

Metodi digitali

Fare ricerca sociale con il web

Richard Rogers

1. La fine del virtuale e la nascita del digitale

Lo studio dei nuovi media e della cultura digitale e lo studio del web dovrebbero cominciare a indagare come la ricerca applicata a Internet possa ampliare i propri orizzonti oltre lo studio della cultura digitale.

Ai suoi esordi internet era percepito come un mondo virtuale.

Nel lavoro sull'uso di Internet a Trinidad e Tobago, gli etnografi Daniel Miller e Don Slater misero in questione il concetto di cyberspazio come luogo a sé stante, i cui “abitanti” esperiscono la possibilità di trasformare le proprie identità indipendentemente dalla loro collocazione geografica nel mondo reale.

Un'altra svolta negli studi su Internet è rappresentata dall'importante progetto di ricerca *Virtual Society?*, che demistificò le presunte capacità trasformative del cyberspazio attraverso numerosi studi empirici sugli utilizzatori della rete. La ricerca si concluse con la formulazione di cinque “leggi della virtualità” che oggi coincidono con il *digital divide*. Anche nelle scienze umane i successivi studi sugli utenti hanno affrontato la questione della divisione tra reale e virtuale.

Qui si considera come si possano ripensare gli studi sugli utenti utilizzando i dati che i software raccolgono di routine.

Fino a oggi gli studi sull'utente si sono basati su metodi che privilegiavano l'osservazione. I dati prodotti dagli utenti vengono registrati e analizzati per offrire risultati “su misura”. L'obiettivo del metodo *cultural analytics*, che prende il nome da Google Analytics, è costruire massicci strumenti di raccolta, stoccaggio e analisi di dati per le *digital humanities*.

Gli strumenti online e i software installati sui computer, come ad esempio i browser, registrano l'attività quotidiana dell'utilizzatore attraverso quella che viene denominata “interattività di registrazione”.

Contributo da parte di Rogers alla definizione di una nuova fase della ricerca su internet, nella quale non conta più la separazione tra reale e virtuale. La base concettuale di partenza è il riconoscimento di internet non solo come oggetto di studio, ma anche come fonte. → es. Google Flu Trends che riapre la discussione sul web come medium anticipatore più vicino al reale di quanto ci si potrebbe aspettare.

Il lavoro che permette di diagnosticare le condizioni della società sulla base di pratiche registrate su internet porta a nuove nozioni teoriche. Il progetto dei metodi digitali introduce l'espressione “territorializzazione online” per descrivere la ricerca che segue il medium, coglie le sue dinamiche e produce asserzioni sul cambiamento sociale e culturale radicate nei dati del web.

1. *Seguire il medium: fare ricerca con i metodi digitali*

Perché seguire il medium? La ricerca su Internet è spesso confrontata con oggetti di studio instabili; questa instabilità è associata al carattere effimero dei siti web. La questione è come renderli permanenti in modo che possano essere studiati con attenzione. Il problema dell'instabilità, tuttavia, va oltre quello della preservazione. Lo studioso di Internet è spesso colto di sorpresa dai cambiamenti improvvisi del medium, come gli aggiornamenti dei software.

La pratica di seguire il medium si pone all'opposto dei tentativi di stabilizzarlo e potrebbe essere definita con un termine rubato al giornalismo e alla sociologia della scienza → *scooping*. Agli studiosi di Internet capita frequentemente di “farsi soffiare lo scoop”: analisti industriali, *watchdogs* e blogger coniano continuamente termini come *googolizzazione*, e spesso giungono a conclusioni che influenzano il lavoro accademico.

A livello teorico, seguire il medium è una forma particolare di ricerca *medium-specific*. La specificità del medium non riguarda solo la suddivisione disciplinare degli studi sui media secondo l'oggetto di studio, ma si riferisce anche alla differenza ontologica dei media; ogni medium ha una specificità legata al modo in cui entra in relazione con i sensi: quindi, i media non si differenziano a priori ma possono essere differenziati da chi li produce.

La Hayles propone un'analisi *medium-specific* basata su uno studio comparativo dei media, che consideri le esemplificazioni materiali delle caratteristiche di un certo medium e analizzi la loro presenza negli altri. La ricerca consiste nello studiare se i tratti caratteristici di un medium siano presenti anche negli altri media.

La specificità del medium che viene proposta riguarda il metodo, sia nel senso che gli strumenti preferiti per studiare un particolare medium, sia nel senso dei metodi del medium stesso. I metodi del web non ancora interrogati dalla ricerca sono anch'essi degni di essere studiati, sia in sé stessi che per gli effetti che producono sugli altri media.

Il primo lavoro nell'ambito dell'epistemologia web riguardava le politiche dei motori di ricerca e cercava di analizzare i criteri con cui questi selezionano le fonti. In un certo senso, la risposta risiede nel modo in cui i link vengono utilizzati: i link non sono altro che oggetti digitali, come le discussioni. La proposta di Rogers è studiare come questi oggetti vengano utilizzati all'interno del medium, e imparare dal metodo utilizzato dal medium stesso. L'obiettivo globale del principio di seguire il medium è di cambiare l'orientamento della ricerca su Internet in modo da considerare la rete come una fonte di dati, metodi e tecniche.

2. *Il link*

Ci sono almeno due approcci principali allo studio dei link:

1. La teoria letteraria dell'ipertesto → insiemi di link ipertestuali creano una moltitudine di percorsi distinti all'interno del testo. Gli studiosi si interessano sia ai nuovi strumenti dell'autorialità, sia alla storia raccontata navigando attraverso i link.
2. La teoria delle reti sociali → di cui fa parte la teoria del piccolo mondo e delle traiettorie sociali. I nodi che formano una traiettoria mostrano la distanza tra gli attori sociali. Gli studiosi si concentrano su come i legami unidirezionali / bidirezionali posizionino gli attori sociali. Un

attore è definito centrale quando c'è un'elevata probabilità che gli altri attori debbano passare attraverso di lui per entrare in contatto tra di loro.

I motori di ricerca trattano i link secondo un approccio scientometrico: i motori di ricerca si interessano alla posizione dell'attore, non necessariamente in termini di distanza dagli altri attori o di come un attore possa essere raggiunto attraverso la rete. I nodi definiscono la posizione dell'attore.

Il profilo degli attori può essere delineato non solo attraverso la quantità ma anche esaminando il tipo di link ricevuti e inviati es. → I siti aziendali tendono a non rinviare ad altri siti, a eccezione dei siti delle loro strutture collettive. Queste tendenze a inviare link all'interno di categorie di organizzazioni rivelano una "politica di associazione". La definizione di un attore in base ai link in entrata e in uscita permette di notare se ci sia qualche divergenza di norma.

L'analisi dei link può servire per un lavoro di campionamento più sofisticato.

Il gruppo di ricerca di Rogers ha contribuito al lavoro dell'*Open Net Initiative* utilizzando un metodo che percorre tutti i siti web di una particolare categoria, registra i link e determina ulteriori siti-chiave, che non compaiono nella lista iniziale → «campionamento dinamico degli URL», per sottolineare la differenza tra la compilazione manuale di una lista di URL e le tecniche più automatizzate per individuare URL significativi. I nuovi siti trovati con questo metodo vengono verificati attraverso le statistiche di connessione per determinare se essi siano bloccati nei vari paesi.

3. I siti web

La maggior parte dei metodi usati per lo studio dei siti web si collocano per così dire «sopra la spalla»: il ricercatore osserva i soggetti mentre navigano o usano un motore di ricerca e poi pone loro una serie di domande. Una delle tecniche è l'*eye-tracking*.

L'altro filone è l'analisi delle funzionalità per cui i siti vengono paragonati e contrapposti in base al livello di interattività. Bisogna vedere se un particolare insieme di funzioni dia come risultato maggiore utenti e attenzione, e, in questo studio, spesso i siti vengono archiviati per essere analizzati approfonditamente.

