede come fonte di dati. Allo stesso tempo, l'operazione di scelta e selezione degli strumenti di raccolta viene svolta al principio, prima della stessa produzione di dati. http://www.nuovetecnologiedellarte.it/cuoredinapoli/ http://www.mediaintegrati.it/cuoredinapoli/
[4] Attraverso questi tool è possibile fare estrazioni di circa 18.000 tweet su un periodo di 7 giorni, ma esistono anche metodi per operare raccolte più lunghe.

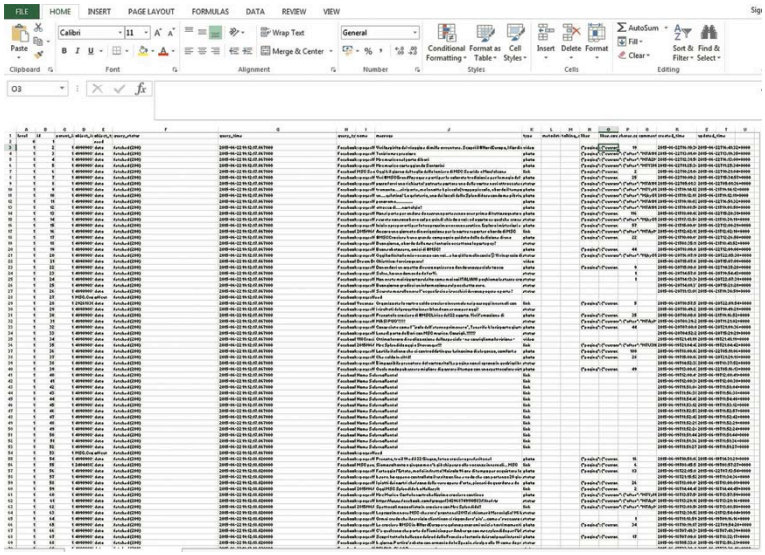

Fig. 3 – dataset

Per il management dei dati, l'analista ha esportato i dati in un formato supportato da Gephi[5] per compiere due operazioni principali:

- Elaborazione dei nodi in base ad alcuni algoritmi basati su criteri di centralità che restituiscono la posizione di un attore nella rete[6]. I nodi corrispondono ad ogni singolo utente e le misure riguardano la sua importanza nella rete.
- Colorazione dei nodi in base all'appartenenza a delle community individuate grazie a un algoritmo di clustering (modularity class) che individua statisticamente i gruppi.

La fase di verifica può aprire nuovi scenari nel ciclo di vita e interrompere la linearità obbligando l'analista a tornare al database per nuove operazioni. Per la visualizzazione, l'analista ha scelto di esportare il grafo in formato interattivo[7].

[5] "Gephi is the leading visualization and exploration software for all kinds of graphs and networks" da https://gephi.org/

[6] Betweenness centrality indica la capacità di orientare la conversazione attraverso la rete; closeness centrality riguarda la vicinanza tra i nodi (un nodo centrale avrà una minore distanza con gli altri nodi). Per un approfondimento si consiglia di consultare http://www.martingrandjean.ch/gephi-introduction/

[7] Grafo in formato interattivo con linguaggio sigmajs: http://www.socialistening.it/networkviz/grexit/network/index.html#