Uno dei compiti cruciali è la riflessione sui mezzi attraverso cui i siti vengono registrati e immagazzinati in modo da rendere accessibili i dati sui cui sono basati i risultati della ricerca. Il programma di ricerca *Digital Methods* si interessa specificamente ai siti web in quanto oggetti archiviati, resi accessibili nel modo più rapido attraverso la Wayback Machine dell'Internet Archive.

Si può sviluppare un metodo di ricerca basandosi sul modo in cui l'Internet Archive organizza i siti web. La Wayback Machine permette infatti di studiare l'evoluzione di una singola pagina nel corso del tempo, per esempio raccogliendo una serie di immagini della pagina (*snapshot*).

Viene da chiedersi se le storie dei motori di ricerca registrate attraverso l'evoluzione delle loro interfacce indichino dei cambiamenti più in generale sul modo in cui vengono organizzate l'informazione e la conoscenza. Per rispondere sarebbe utile applicare un approccio comparativo allo studio dei media.

4. I motori di ricerca e le sfere

Se si inserisce il nome di una persona su Google, spesso sono le forme di presenza che quella stessa persona ha creato sul web a comparire per prime, mentre quello che altri hanno scritto su quella persona appare in fondo alla ricerca. Tuttavia, con l'archiviazione delle *queries* effettuate su un motore di ricerca, un terzo insieme di tracce può contribuire a definire un individuo o un insieme di dati che si riferiscono a un individuo: la cronologia delle ricerche.

Un filone di studi sui motori di ricerca, riassunto dal neologismo *googlizzazione*, è la critica di matrice politico-economica, che considera come il modello Google possa diffondersi alle varie industrie culturali.

Solo una piccola percentuale di utenti personalizza in numero di risultati per pagina; per questi motivi il potere di un motore di ricerca risiede in una combinazione delle sue pratiche di classificazione e dell'apparente rispetto da parte dell'utente dell'ordine dei risultati forniti. Il modello di Google si basa sull'interattività di registrazione, in cui le preferenze e le cronologie sono registrate, immagazzinate e utilizzate sempre di più per fornire risultati «su misura».

Alcuni anni fa Google rese disponibile un'API che rendeva possibile la raccolta di dati. Si poteva lanciare un numero limitato di *queries* al giorno e i risultati potevano essere utilizzati per la ricerca accademica. Google disconobbe il servizio, ma lo reintrodusse sottolineando che le *queries* automatizzate e la conservazione dei dati sono contrari alle regole del servizio.

I primi risultati di una *query* sono quelli più cliccati dagli utenti; per questo motivo le organizzazioni fanno largo uso delle tecniche di ottimizzazione per i motori di ricerca, in modo da incrementare la visibilità dei propri siti. Per fare ciò ci sono tecniche → *white hat* e *black hat* che li obbligano a eliminare alcuni siti finché questi non si adeguano nuovamente alle loro regole.

Per quanto Google possa essere dominante, esistono anche altri motori di ricerca che sono leader in particolari sezioni o sfere del web, cosa che probabilmente viene spesso sottostimata.

Quando si pensa al web in termini di sfere si fa riferimento alla blogosfera; la radice *-sfera* richiama idealmente alla sfera pubblica. Una sfera può essere considerata come una serie di fonti delimitate da un dispositivo. Nell'analizzare una sfera bisogna considerare quali fonti siano più influenti, non solo complessivamente ma anche per ciascuna *query*. L'analisi comparata delle sfere mette a confronto le fonti che ciascuna sfera rimanda per la stessa *query*. Questo metodo permette di analizzare le conseguenze dell'uso di link, tag, aggiornamenti ecc. da parte di ogni motore di ricerca.

5. Web nazionali

I protocolli e i principi su cui si basa internet, in particolare la commutazione di pacchetto e il principio *end-to-end*, rafforzarono inizialmente la nozione di cyberspazio come luogo libero da costrizioni fisiche; la sua stessa architettura è costituita da uno spazio slegato dagli Stati-nazione.

In realtà la geografia è stata iscritta nel cyberspazio fin dall'inizio, basti considerare la localizzazione dei *root servers* e la distribuzione di indirizzi IP in serie, che in seguito ha reso possibile l'applicazione della tecnologia di localizzazione basata sugli indirizzi geo-IP. La tecnologia geo-IP può essere utilizzata anche dalla ricerca su internet per indagare cosa esso possa rivelare rispetto alle condizioni sociali dei diversi paesi.

Per studiare il web su scala nazionale si possono anche usare i dati che vengono raccolti di routine es. → Alexa, che attraverso il servizio *Top Sites* assume dati sui siti più visitati nei diversi paesi. Questa barra degli strumenti forniva statistiche sui siti che l'utente aveva caricato, in cui aveva navigato o a cui si era registrato, e gli URL registrati venivano confrontati con quelli già presenti nel database di Alexa. Quelli che non risultavano già inseriti nel database venivano analizzati e indicizzati: così è nato l'Internet Archive.

L'Internet Archive è stato concepito per il *surfing*, un tipo di uso di internet che probabilmente ha dato origine alle *queries* sui motori di ricerca.

6. Studi postdemografici sui social media

I social network offrono molte opportunità alla ricerca sociale e culturale.

Uno studio condotto su più vasta scala mette a confronto le reti sociali online con le reti sociali preesistenti nel mondo reale.

Mentre la ricerca sull'opinione pubblica è stata a lungo associata ai sondaggi, le informazioni contenute sui profili social potrebbero fornire una prospettiva diversa sulla composizione e le caratteristiche del pubblico. Per questo studio vengono utilizzati gli *usernames*. Analizzando quale combinazione di servizi è associata a un nome utente si entra nel campo della ricerca sugli argomenti correlati, spesso utilizzata sul web per consigliare prodotti, servizi, informazioni o contatti sui social network.

7. Wikipedia e il contenuto creato dalla rete

Gli autori di Wikipedia non sono professionisti, e tuttavia il risultato è sorprendentemente simile a quello di un'enciclopedia tradizionale. Una ricerca ha mostrato che sulle piattaforme del web 2.0, tra le quali Wikipedia, il rapporto redattore per utente è molto basso, sfatando così il mito del "contenuto generato dall'utente".

Attraverso uno studio dei dati demografici dei redattori di Wikipedia è possibile relativizzare la presunta differenza tra dilettanti ed esperti. Nonostante siano tutti volontari non retribuiti, i redattori sono estremamente coinvolti e attenti.

Il progetto Wikiscanner studia come vengano scritti gli articoli di Wikipedia. Dà la caccia ai redattori anonimi, cercando il loro indirizzo IP e confrontandolo con un database di localizzazioni di indirizzi IP.

Ricerca sulla qualità di Wikipedia introducendo i concetti di:

1. «contenuto realizzato da una rete» → contenuti mantenuti da autori umani e da strumenti non umani, come i bot e i software di allerta, che ripristinano la versione originale di una voce modificata senza permesso o notificano alla community wikipediana i cambiamenti che vengono apportati.
2. «tecnicità del contenuto» → sui bot che sono i visitatori più frequenti di Wikipedia. L'importanza dei bot e dei software di allerta come agenti di vigilanza supera quella dei redattori umani.

È possibile studiare l'evoluzione di un articolo, i materiali disponibili sono soprattutto la cronologia delle revisioni o la pagina di discussione.

8. La fine del virtuale: territorializzare le asserzioni online

L'obiettivo è dare il via a una trasformazione dei modi e delle ragioni per cui fare ricerca con internet.

Rogers ritiene sia possibile imparare dai metodi utilizzati dal medium, spostando la discussione riguardante la teoria della specificità del medium dal piano ontologico, ovvero che riguarda le sue proprietà e caratteristiche, a quello epistemologico, ovvero del metodo. Internet e il web possiedono i propri oggetti ontologici, come ad esempio i link e le tag. L'epistemologia del web studia tra l'altro come questi oggetti nativi digitali vengano utilizzati dai dispositivi.

I metodi correnti di Internet possono essere distinti tra quelli che seguono il medium e quelli che ri-mediano o digitalizzano metodi preesistenti.

La ricerca che ha per oggetto internet non deve essere limitata alla cultura online e ai suoi utenti; internet può essere ripensato come fonte di dati che forniscono informazioni sulla società e la cultura.

2. Il link e la politica dello spazio web

1. La morte del cyberspazio

Oggi i software web conoscono la localizzazione geografica di ogni utente e agiscono di conseguenza. Per rendersene conto, basta provare a utilizzare il browser google.com in Italia: si verrà reindirizzati automaticamente su google.it.

Con l'avvento nel web degli strumenti consapevoli della localizzazione geografica, il cyberspazio più che un'esperienza di spostamento diventa un'esperienza di rientro, dal momento che l'utente viene «riportato a casa» di default.

Prima che il web fosse legato al territorio secondo una geografia della localizzazione, internet offriva forme o organizzazioni spaziali che non erano basate sulle coordinate geografiche di un luogo. La cartografia del web è definita per l'utente dalla navigazione più che dal territorio o dal cyberterritorio.

Collegarsi a siti che esprimono le stesse opinioni, aderire a un insieme di siti connessi tra loro tramite link o installare un *crawler* e un dispositivo di visualizzazione grafica per evidenziare la dimensione di un movimento interconnesso o di una rete collegata a una tematica: tutte queste azioni creano e mappano la politica sul web.

Le cartografie possono essere realizzate per visualizzare diverse questioni politiche: es. → le mappe del traffico in Internet mostrano le politiche economiche di ingegneria della rete. Un altro esempio di geografia politica online è la mappa dei tredici *root servers*, che evidenzia il divario tra nord e sud del mondo e il controllo esercitato sul web dagli Stati Uniti e dai loro alleati.

La rappresentazione geografia che permette di visualizzare le questioni politiche online è meno significativa delle politiche che emergono dalle connessioni, di qualunque natura sia questo legame.

2. Notti stellate: collegare singoli siti web uno all'altro (attraverso i link) nel periodo dell'iperspazio

La visione del web come universo inizialmente coincise con l'idea del web come un iperspazio.

Oltre ad analizzare il traffico su Internet e costruire le mappe di localizzazione dei server, lo studio dei link serva a radicare lo spazio del web.

Negli anni '90 si ebbe l'intuizione che i siti web costruivano i link in modo selettivo piuttosto che casuale. C'è un grado di scelta nel costruire i link: dal punto di vista sociologico o politico. Uno studio sull'uso del web rivelò che i link hanno a che fare con le politiche organizzative, specialmente nel caso delle grandi aziende e delle istituzioni governative.

A mano a mano che i collegamenti casuali hanno ceduto il passo a quelli mirati, la cartografia dei link tra singoli siti web ha permesso di far emergere il loro significato. In questa cartografia pre-rete, ogni sito era analizzato singolarmente per valutarne la reputazione e le relazioni con gli altri siti. es. → Comparando i siti di tre grandi aziende, lo status di ciascuno di essi si differenziava in funzione dei link in entrata e delle associazioni dei link in uscita: una delle tre aziende emergeva in modo particolare perché il suo sito riceveva link anche da siti istituzionali e da organizzazioni non governative, mentre le altre due li ricevevano solo da altre aziende.

Seguendo il principio che non tutti i link si equivalgono, si è studiata la socialità e temporalità della loro costruzione. I link sono stati classificati come:

1. Cordiali → sono collegamenti che rinviano a progetti partner o affiliati e ad altre fonti di informazione “amiche”
2. Critici → stanno scomparendo
3. Di aspirazione → fatti da piccole organizzazioni verso i siti di entità sconosciute.

Questi insieme di siti collegati tra loro attraverso i link divennero noti come “spazi”; questo segnò una cesura netta con il concetto di “cyberspazio”, suggerendo che le pratiche di *linking* ne smantellassero l’apertura illimitata.

3. *Dalla politica dei percorsi di navigazione all'autorità della lista*

I grafici non direzionali fanno pensare al web come a un «piccolo mondo», in cui le distanze tra i siti sono misurabili e possono essere descritte in termini di gradi di separazione. Le mappe di link erano quindi concepite come mappe di percorsi di navigazione.

Considerare il web come un insieme di percorsi non è una forzatura. I link bidirezionali sono effettivamente meno frequenti di quelli unidirezionali, ma a prescindere dalla loro direzione, i ricercatori hanno identificato delle pratiche politiche nei loro percorsi.

Gli algoritmi del motore di ricerca più diffuso trattano i link in uscita come segni di adesione più che come tappe di un percorso. L'aspetto più sorprendente è che i link in uscita sono considerati come misure di autorevolezza del sito a cui rinviano; quindi la maggior parte degli attori che davano ordine al web condividevano la visione dei grandi pionieri dell'ipertesto e dei generatori casuali di siti, che consideravano il web come uno spazio in cui l'utente poteva tracciare un percorso, diventando autore e protagonista di un viaggio.

Per quanto riguarda i motori di ricerca, le liste vengono generate sulla base dei link che collegano i siti tra loro, e la posizione di un sito all'interno di una lista dipende da quanti link riceve da altri siti che l'algoritmo considera autorevoli.

Abbandonati lo *storytelling* e l'idea di percorso attraverso il web derivante dalle teorie letterarie sull'ipertesto, si entrava nell'ambito dell'informazione.

Alla fine degli anni '90 i link definiti in “entrata” o “uscita” non erano visibili chiaramente all'interno dei siti. I motori di ricerca che incoraggiavano le *queries* di tipo booleano rendevano possibile una ricerca sofisticata dei link in entrata. I file di log di un sito, che una volta erano considerati come uno strumento privilegiato per gli studi su Internet, sono oggi di solito invisibili al pubblico. Un'analisi approfondita del numero di visite ricevute da un sito può essere utile per studiare un singolo sito o per paragonare tra loro un numero limitato di siti.

A parte gli studiosi delle reti e gli sviluppatori di algoritmi, solamente pochi ricercatori che studiavano le politiche del web riuscivano ad analizzare questo tipo di link, utilizzando dei *crawlers* sviluppati appositamente, che li rilevavano percorrendo una serie di siti: i link in uscita di un sito venivano ricostruiti a partire dai link in entrata di altri siti.

L'abbandono dell'idea del web come cyberspazio e dei mappamondi politico-geografici dei web, cartografati e non è stato un cambiamento importante nella definizione dello spazio web. In linea con le idee, che circolavano in quel periodo, sul suo potenziale pluralismo, il web divenne uno spazio paritario, dove venivano sfidate le gerarchie di credibilità istituzionali e gli attori non istituzionali si trovavano spesso tra i primi risultati forniti dai motori di ricerca.

Google invece conferì questa autorità ad altri siti, attraverso i link e il testo del puntatore del link. Il conteggio dei link e il fatto di conferire autorità ad altri siti attraverso di essi sono oggi i principi base della maggior parte degli algoritmi dei motori di ricerca.

4. Lo spazio dei motori di ricerca e la nuova politica della sfera

L'idea della sfera era intesa sul web come "grande conversazione". Mappare le discussioni corrispondeva alle assunzioni del potenziale neopluralista, al ricco contenuto del dibattito pubblico online e allo spirito deliberativo e democratico.

Le concezioni dello spazio web e del modo in cui è organizzato devono oggi tener conto del fatto che i motori di ricerca sono sfere delimitanti e che i proprietari dei siti devono cooperare con essi per essere inclusi in una sfera.

Oggi gli algoritmi dei motori di ricerca e il comportamento dei proprietari dei siti cooperano sempre più alla costruzione delle sfere. Quando i proprietari dei siti linkano in maniera impropria, i motori di ricerca non funzionano più.

La pratica del *Google bombing* e altre forme di mancanza di cooperazione mostrano come Google e altri algoritmi analoghi al PageRank vorrebbero che i proprietari dei siti si comportassero riguardo ai link. Le opinioni di un motore di ricerca su quale dovrebbe essere il comportamento corretto dei proprietari di siti o degli utenti hanno conseguenze importanti sulla politica dello spazio web.