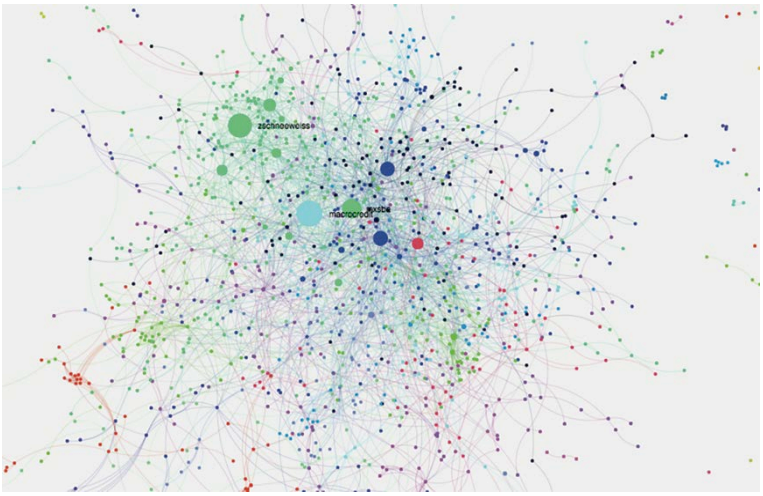

Fig. 4 – Grafo interattivo

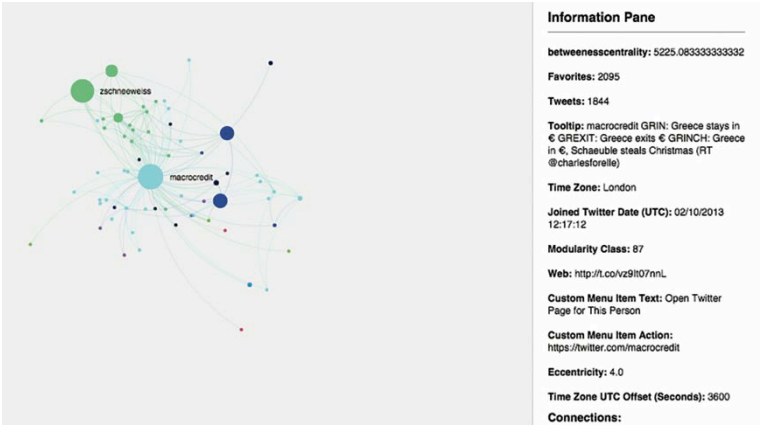

Fig. 5 – Nodo isolato

Anche la fase di visualizzazione può aprire nuovi scenari di intervento dell'analista nel ciclo di vita. Quando, ad esempio, il nodo principale si rivela un account ufficiale, l'analista potrebbe scegliere di escluderlo lasciando emergere nodi e relazioni in precedenza sottovalutati. Nel caso specifico, è emerso un cluster importante di partecipanti (quelli in color verde chiaro) afferenti alla sfera Bloomberg, una tra le più note agenzie di stampa internazionale. Esplorando i nodi principali e la breve descrizione che accompagna gli account è chiara una "competenza" nell'ambito economico e geopolitico internazionale:

- Zoe Schneeweiss, Bloomberg Economy Editor in Switzerland;
- Maxime Sbaihi, Euro-Area Economist, Bloomberg Intelligence;
- Alberto Gallo – Head of Macro Strategies at Algebris Macro Credit Found;
- Yannis Koutsomitis, European affair Analyst;
- Simon Asselbergs, la cui descrizione rimanda a una visione sarcastica del mondo finanziario, ha un profilo twitter privato, ragion per cui è stata accompagnata una ricerca ad hoc[8]
- The Greek Analyst, rimanda a un blog[9] dedicato all'economia e alla geopolitica con focus sulla Grecia e sull'Europa.

[8] http://cv.simon-asselbergs.com/?page=skills
[9] https://greekanalyst.wordpress.com/

Nell'ambito del lavoro del laboratorio Data Factory, il caso è stato inserito in un progetto più ampio di ricerca volto a testare metodologie quali-quantitative. La validità del caso, infatti, non risiede tanto nell'individuazione dei sei principali influencer della conversazione e degli specifici cluster, quanto nel procedimento sotteso di prove ed errori che ha accompagnato l'utilizzo dei software e corretto la direzione di ricerca: individuare gli influencer è un punto di partenza per nuovi procedimenti analitici. Sulla base degli stessi dati, infatti, si è deciso di procedere a una seconda visualizzazione (Fig. 6) realizzata con il software Tableau[10] al fine di cogliere la complessità e il valore del portato semantico dei dati analizzati, nonché il quadro completo di competenze. In particolare:

- l'intero flusso della conversazione, individuando i momenti di maggiore attività e intensità di partecipazione;
- l'andamento di specifici temi, che consente una sorta di "co-topic" analysis (Fig. 7);
- tutti i tweet di un utente (Fig. 8);
- tutte le lingue dei tweet.