Le conseguenze dei proprietari/amministratori e degli utenti dei siti hanno piuttosto a che fare con la moltiplicazione degli spazi web. Quando un motore di ricerca non è in grado di gestire il comportamento del proprietario/amministratore e/o dell'utente in un nuovo spazio il web si trasforma in una serie di sottospazi.

Lo studio della politica dello spazio web ormai è trasversale, e analizza es. come lo stesso sito si posizioni tra i risultati di *queries* effettuate nelle diverse sfere. Questo aspetto solleva domande sugli effetti dei nuovi media che sono congiunti al potenziale neopluralista del web, desunto dalla teoria della sfera pubblica.

L'idea di un controprogetto di Google non è alimentata tanto dall'urgenza di uno spirito di resistenza alle sue derive commerciali nell'interesse pubblico, quanto dalla netta predominanza dei siti statunitensi.

5. Cartografia delle reti e analisi pubbliche dei siti

Mappando le reti, gli studiosi puntano la loro attenzione sul reale. L'informazione proveniente da internet non viene più considerata proveniente da una dimensione virtuale o degna di un particolare status. Qui il lavoro si fonda su analisi di molteplici siti.

Ci sono due tipi di cartografia delle reti politiche che fanno uso di questa metodologia: la cartografia sociale e quella tematico-professionale. Uno degli scopi della cartografia delle reti è rendere visibile ciò che è nascosto.

La concezione del web come spazio che può rivelare una rete sociale, unita al ritorno dell'informalità del web ha dato vita a uno sguardo indagatore.

Il lavoro svolto dal gruppo di Rogers si distingue facendo emergere un insieme di connessioni: anche quando hanno provato a rendere il web uno spazio di dibattito, hanno trovato solo una giustapposizione di affermazioni, ma scarsi scambi tra organizzazioni. Il web non sostituisce un luogo fisico o un evento in cui gli attori del dibattito possono incontrarsi di persona.

Al calo dell'interesse verso gli approcci deliberativi dello spazio web politico è corrisposto un apprezzamento per le forme di politica di rete. Mappare il web è diventato quindi un mezzo per fissare la mobilità degli attori all'interno delle reti, oltre che per interrogarsi sull'estensione della mobilitazione e dell'attenzione pubblica.

Una precedente ricerca sui movimenti sociali aveva sviluppato l'idea di un potenziale movimento di «galleggiamento libero», ovvero di una serie di pubblici suscettibili, in particolari condizioni, di aggregarsi per formare un movimento. I movimenti sono fenomeni infrastrutturali.

La nozione di attori fondata sulle reti piuttosto che sulle istituzioni o altri contesti radicati nel mondo fisico, solleva il dubbio che essi si ricordino di cosa stia accadendo sul campo. Le difficoltà maggiori nella cartografia delle reti politiche dello spazio web riguardano il modo in cui le mappe delle reti costruite intorno a uno specifico tema possano rappresentare non tanto quello che accade offline, quanto offnetwork.

6. Questioni per lo studio della politica nello spazio web recente

Queste mappe ricostruiscono le reti in modi che spesso si differenziano dai modelli infrastrutturali di rete che le hanno precedute e, in particolare, quello teorizzato da Baran che distingue reti centralizzate, decentralizzate, distribuite, a catena, a stella e multicanale. Ogni distribuzione spaziale riconfigura la rete come spazio per compiere un lavoro che non sia semplicemente mantenere efficace un flusso comunicativo o reagire a un comando eseguendo delle operazioni. Questi spazi sono basati non solo sulla fisica della mappa di rete, ma anche sulla metafisica della sfera non geometrica.

Con i primi mappamondi del cyberspazio la rete ha assunto contorni più definiti. Il web non era più descritto in termini di matrici e corridoi, ma di territori e isole. Quando comparvero i software per cartografare lo spazio web, si realizzò l'autospazializzazione.

L'introduzione dei grafi ha a sua volta inferito con le rappresentazioni basate su mappe circolari e sul concetto delle tavole rotonde virtuali.

La recente "svolta localizzatrice" della rete ha segnato la fine del cyberspazio e del virtuale come spazi politici. Da quando il cyberspazio è stato agganciato allo spazio geografico, i tentativi di conservare la sua sovranità sono stati spinti lontano.

L'attuale periodo della localizzazione è caratterizzato dalla nascita di metodi all'interno degli strumenti del web usati per rivelazioni e scandali.

es. → tendenza delle community online a essere concentrate geograficamente su un'unica piattaforma: es. Myspace in USA. Gli studi hanno sottolineato il rafforzamento delle strutture di classe nelle popolazioni di utenti nelle piattaforme social dei vari paesi. Inoltre, i ricercatori hanno trovato una miniera di dati nei profili e hanno fatto incetta soprattutto in quelli degli amici per svolgere analisi sui social network.

Il web è sempre più radicato nelle specificità geografiche e linguistiche attraverso le piattaforme e la gestione dello spazio.

3. Il sito web come oggetto d'archivio

Non si riflette quasi mai sul fatto che alla sua nascita il web fosse un'infrastruttura in attesa di contenuto e non viceversa. L'incoraggiamento a creare e mettere online nuovi contenuti arrivò sotto forma di premi "fatti in casa". Gradualmente cominciarono a essere conferiti premi per categorie. Man mano che i premi proliferavano, ad alcuni di essi venne dato un *imprimatur*, per renderli più prestigiosi o una rivendicazione di autenticità.

Un'altra pratica diffusa era la formazione di lista professionali di link e delle directory; le liste di link selezionati e organizzati per categoria potrebbero essere considerate le prime guide del web. Selezionare, valutare e classificare i siti potrebbe essere considerata una primitiva forma di analisi.

1. Il sito web archiviato e la predominanza del contenuto

In alcune aree degli studi sul web, il singolo sito viene privilegiato rispetto ad altri oggetti o spazi del web, perché è lì che si trova il «contenuto». Il sito è l'unità fondamentale di analisi del web. Di solito per salvare il contenuto, gli archivisti devono distruggere la maggior parte del sito.

Neppure vengono registrate le entità che circondano il sito, come i *cookies* o i siti interconnessi al sito che si vuole archiviare. Queste relazioni complesse si instaurano quando il sito si carica: durante questa operazione, un URL breve può rinviare a un URL di destinazione, che mette in funzione uno o più *adserver* e i *web bugs* 1x1 i quali a loro volta piazzano o leggono i *cookies* e contano i risultati: la lista degli URL che vengono caricati quando si carica un sito è visibile nel log di attività del browser.

Generalmente oggi il sito web archiviato termina con il contenuto inserito dall'autore. In un certo senso tutti gli elementi tipici dei nuovi media non sono resi disponibili per la posterità.

2. Navigare nel web del passato

Internet non esiste semplicemente in una forma che si presta all'archiviazione, ma è costruito come un oggetto da studiare nell'archiviazione.

Quando gli utenti abituati ai normali motori di ricerca utilizzano l'Internet Archive, quello che risalta ai loro occhi è il modo in cui si possono lanciare delle ricerche attraverso la Wayback Machine; inserendo l'URL la macchina fornisce una lista di pagine archiviate associate in passato a quell'URL.

In precedenza, i siti web archiviati erano visualizzabili attraverso la barra degli strumenti di Alexa, che indicava se una versione archiviata di un sito era disponibile.

Si può affermare che tutti i mezzi di navigazione dell'Internet Archive nella Wayback Machine derivino da un principio di fluidità, che consente di navigare da una pagina all'altra senza interruzioni e preserva anche Internet in quanto «cyberspazio».

Conservare la navigazione nell'Internet Archive è quindi una maniera di fare la storia del web; e lo è anche nel senso che a volte la navigazione nella Wayback Machine è più fluida di quanto non fosse nel web.

Una volta caricate le pagine disponibili nell'archivio per l'URL richiesto, si può cliccare sulle pagine fornite oltre che su ulteriori pagine di altri siti. Quando un utente clicca su un link, viene caricata la pagina più vicina alla data della pagina di origine; se non è disponibile alcuna pagina archiviata, in Wayback Machine permetterà di accedere alla pagina web «viva».

3. *La Wayback machine e la biografia del sito web come metodo storiografico*

Con la forma attuale della Wayback Machine si può studiare l'evoluzione di una singola pagina; si può anche risalire a una vecchia versione a scopi probatori o di verifica. Questo sembra essere l'uso principale che si fa della Wayback Machine nell'ambito della ricerca accademica.

La biografia di un sito può però essere utilizzata anche al di fuori del campo probatorio. Si possono ad esempio esaminare i registri pubblici per questioni di proprietà.

Riveste interesse la collezione di siti “parcheggiati” in attesa di proprietari, costituiti da un *template* e da contenuti generici raccolti e archiviati. Ottenere i file di log di un sito può essere interessante per i ricercatori che vogliono informazioni sull'andamento delle visite.

Si può anche adottare un metodo “a strati” che ha eliminato i contenuti generati dagli utenti da alcuni siti web 2.0, in modo da lasciare solo il *template* sottostante. In questo caso i siti sono stati sbucciati come cipolle, rivelando aspetti in comune nella forma e nella struttura, e sottolineando criticamente la similarità soggiacente.

Un metodo simile è l'analisi delle funzionalità, in cui si crea un codice crittografico di tutte le funzionalità di un sito e in seguito si controlla una serie di siti per verificarne la presenza o l'assenza, creando una matrice di funzionalità.

Una volta introdotta la variabile temporale in questi e altri tipi di analisi, la Wayback Machine diventa indispensabile per ricostruire la biografia di un sito.

Tutte le singole pagine disponibili tratte da <http://www.google.com> sono state registrate dalla Wayback Machine e inserite in un film e in un'infografica. L'analisi si concentra sull'area dell'interfaccia posta sopra la maschera di ricerca esaminando quali servizi di ricerca siano stati privilegiati da Google nel corso del tempo sui tab della home page.

A parte le genealogie whois, le anatomie, le analisi delle funzionalità, l'epistemologia dell'interfaccia e le sue implicazioni politiche, si possono registrare e interpretare i cambiamenti nella sostanza di un sito. A essere analizzata è la sostanza del menu principale che è situato sulla home page e organizza il contenuto del sito. L'analisi dei contenuti del menu è l'unico metodo tra quelli descritti finora che si avvicina alla pratica di ricerca basata sulla lettura dei siti web.

A livello metodologico seguì l'analisi lessicale: data una serie di sinonimi, verificarono se il termine utilizzato fosse più o meno duro o estremista. I giornalisti scoprirono che i siti delle organizzazioni di destra più moderate tendevano ad allinearsi gradualmente ai toni e ai contenuti di quelli dell'estrema destra.

Con strategia analitica, si può stilare una lista di siti già archiviati e fornire i mezzi per accedervi, lanciare *queries* e analizzarli in altro modo: un metodo di studio del sito come oggetto archiviato.

4. Dalle storiografie bibliografiche a quelle nazionali basate su eventi

Al posto di siti in attesa di contenuti, oggi i contenuti attendono gli utenti come nelle biblioteche i libri aspettano chi li prenda in prestito.

L'Internet Archive e la Wayback Machine sono molto citati: una *query* su un motore di ricerca con le parole produce numerosi risultati. La grande maggioranza sono citazioni di informazioni o di pezzi di letteratura scientifica sui metodi e le tecniche dell'archiviazione del web.

La questione della mancanza di «interesse da parte dei ricercatori verso gli archivi web» è stata sollevata dagli studiosi di archiviazione del web. Una delle osservazioni più pertinenti riguarda il tipo di web che andrebbe archiviato prioritariamente nel futuro. Secondo alcuni gli archivi web potrebbero essere più attraenti per gli studiosi di scienze umane se i siti contenuti fossero costituiti da materiali digitalizzati.

Come criterio di selezione dei materiali da archiviare, il salvataggio di siti che contengono media storici digitalizzati ha i suoi adepti. Costruire archivi web sembra un'attività basata su singoli progetti piuttosto che sulla continuità.

Se si dovessero definire in generale le collezioni speciali, si potrebbe affermare che esse sembrano incorporare un secondo metodo storiografico all' archiviazione del web: la storia basata su eventi. In effetti registrare «eventi importanti, come elezioni o catastrofi» è diventata una pratica consolidata. Questo impegno deriva dal lavoro del webarchivist.org.

Gli eventi pongono probabilmente le maggiori sfide agli archivisti, e contemporaneamente creano una «febbre archivistica» per l'urgenza di intraprendere l'archiviazione, dal momento che la combinazione tra la natura effimera del web in generale e i rapidi cambiamenti dei siti, quando si verificano degli eventi straordinari in particolare, causa una perdita costante di contenuti per la posterità.

La websfera è dinamica in due sensi fondamentali:

1. Per l'archivista che individua continuamente nuovi siti o nuove risorse web da includere
2. I siti che la compongono rimandano continuamente ad altri siti.

La websfera è delimitata dal tema e dalla dimensione temporale o periodicità.

Il metodo di ricerca per raccogliere i siti web utilizzato attualmente può essere avvicinato al cosiddetto «campionamento a valanga»: trovano gli URL cercando e navigando attraverso i link tra siti web collegati tematicamente; altri URL vengono loro segnalati anche attraverso il *crowd-sourcing*, li controllano per decidere se includerli nella raccolta. Ai siti vengono poi attribuite delle etichette in modo da creare dei metadati.

I costruttori dell'Internet Archive e delle collezioni speciali che usano le websfere stanno cedendo il posto di principali archivisti del web alle biblioteche nazionali che stilano liste di siti da salvare.

Almeno nell'ambito dell'archiviazione del web, i catalogatori di Internet, i bibliotecari del web, i compilatori di liste di link e altri redattori del web hanno contribuito a definire cosa sia un sito olandese in modo da rendere necessario un intervento manuale.

Normalmente i siti che vengono archiviati sono statali, di istituzioni culturali nazionali e di università, una sorta di *establishment*.

Uno degli obiettivi della riflessione sulle conseguenze delle pratiche di analisi manuali dei siti riguarda il tipo di web ottenuto una volta ultimata l'archiviazione, e il tipo di ricerche che è possibile svolgere su di esso e attraverso di esso, come ho descritto in precedenza in termini storiografici: storie o biografie di singoli siti.

5. *Rievocare uno stato passato del web*

Cosa si poteva ricavare dalla storia di un singolo sito, a parte cercare al suo interno delle prove per qualche procedura legale? Con i documentari *screencast*, o montando in sequenza i risultati della Wayback Machine, l'analisi dei siti si arricchisce della dimensione temporale. Si esplicitano inoltre le implicazioni della Wayback Machine, che spinge a raccontare la storia di un sito, e attraverso di essa la storia del web.

Si voleva anche costruire un compilatore di collezioni di siti già archiviati. La costruzione di una tale raccolta si poneva in continuità con la tendenza dell'archiviazione del web a fornire strumenti che permettano agli utenti di contribuire all'archiviazione stessa; avevano il desiderio di aggiungere un altro metodo storiografico che fosse sensibile anche alle necessità della storia del web.

Come porzione di web su cui lavorare venne scelta la prima blogosfera, insieme a Eatonweb che venne utilizzata per datare la fine della prima epoca della blogosfera.

L'ultima lista completa dei blog compilata da Eatonweb funge da lista degli URL che compongono la prima blogosfera. Ogni URL è stato inserito nella Wayback Machine, stabilendo così la percentuale di web degli esordi che è stata archiviata.

Ognuno dei siti archiviati che appartenevano alla blogosfera degli esordi è stato percorso da un *crawler* per registrare i link in uscita. Usando un software di cartografia dei link, abbiamo creato una mappa a *clusters* di questa prima blogosfera, includendo non solo i siti che si trovano nell'archivio ma anche quelli mancanti. Sulla mappa, i blog mancanti che facevano parte della prima blogosfera sono riapparsi con il loro nome, e sono diventati visibili anche i link che portavano a essi.