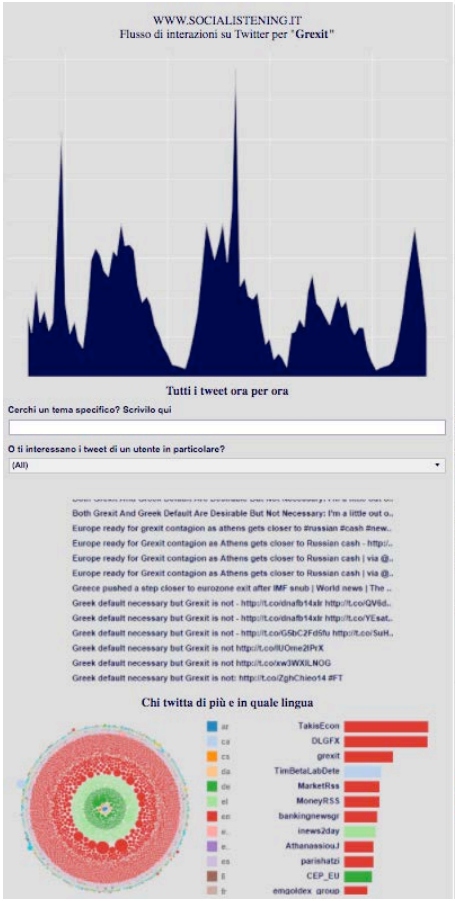

Fig. 6 – visualizzazione con Tableau

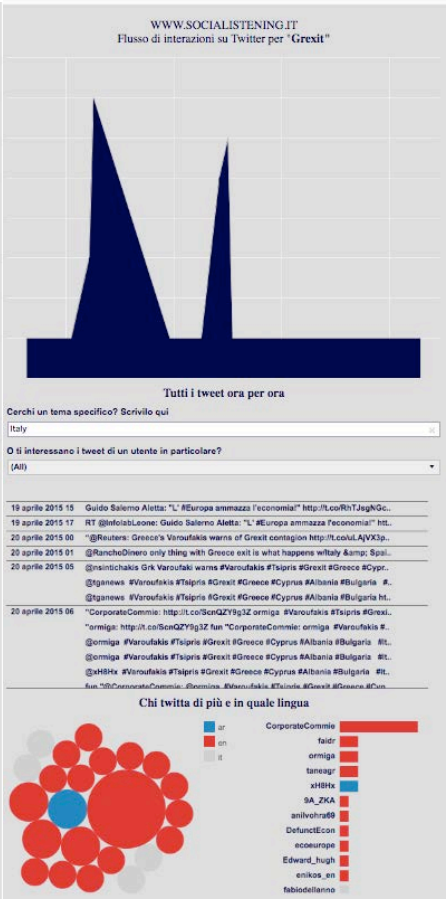

Fig. 7 – "co-topic" analysis

[10] https://www.tableau.com/

Fig. 8 – @macrocredit tweets.

Uno dei problemi principali legati all'analisi dei big data è quello di una eccessiva semplificazione incapace di rappresentare realmente la complessità socio-culturale cui si riferisce (Highfield, Leaver 2015; Zeynep 2014). A questo, è necessario rispondere sia con analisi multi-piattaforma, sia con procedimenti analitici incrociati sugli stessi dati, senza mai dimenticare l'aspetto off-line dell'attività umana.
Il procedimento effettuato non vuole essere un modello né esaustivo, né conclusivo dell'analisi dei social big data, lasciando comunque aperte nuove direzioni di ricerca (es. linguistica computazionale; text mining; topic modelling). Senza dubbio, ha l'intento di muovere i primi passi nella direzione del rispetto della complessità, promuovendo procedimenti integrati. In ultimo, le due visualizzazioni sono state condivise sul web, dove restano disponibili per la consultazione e la condivisione. Il ciclo di vita dei big data è, cioè, un processo tutt'altro che lineare e chiuso, in cui, conclusa la fase di visualizzazione, si aprono gli spazi della diffusione e della circolazione.

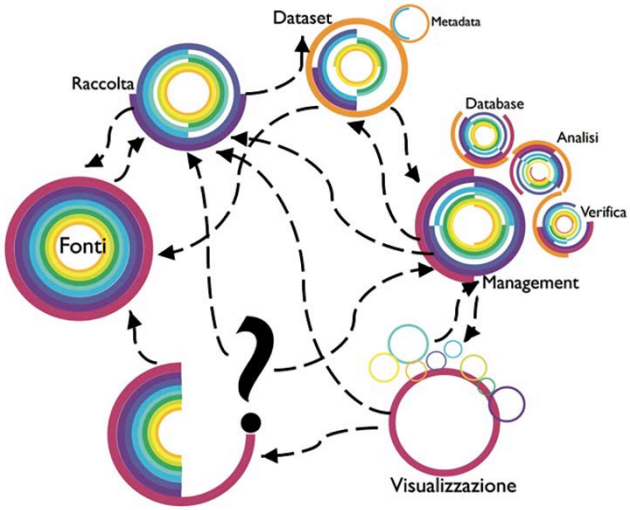

Fig. 9 – Ciclo di vita dei Big Data con relazioni tra le fasi visibili (Rappresentazione realizzata dall'autrice)

Ne segue l'evidenza di alcuni snodi teorici fondamentali:

- le caratteristiche di veridicità e valore dei dati, riconosciute come cruciali (Zikopoulos *et al.* 2013), non appartengono all'insieme dei dati in quanto tali, ma dipendono dal modo in cui questi vengono trattati, organizzati e presentati, palesando così, le istanze soggettive presenti nel discorso dei dati.
- la questione interpretativa (Boyd, Crawford 2011; Gonzalez-Bailon 2013), solitamente messa in luce in riferimento al lavoro dell'analista, nella fase che segue quella della visualizzazione, non può tralasciare la considerazione delle competenze e delle esperienze del lettore.

1.1 Alla scoperta della testualità

Il ciclo di vita dei big data può essere definito un processo di disambiguazione testuale per gradi di leggibilità e complessità:

1. COMPLESSO ILLEGGIBILE (dati prima e durante la raccolta)
2. SEMPLICE ILLEGGIBILE (database e management)
3. COMPLESSO LEGGIBILE (visualizzazione e circolazione)

Definiamo "*complessità*" la "caratteristica di un sistema […] concepito come un aggregato organico e strutturato di parti tra loro interagenti, in base alla quale il comportamento globale del sistema non è immediatamente riconducibile a quello dei singoli costituenti, dipendendo dal modo in cui essi interagiscono" (Enciclopedia Treccani.it).
La "*leggibilità*" è, invece, "[…] Il fatto d'esser leggibile, con riguardo alla scrittura, oppure alla comprensibilità" (Enciclopedia Treccani.it).