4. La googolizzazione e il motore di ricerca “innocente”

1. La googlizzazione e il modello “servizi in cambio di profili”

“googlizzazione” è un termine introdotto per designare il crescente insinuarsi delle tecnologie di ricerca e dell'estetica dell'azienda. Si indica, attraverso una critica, come Google si stia accaparrando uno dopo l'altro vari servizi online. Nell'ambito specifico degli studi sui media la googolizzazione viene intesa come una forma di analisi di Google in quanto medium di massa, che invita a riflettere sui criteri con cui vengono analizzati i vecchi media. Uno di questi criteri es. è la separazione tra produttori/ distributori da un lato e consumatori dei media dall'altro.

I servizi offerti da Google sono solo apparentemente gratuiti, quando utilizziamo «le ricerche sul web, l'e-mail, le piattaforme di blog o YouTube, Google registra le nostre abitudini e i nostri gusti in modo da adattare efficacemente la pubblicità al target»

Per studiare il dilagare della googolizzazione si deve indagare se il modello “servizi in cambio di profili” stia trasformando gli altri media.

Turow afferma che con Internet la pubblicità sta gradualmente abbandonando i mezzi di comunicazione di massa (*broadcast*) per la «vendita diretta». Privata ormai del contatto umano, la costruzione della relazione con il cliente si basa oggi sulla forma tecnologica scelta per raccogliere i dati dell'utente/cliente e sulla conseguente personalizzazione dei saluti, degli avvisi, delle pubblicità e dei consigli.

Il database contiene delle «chiazze» dell'utente; informazioni su abitudini e interessi, i tratti per esempio dalle *queries* fatte sul motore di ricerca, che sono utilizzate per tracciare un profilo sulla base di una piccola raccolta di pezzi di informazioni. Riunire questi pezzi de-anonimizza l'utente solo parzialmente. La profilatura dei gusti deriva essenzialmente da alcune parole-chiave tratte dalle *queries* e dalla posizione geografica, attraverso i codici postali associati all'account.

Quando l'industria è stata googolizzata ci fu il passaggio dell'interattività di consultazione all'interattività di registrazione. Nell'interattività di consultazione l'utente svolge ricerche e sceglie tra informazioni precarie; l'anonimità dell'utente non entra in gioco.

Con l'interattività di registrazione, l'informazione fornita dipende dalle impostazioni personali e dalla quantità di risultati della versione più leggera. Dal momento che le impostazioni si uniscono alle cronologie personali, la «familiarità» tra il motore di ricerca e l'utente a volte fornisce risultati che possono sembrare perturbanti.

2. La googolizzazione “back-end”

I risultati del motore di ricerca sono analizzati non solo per quello che includono o escludono ma anche per il tipo di storie che questi risultati raccontano. L'idea di costruire una storia a partire dai risultati dei motori di ricerca richiama gli scritti teorici dell'ipertesto letterario che considerano i passi datti da chi naviga in rete come mezzi di autorialità.

Invece di offrire uno spazio in cui si scontrano versioni alternative del reale, Google fornisce risposte già

note. La prevedibilità dei risultati ha messo fine all'idea che sul web tutti gli attori siano indipendenti dalla loro reputazione e dal loro prestigio: Google è diventato giornalistico nel senso che si basa sulle stesse fonti dei media ufficiali e degli *agenda-setters* dotati di grandi risorse economiche.

Quali tipi di fonti vengono consigliate per una *query*? Come si può pensare attraverso google e al suo sistema di raccomandazione? Date tutte le pagine che contengono una certa parola-chiave, il motore di ricerca fornisce quelle che meritano di essere inserite tra i primi risultati.

Gli studiosi della googolizzazione hanno riscontrato che questa creazione di status narrativo si propaga attraverso le altre piattaforme: viene da chiedersi se l'algoritmo *back-end* abbia preso il posto dei tradizionali creatori di status.

3. La googolizzazione “front-end”

Powezeck, designer delle interfacce di Technorati, ritiene che troppi motori di ricerca abbiano cercato di assomigliare a Google. Il suo argomento può essere interpretato come una preoccupazione riguardo alla googolizzazione delle interfacce. Bisognerebbe fare attenzione all'omogeneità crescente dell'home page.

La semplicità della maschera di ricerca di Google, con i suoi due pulsanti, uno per le ricerche sul web e l'altro, il famoso «mi sento fortunato», che è un omaggio all'iperspazio, è indubbiamente affascinante.

Nei suoi primi dieci anni Google ha fatto dei sottili cambiamenti alla struttura della home page. Alcuni servizi sono stati valorizzati o messi in secondo piano; altri progetti acquisiti hanno visto brevemente la luce per essere accantonati.

I risultati dei motori di ricerca che non sono sponsorizzati sono detti “organici”. In un altro caso che riguarda le sue relazioni con gli esseri umani, Google si è trovato in sintonia con Yahoo! su un progetto di importanza cruciale per bibliotecari e redattori: entrambi hanno degradato la propria directory.

Studiare la googolizzazione significa quindi indagare come sottili cambiamenti nell'interfaccia implicino una politica della conoscenza, in particolare i meccanismi svalorizzanti, attraverso la relega dei servizi editoriali nei recessi più profondi di un sito. Il fatto che sia Google che Yahoo! abbiano seppellito le loro directory è indice di un fenomeno ben più vasto: l'estromissione dei redattori esperti umani dal web, in cui rientra anche la progressiva scomparsa dei catalogatori del web retribuiti.

Un'ulteriore questione è l'impatto del potere crescente dell'utente nei confronti delle competenze editoriali o della purezza dell'algoritmo. Le ricerche sul web stanno diventando sempre più personalizzate.

Per raggiungere questo risultato, l'utente del motore di ricerca viene “registrato”, anche nel senso delle parole scelte da Google per le impostazioni.

4. Il motore di ricerca “innocente”

Gli studiosi dei media si sono chiesti come reinterpretare il ruolo del *gate keeper* alla luce delle classificazioni, determinate dalle reti di link, e delle cronologie delle ricerche. Una riflessione sulle nuove forme assunte dal *gatekeeping* dovrebbe cominciare dai casi di siti che sono stati deindicizzati.

A proposito di un *crawl* madre del 2006 chiamato *bigdaddy*, Cutts scrive che «il nostro algoritmo ha scarsa fiducia nei link in entrata e in uscita di alcuni siti, per esempio quando sussiste un numero eccessivo di link reciproci, che portano a zone del web piene di sparo».

Le ricerche personalizzate eliminano i risultati comuni a tutti per le stesse *queries* aggiungerei che la personalizzazione discolpa il motore di ricerca, perché il biasimo o la responsabilità dei risultati ricadono in parte sull'utente.

A differenza di prima, oggi l'utente ha la possibilità di retroagire sul medium perché è coautore dei risultati delle sue stesse ricerche.

5. La ricerca in internet come ricerca sociale: la distanza della fonte e l'analisi comparata delle sfere

1. Lo studio delle “queries” (“search research”) VS la “query” come strumento di studio (“search as research”)

Lo smantellamento del web modificato da esseri umani è stato analizzato attraverso i cambiamenti gradualmente all'interfaccia di Google, partendo dal presupposto che la sua storia possa in parte raccontare anche la storia del web.

Come è stato scoperto e denunciato verso la fine degli anni '90 i motori di ricerca non indicizzano l'intero internet, e possono es. trascurare i siti “orfani”, ovvero quei siti che non ricevono link. La consapevolezza dell'incapacità di percorrere tutto il web portò a elaborare nozioni come quella di *dark web* → definizione che esprimeva una critica al fatto che i motori di ricerca non riuscissero a raggiungere una parte del web, e anzi lo oscurassero attraverso le loro pratiche di esclusione.

I motori di ricerca spingevano ai primi posti le fonti che ricevevano molti link e tali risultati venivano chiamati *organici*.