Chiaramente, il proposito di rendere leggibile la complessità, non necessariamente trova il suo compimento. Scopo del processo di visualizzazione è la ricerca della traduzione formale più adatta affinché emergano relazioni e significati, ma il processo di disambiguazione non può ritenersi completo quando vengono rappresentati dati semplici (o pattern isolati) e, solo per questo, leggibili. Il lavoro di Manovich (2010), ad esempio, si è basato a lungo sulla questione della visualizzazione per risolvere i problemi legati alla "data reduction" ammettendo che, anche laddove si riesca a preservare la forma originale dei dati attraverso una riorganizzazione senza riduzione, sono inevitabili trasformazioni dei dati, come ad esempio, la misura.
Se già la dimensione testuale emersa attraverso la ricostruzione del ciclo di vita consentiva uno spazio di riflessione per un approccio semiotico ai big data, la questione culturale (Manovich 2015) ne evidenzia ulteriormente la necessità. In questo senso, la semiotica raccoglie la sfida di pensare i Big Data[11], non tanto sulla base di una metafora "*considerare i Big Data come se fossero testi*"[12], ma, in seno alla semiotica della cultura, proprio in ragione della loro testualità, intendendo con testo "qualcosa che il mondo, la cultura, riconosce come tale" (Lorusso 2010, p. 20). La possibilità di un approccio testuale ai big data esiste, perciò, non in ragione di una simulazione, ma di un approccio che non ha mai derivato la sua specificità da quella del suo oggetto analitico[13].
Quando si parla di *big cultural data*, tutte le analisi devono rispondere del fatto che il dato sia solo un sostituto, una parziale rappresentazione dell'oggetto culturale che ci si propone di analizzare. Proprio per questo, il contesto (Lotman 2006, Lorusso 2010, Marrone 2011, 2013b) è un'informazione essenziale al fine di intendere il rapporto e le dinamiche tra il testo e il suo sfondo. Viene considerato, cioè, il modo in cui il senso del discorso dei big data si manifesta tanto nelle strutture interne al testo, quanto in quelle esterne, nelle relazioni e nei fenomeni culturali agganciati, usi, pratiche, circolazione, "la vita del testo nel mondo" [14]. L'ultima fase del ciclo di vita, quella della diffusione e circolazione, obbliga a un ripensamento della questione interpretativa in termini di competenze del lettore e esperienze del lettore. Cosa accade quando il testo confezionato dagli analisti circola attraverso i media e incontra i suoi destinatari? Cosa accade quando i destinatari non possiedono tutte le competenze necessarie a decifrare i processi e le fasi sottesi al testo visualizzato che incontrano e che, per questo, interpretano male, o in maniera parziale?

---

[11] Sfida lanciata da Paul Buissac nel 2012 dalle pagine del blog SemiotiX: Occupy Semiotics http://semioticon.com/semiotix/2012/12/occupy-semiotics/
[12] Carlos A. Scolari, Occupy Semiotics (Hacia una semiotica del big data) http://hipermediaciones.com/2012/12/16/occupy-semiotics-big-data/; anche gli ultimi lavori Los ecos de McLuhan: ecología de los medios, semiótica e interfaces. (2015); Prologo in Cultura Transmedia (Jenkins, 2015).
[13] Cfr. Lorusso A. M (2010) p.20
[14] Il riferimento è a Lotman (2006), vedi anche Lotman in Lorusso A.M. (2010).

Un esempio illuminante è la mostra presso la Public Library di New York del progetto *On Broadway* (Manovich 2015a). Durante la mostra, i visitatori hanno "personalizzato i big data" selezionando una particolare zona della città, per loro significativa, e navigandola attraverso l'interfaccia interattiva. Sono i "prodotti culturali" considerati non più "come dati a partire dai quali elaborare statistiche sulla loro circolazione o individuare i meccanismi economici della loro diffusione, ma come il repertorio in base al quale i fruitori li utilizzano secondo modalità proprie. Questi fatti pertanto non sono più i dati dei nostri calcoli, bensì il lessico delle loro pratiche" (De Certeau 1990, p. 65).

2. The dark side of Big Data. Il ruolo anti-ideologico della semiotica

Il ciclo di vita dei big data palesa le istanze enunciative del discorso, dagli analisti impegnati nelle prime fasi di reperimento e analisi, alle figure coinvolte nella circolazione (es. giornalisti) e uso (utenti finali), senza contare la necessità di considerare i diversi strumenti tecnologici impiegati nelle varie fasi del processo, da quelli atti a registrare dati, agli ambienti digitali (e non) in cui hanno luogo i comportamenti che li originano, ai software utilizzati per processarli e visualizzarli.

"Analizzare la struttura di un testo artistico significa, in questo quadro, spiegare come esso 'diventi portatore di un pensiero determinato, di un'idea' e, nello stesso tempo, 'come la struttura del testo si rapporti a quest'idea' (Lotman, 1976, p. 11)" (Marrone 2013b, p. 439).

I Big Data sono "tracce documentali" il cui interesse risiede nel "modo in cui queste tracce fanno sistema, le relazioni che le collegano e il modo in cui danno identità a chi le utilizza"[15] (Schöch 2013). Chi utilizza i Big Data? Per capirlo, bisogna chiedersi piuttosto: chi avvia l'azione? In risposta, possiamo immaginare due schemi di azione. La prima, in cui un soggetto (un'azienda, un'istituzione, una testata giornalistica) incarica un altro soggetto di realizzare un'indagine per un proprio scopo. Il soggetto mandante dell'analisi potrebbe essere anche il ricevente ultimo dei dati, che non circoleranno pubblicamente; oppure, i dati analizzati e visualizzati saranno poi condivisi, resi fruibili da una porzione estesa di pubblico. Il più delle volte, in questi casi, la circolazione riguarderà solo i dati nella loro ultima forma, visualizzata. La seconda, in cui è il soggetto stesso o un suo dipartimento interno a dare avvio all'indagine. In questo caso, cioè, non c'è un soggetto esterno mandante dell'azione, ma è la stessa azienda, università, testata giornalistica, a decidere di avviare la ricerca.
In conclusione, i big data sono utilizzati per scopi diversi e con diversi gradi di coinvolgimento da[16]:

- Aziende;
- Istituzioni;
- Università e gruppi di ricerca;
- Utenti finali;
- Testate giornalistiche.

È l'ultima fase del ciclo di vita a illuminare la necessità di un approccio anti-ideologico ai big data, volto a svelare da un lato, il modo in cui il testo finale si rende portatore di un'idea, dall'altro, il modo in cui questa idea viene accolta/rifiutata/utilizzata nelle pratiche dei fruitori secondo le loro modalità, ma soprattutto, il modo in cui i big data conferiscono identità a chi li utilizza.
La spesa in armi dei paesi europei. I percorsi dei turisti in città. La città in cui si sorride di più nei selfie.

[15] Traduzione personale
[16] Gli esempi sono molteplici: Obama è stato definito "The big data President" (The Washington Post, 2013); laboratori universitari (UniTo, http://www.despina.unito.it/; Unisa, http://www.unisa.it/news/index/idStructure/2405/id/15260; MIT Media Lab, https://www.media.mit.edu/; Cambridge, http://www.bigdata.cam.ac.uk/); giornalismo (datajournalism.it; followthemoney.lastampa.it/; http://www.opentg.it/controlla-i-tg/).

Quale isotopia comune attraversa il "*data-driven discourse*"? I big data forniscono inedite capacità conoscitive dell'ambiente circostante, che possono poi essere tradotte per obiettivi di business o di dissertazione accademica, essere usate a scopo civico o politico, eticamente o con obiettivi di manipolazione.
Gianfranco Marrone ha scritto che per la scena informativa

> "i regimi del sapere si configurano diversamente a seconda delle moltiplicazioni e delle relative unificazioni dell'osservatore e dell'informatore. È chiaro infatti, per esempio, che se le fonti di informazione sono molteplici e contraddittorie, il sapere tende a sgretolarsi, a perdere di credibilità, mentre se l'informatore è unico viene prodotta una specie di discorso realistico, veritiero." (Marrone 2001, p. 119)