I motori di ricerca oscurano ancora il web; tuttavia, oggi c'è meno consapevolezza critica e meno comprensione dei meccanismi che privilegiano alcune fonti rispetto ad altre. Il motivo è che il contenuto, posto a metà strada tra informazione e pubblicità, cerca continuamente di salire ai primi posti dei risultati forniti dai motori di ricerca.

I cambiamenti apportati all'algoritmo di Google hanno lo scopo di penalizzare i siti costruiti apposta per i motori di ricerca. Le *content farms*, o «fabbriche di contenuti», di scarsa qualità sono state degradate nel ranking e sono state equiparate dai motori di ricerca alle *link farms*, o «fabbriche di link», come principali fonti di inquinamento del web. Le *content farms* sono più recenti, ma le *link farms* continuano a esistere.

2. Il motore di ricerca come macchina per la ricerca o strumento per il consumatore

Da questa prospettiva i motori di ricerca diventano macchine socioepistemologiche, che determinano la posizione di una fonte per un certo tema, materia o argomento.

Il prestigio online è costruito su un'associazione diretta e nominale, ma anche su una serie di altri criteri come numero di visualizzazione, frequenza degli argomenti e post, età del sito, numero di “mi piace” ricevuti. Le combinazioni chiave applicate alla reputazione dell'autore si differenziano in certo modo per ogni spazio o sfera del web.

L'importanza di una fonte rispetta la cultura delle sfere, nel senso che si basa sulle differenze delle pratiche degli amministratori dei siti nella websfera, dei blogger nella blogosfera e dei mezzi di informazione nell'infosfera.

Dal nostro punto di vista, bisogna chiedersi se sia possibile utilizzare il prestigio di una fonte, creato dai motori di ricerca, per la ricerca sociale, e con quali fini.

Perché la ricerca attraverso le *queries* possa diventare ricerca sociale, bisogna chiedersi più in generale come interrogare Google e come interpretare i risultati che fornisce.

Nel valutare se sia possibile usare i motori di ricerca come strumenti per la ricerca scientifica ai fini di identificare le fonti più rilevanti in rapporto a una *query*, è essenziale lo statuto della rilevanza. Le aziende che possedevano i motori di ricerca hanno gradualmente cambiato la definizione di rilevanza di una fonte, passando da uno schema di valutazione basato sul conteggio dei link in entrata a un criterio che prendeva in considerazione anche il conteggio dei clic e il fatto che le fonti fossero state messe online di recente. Con questo sistema le fonti vengono spinte verso i primi risultati se gli utenti le hanno cliccate nel corso di precedenti ricerche. Le fonti salgono nell'ordine dei risultati se sono recenti, elevando in tal modo lo status epistemologico dell'attualità.

3. La “dimensione locale” secondo Google

Una seconda serie di osservazioni riguarda i casi in cui la dimensione locale di Google può essere utilizzata per la ricerca sociale, specialmente per quanto riguarda i domini locali e il modo in cui si valutano le fonti.

La questione è se Google sia uno strumento localizzante o globalizzante, oppure qualcos'altro ancora. Si può cominciare con il chiedersi se, quando lavora come motore di ricerca locale, Google fornisca come risultati solo fonti provenienti da quell'area o se fornisca risultati nella lingua locale.

Con i suoi circa 150 domini locali, Google può essere considerato una macchina globalizzante. In uno studio derivato dall'analisi del senso del locale per Google bisogna chiedersi se per fare analisi comparate tra paesi si debba utilizzare come *corpus* qualche dominio locale o tutti i domini locali, quelli in cui le lingue sono relativamente specifiche di un paese, o quelli che sono in grado di organizzare i risultati locali. Tipi di studi:

1. *The world according to Google* lanciava le *query* [“diritti umani”] in oltre 150 domini locali di Google; solo 25 hanno fornito una maggioranza di risultati provenienti dal paese di origine.
2. L'analisi ha cercato di chiarire ulteriormente la nozione di locale secondo Google, interrogando in determinate lingue temi riguardanti il bacino del Rio delle Amazzoni, per verificare quali fonti abbiano il privilegio di discutere i problemi locali. La prima *query* è stata lanciata nei domini locali di tre paesi, sul cui territorio si estende il bacino del Rio delle Amazzoni: Colombia, Perù e Venezuela. La maggior parte dei risultati erano fonti provenienti dalla Spagna.
L'aspetto ancora più importante è che i risultati erano molto simili in tutti i paesi latinoamericani, come se esistesse un'unica serie di risultati per tutta l'America Latina: il concetto di locale per Google è molto più ampio di quello definito dai domini dei singoli paesi.

I domini locali di Google sono stati interrogati per la parola *diritti* nelle rispettive lingue, poi sono stati estratti i tipi di diritti lasciandoli nell'ordine in cui Google li aveva forniti. Attraverso una formulazione aperta della domanda, ai motori di ricerca è stato chiesto di classificare delle preoccupazioni sociali invece che la sola informazione. I ricercatori potevano scegliere una particolare impostazione per facilitare una prospettiva di ricerca che si distinguesse dall'analisi o dallo studio di Google.

4. Lo studio delle “queries”

Altri oggetti interessanti per gli analisti sono caratterizzati da un uso alternativo del motore di ricerca, come per esempio il *Google bomb* (bombardamento di Google), che porta ai primi posti dei risultati di

Google un sito che normalmente non dovrebbe comparire per la *query* in questione. Tra i numerosi casi di *Google bomb*, ce ne sono alcuni in cui la manipolazione del motore di ricerca viene fatta per scopi politici più che per scopi commerciali.

Un altro genere di *query* che è stato studiato è quello che ottiene risultati offensivi. Il fatto che Google lo mantenga fa sì che il risultato offensivo possa essere visto come prova della mancanza di controllo manuale sui risultati.

Rogers considera Google autore dei suoi risultati, nel senso che determina la posizione delle fonti per ciascuna lista di risultati, sebbene la posizione non sia attribuita manualmente da un operatore. Il motore di ricerca continuerà a fornire il risultato offensivo.

Riuscire a entrare tra i risultati forniti dal motore di ricerca è un traguardo difficile. la personalizzazione costituisce un indebolimento, perché implica che l'utente sia coautore dei risultati delle sue stesse ricerche. Nel prepararsi a utilizzare il motore come macchina per la ricerca e nel concepire le *queries*, è necessario eliminare prima tutte le impostazioni personali.

Il motore di ricerca è coinvolto in un dibattito sul suo ruolo come valutatore e come autore di una lista di fonti, sia che i contenuti dei siti siano goliardici sia che siano offensivi.

5. Interpretare i risultati per la ricerca sociale

Il punto di partenza della ricerca in Internet come ricerca sociale può essere sintetizzato con una formula un po' estrema: quando guardiamo i risultati di Google non vediamo Google, ma la società. I motori di ricerca mettono in mostra molte cose, la competizione tra le fonti, la loro longevità e l'impegno necessario per ottenerla.

La progettazione della *query* è la pratica di rivolgere una richiesta al motore di ricerca in modo che i risultati possano essere interpretati come indicazioni e scoperte. Per preparare l'analisi vera e propria, il ricercatore che utilizza i metodi digitali deve installare un browser di ricerca «pulito». Un'alternativa alla scomodità di disconnettersi dal motore di ricerca è utilizzare uno *scraper* → interroga la versione ncr (*no country redirect*) di Google, che reindirizza sulla versione pura google.com

Bisogna considerare la *query* come una domanda scientifica posta al motore di ricerca e formularla attentamente. Il punto fondamentale è stabilire quali parole-chiave e quali operatori di ricerca utilizzare per interrogare una serie di fonti.

L'indicizzazione, il lavoro di ricerca si basa soprattutto sulle capacità del motore di ricerca di indicizzare i singoli siti web, in modo che il ricercatore possa contare il numero di volte in cui un determinato termine compare in un sito.