Tutto torna al principio, a quell'information explosion da cui si fa incominciare la storia dei big data. Il moltiplicarsi delle fonti di informazione, in particolare con il web, ha minato i regimi di credibilità che hanno da sempre sostenuto i discorsi politici, economici, giornalistici. La mancanza di distinzione tra diffusione, popolarità e veridicità, la moltiplicazione delle possibilità espressive, la distanza sempre più infinitesimale tra utenti e fonti, tutto ciò ha eroso fortemente non solo la credibilità, ma anche l'autorevolezza delle fonti di informazione. Nessuna fonte è autorevole, o meglio, tutte le fonti sono autorevoli, fino a tweet contrario.
Il "data-driven discourse" sembra costruire un'*isotopia di credibilità* proponendosi come informatore unico e trasversale. Fonte dell'informazione non è più, pertanto, l'azienda, il politico, la testata giornalistica, ma i dati stessi. Per riabilitare la propria autorevolezza, i soggetti "delegano" la produzione di credibilità ai dati, proponendosi esclusivamente come "ambasciatori", portatori di questo discorso. Ma questa operazione è un'ulteriore mediazione, un *camouflage* (Paolo Fabbri, 2008). La pretesa di oggettività fonda le sue radici in un'operazione di compressione (ma non sintesi) delle molteplici fonti in una sola, i dati. Il discorso scientifico neutralizza le istanze enunciative nell'uso dell'impersonale, allo stesso modo, il discorso dei dati nasconde istanze enunciative e, con esse, il processo di produzione dei dati, che così si presentano come "naturali", puri, appunto, "dati". La stessa tecnologia impiegata nelle diverse fasi del processo si fa portatrice di questa ideologia, come se lo strumento fosse garanzia di oggettività. Al contrario, ogni strumento ha in sé categorie e deviazioni analitiche: in primo luogo, suggerendo "forme" di comportamenti, rendendo disponibili e diffuse determinate funzioni "aggregatrici" e non altre (hashtag, reaction, retweet, mentions, like etc.); in secondo luogo, il comportamento umano è infinitamente più complesso e l'uso che fa di queste funzioni è spesso arbitrario (es. ironia, sarcasmo, visibilità etc.) e non misurabile, o almeno, non senza l'integrazione di analisi qualitative profonde.
Il meccanismo che tende a rendere naturale ciò che è costruito strategicamente è l'ideologia[17], a cui la semiotica risponde con un lavoro di demistificazione, ripercorrendo le stratificazioni testuali, le catene connotative, le mediazioni.

[17] Il tema dell'ideologia è centrale in tutto il lavoro di Roland Barthes (1957, 1998)

Bibliografia

Nel testo, l'anno che accompagna i rinvii bibliografici è quello dell'edizione in lingua originale, mentre i rimandi ai numeri di pagina si riferiscono alla traduzione italiana, qualora sia presente nella bibliografia.

Barthes, R., 1957, *Mythologies*, Paris, Editions du Seuil; trad. it. *Miti d'oggi*, Torino, Einaudi, 2005

Barthes, R., Marrone, G., a cura, 1998, *Roland Barthes. Scritti: società, testo, comunicazione*, Torino, Einaudi.

Boyd, D., Crawford, K., 2011, "Six Provocations for Big Data", in *Social Science Research Network*, www.ssrn.com

Buissac, P., 2012, "Occupy Semiotics", SemiotiX, in *semioticon.com/semiotix*, consultato il 3 settembre 2015

De Certeau, M., 1990, *L'invention du quotidien. I Arts de faire*, Paris, Editions Gallimard; trad. it. *L'invenzione del quotidiano*, Roma, Edizioni Lavoro, 2012.

Drucker, J., 2011, "Humanities Approaches to Graphical Display", in *DHQ: Digital Humanities Quarterly*, vol. 5.1, disponibile su www.digitalhumanities.org/dhq/

Fabbri P., 2008, "Paolo Fabbri: Estrategias del camuflaje", Intervista di Tiziana Migliore, in *Revista de Occidente*, Fundación José Ortega y Gasset, Madrid, Número 330, Noviembre 2008; trad. it "Lo sguardo dell'altro. Strategie del Camouflage", 2008 in www.paolofabbri.it/interviste

Forbes, 2013, "A very short history of Big Data", 9 marzo

Gonzalez-Bailon, S., 2013, "Social Science in the Era of Big Data", in *Social Science Research Network*, www.ssrn.com

Highfield, T., Leaver, T., 2015, "A methodology for mapping Instagram hashtags", in *First Monday*, vol. 20, nn. 1-5.

Jenkis H., Green J., Ford S., 2015, *Cultura Transmedia. La creación de contenido de valor en una cultura en red*, Barcelona, Gedisa.

Lazer, D., Kennedy, R., King, G., Vespignani, A., 2014, "The Parable of Google Flu: Traps in Big Data Analysis", in *Science*, vol. 343, pp. 1203-1205

Lazer, D., *et al.*, 2009, "Computational Social Science", in *Science*, vol. 323, N. 5915, pp. 721-723

Lorusso, A.M., 2010, *Semiotica della cultura*, Bari, Editori Laterza

Lotman J.M, Sedda, F., a cura, 2006, *Tesi per una semiotica delle culture*, Roma, Meltemi.

Manovich, L., 2010, "What is Visualization?", in *Visual Studies*, vo. 26, No 1, disponibile su www.academia.edu

Manovich, L., 2015a, "Exploring urban social media: Selfiecity and On Broadway", disponibile su www.academia.edu

Manovich, L., 2015b, "The Science of Culture? Social Computing, Digital Humanities and Cultural Analytics", disponibile su www.academia.edu

Marbach, G., 1996, Lemma "Marketing", in *Enciclopedia Treccani, Enciclopedia delle scienze sociali*, disponibile su www.treccani.it/enciclopedia/marketing_(Enciclopedia-delle-scienze-sociali)/, consultato il 3 febbraio 2016

Marrone, G., 2001, *Corpi Sociali. Processi comunicativi e semiotica del testo*, Torino, Einaudi.

Marrone, G., 2011, *Introduzione alla semiotica del testo*, Roma, Bari, Editori Laterza.

Marrone, G., 2013b, "Divergenze parallele. La nozione di testo in Greimas e *Lotman"*, in P. Fabbri, D. Mangano, a cura, *La competenza semiotica: basi di teoria della significazione*, Roma, Carocci, pp. 427-443.

Schöch, C., 2013 "Big? Smart? Clean? Messy? Data in the Humanities", in *Journal of Digital Humanities*, Vol. 2, No. 3, journalofdigitalhumanities.org

Scolari, C. A., 2012, "Occupy Semiotics (Hacia una semiotica del Big Data)", in hipermediaciones. com: consultato il 3 settembre 2015

Scolari, C. A., 2015, "Los ecos de McLuhan: ecología de los medios, semiótica e interfaces." in Palabra clave, http://palabraclave.unisabana.edu.co/ consultato il 10 gennaio 2016Scolari, C., A., 2015, "Prologo" in "Cultura Transmedia" in H. Jenkins, J. Green, S. Ford, 2015, pp. 9-13

The Washington Post, 2013, "Obama, the 'big data' president", june 14

Zeynep, T., 2014, "*Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls*.", in "ICWSM '14: Proceedings of the 8th International AAAI Conference on Weblogs and Social Media", http://arxiv.org/pdf/1403.7400v2.pdf

Zikopoulos, P. C., *et al.*, 2013, Harness the Power of Big Data, The IBM Big Data Platform, McGrawHill

Sitografia

www.academia.edu
www.datajournalism.it/noi-altrove-a-stronger-loving-world/
www.digitalhumanities.org/
followthemoney.lastampa.it/
www.forbes.com/
gephi.org/
hipermediaciones.com/2012/12/16/occupy-semiotics-big-data/
journalofdigitalhumanities.org/
manovich.net/
www.martingrandjean.ch/gephi-introduction/
www.mediaintegrati.it/cuoredinapoli/
www.nuovetecnologiedellarte.it/cuoredinapoli/
www.on-broadway.nyc/
www.opentg.it/controlla-i-tg/
research.doing.com/
semioticon.com/semiotix/2012/12/occupy-semiotics/
www.socialistening.it/networkviz/grexit/network/index.html#
www.tableau.com/
telefonicacatalunya.com/big-bang-data-un-mundo-datificado-e-hiperconectado/
www.treccani.it
www.washingtonpost.com/opinions/obama-the-big-data-president/2013/06/14/1d71fe2e-d391-11e2-b05f-3ea3f0e7bb5a_story.html