Utilizzando gli intervalli di data, si può iniziare ad analizzare i cambiamenti delle strategie di gestione delle campagne nel corso del tempo. I risultati possono anche essere letti e interpretati in base alla composizione dei tipi di fonti che appaiono ai primi posti, o in base alla familiarità delle fonti fornite con altri media.

6. La distanza della fonte: lo studio epistemologico dei risultati del motore di ricerca

Alcune fonti vincono la competizione per entrare tra i primi risultati. Le critiche alle nuove gerarchie hanno dato vita ad alcuni progetti di *new media art* che mette in ordine casuale i risultati di una *query* su Google, basandosi sulla domanda «e se quello che stai cercando si trovasse nella 53ª pagina?».

La “distanza dalla fonte” → illustra il modo in cui caratterizzare la priorità accordata a certe fonti dal motore di ricerca. Rappresenta lo studio della distanza di una fonte dal primo risultato per una certa *query*. Con questo metodo è possibile studiare se le fonti che si trovano ai primi posti siano dello stesso tipo o condividano una particolare tendenza. Il metodo possiede anche una dimensione temporale, per valutare se per una determinata *query* le fonti che si trovano ai primi posti si mantengano stabili o manifestino una certa variabilità. La distanza dalla fonte permette di studiare i motori di ricerca, evitando di concentrare l'attenzione unicamente sui primi risultati. La domanda è se per un dato tema le stesse fonti abbiano la stessa posizione dominante anche sul web.

Google scraper è stato sviluppato per trasformare la ricerca in Internet in ricerca scientifica, registrando i risultati forniti dal motore di ricerca per una qualsiasi *query* e salvandoli per una successiva analisi. Lo strumento permette anche di esaminare la predominanza di una fonte per una data *query* e le questioni legate alla stabilità o variabilità dei risultati forniti dal motore di ricerca nel corso del tempo.

Per prima cosa il ricercatore interroga il motore di ricerca. I primi 100 o più risultati ottenuti vengono privati del testo di descrizione e di tutto il resto, in modo che rimangano solo gli URL. Ogni URL, o ogni dominio, viene inserito in Google Scraper e interrogato. Google scraper interroga ogni sito per ciascuna parola-chiave e fornisce i risultati sotto forma di nuvola di siti. Il risultato è una nuvola di siti ordinati da google, nel senso che essi si presentano nello stesso ordine in cui li ha forniti il motore di ricerca. Ogni nome del dominio è ridimensionato in base al numero di volte che menziona la parola-chiave.

Google Scraper è stato ribattezzato dispositivo lippmanniano, in onore di Walter Lippmann e del suo tentativo di trovare degli strumenti per individuare le inclinazioni, le prese di posizione e più gli schieramenti. Vuole essere uno strumento di uso comune per analizzare gli spazi tematici.

Le nuvole di tag possono essere rinominate “nuvole di temi” (*issue clouds*) o “nuvole di fonti” (*source clouds*) a seconda di quale output venga scelto dall'utente.

La varietà degli algoritmi l'andata gradualmente diminuendo, almeno tra i motori di ricerca più usati, con l'affermarsi in tutto il mondo di Google e con la chiusura e il riposizionamento dei motori di ricerca nazionali. Google ha dato inizio a una nuova forma di concentrazione mediatica che può essere definita “concentrazione algoritmica”. I motori di ricerca hanno logiche di classificazione diverse per la websfera, la blogosfera e l'infosfera, in cui le fonti vengono privilegiate in base a diversi mix di variabili e di segnali.

7. L'analisi comparata delle sfere

L'uso del termine *comparato* nella ricerca sociale si riferisce di solito a confronti tra paesi. Negli studi sui media esso può riferirsi all'analisi che paragona le diverse forme mediatiche. Questo tipo di analisi studia anche come ciascun medium catturi l'attenzione del pubblico.

Negli studi comparati sui media ci si concentra su come la storia o la narrazione rimangano le stesse, nonostante i contenitori mediatici diano loro forme diverse. L'analisi comparata delle sfere si basa sul

confronto di sostanza, copertura e argomento attraverso le diverse forme mediatiche, e applica i risultati al web.

Una sfera è costituita da una serie di fonti delimitate da un dispositivo. Per studiare una sfera è necessario delimitarla. L'aspetto importante di questo tipo di analisi è che essa permette di considerare le conseguenze del trattamento dei link. È quindi possibile verificare se per esempio alcune particolari fonti tendano ad avere posizioni predominanti in alcune sfere ed essere assenti in altre.

Il gruppo di ricerca di Rogers si è concentrato sul cambiamento climatico. → Sono state individuate e sottoposte a triangolazione delle parole-chiave per l'oggetto di studio in questione; è stata lanciata una *query* nei motori dominanti della websfera, blogosfera e infosfera. Il nome precedentemente individuato è stato poi utilizzato per interrogare ognuno dei risultati ottenuti nella prima fase. Oltre alla risonanza delle parole-chiave in ciascuna sfera, nel corso dell'analisi sono state anche registrate, conteggiate e ridimensionate delle immagini, creando una nuvola di immagini. I risultati ottenuti:

- Sfera dell'informazione → l'animale più legato al cambiamento climatico è stato l'orso polare, risultato amplificato nella blogosfera.
- Websfera → le parole-chiave legate al cambiamento climatico sono più distribuite

8. La ricerca in internet come ricerca sociale

L'introduzione di gerarchie di fonti e pratiche di esclusione in un medium considerato egualitario e democratico è stato il principale carattere messo in luce dalle prime obiezioni al motore di ricerca. Più recentemente, il potere di Google è stato identificato con l'autorità del suo algoritmo o con la credenza nel valore epistemologico dei risultati forniti dal motore di ricerca.

Si può applicare il lavoro del motore di ricerca ai modi in cui delimita gli spazi del web? Risposta non immediata perché motore di ricerca e utente collaborano nella produzione dei risultati. Per fare confronti tra paesi utilizzando i domini locali di Google, è necessario prendere familiarità con il senso del locale secondo Google. Bisogna condurre delle analisi nelle principali aree linguistiche presenti online. Esiste una gerarchia tra quelli che forniscono più risultati locali e quelli che ne forniscono meno, con il nord del mondo che fornisce il maggior numero di fonti locali per la *query*.

Mentre l'analisi del senso del locale di Google diventa in sé un lavoro di ricerca, esiste una seconda pratica che è maggiormente in linea con il tentativo più vasto di fare ricerca *con* il web e non solo *sul* web. Il dispositivo lippmanniano sfrutta la capacità del motore di ricerca di indicizzare i singoli siti per poterli interrogare per parole-chiave.

Per scoprire quali fonti citino per nome gli scienziati scettici sul cambiamento climatico e quali no, si possono seguire due metodi:

1. «distanza della fonte» che misura quante le fonti che menzionano gli scettici siano lontane dalla cima dei risultati per la *query*
2. Allineamento dei temi; una risposta all'appello di Lippmann per trovare un sistema semplice e oggettivo, finalizzato a determinare la posizione di un attore sociale.
3. Si basa sul fatto che il motore di ricerca suddivide il web in sfere. Queste sono considerate serie di fonti delimitate da un dispositivo e motore di ricerca, classificate e fornite come risultati di *queries*.

6. I web nazionali

Come caso studio è stato scelto l'Iran, un paese in cui la censura di Stato ha un peso particolarmente rilevante; nonostante ciò, il web iraniano si è rivelato molto reattivo. Intere porzioni del web iraniano brulicano di blog in attività, mostrando un'attiva evasione dalla censura e l'uso costante di un linguaggio considerato critico dal regime.

1. Come studiare i (domini) web nazionali

Il concetto di web nazionale riassume la transizione di internet dal cyberspazio.

Basandosi su un lavoro di caratterizzazione del web, questo metodo implica:

- una serie di dibattiti metodologici su come studiare un web nazionale → questo metodo tiene conto sia della varietà di modi con cui gli utenti fanno esperienza di Internet, sia delle concomitanti pratiche di raccolta di dati sul web.
- fornisce le ragioni complessive di tale studio.

I prodotti di questa attività di raccolta e conteggio sono di solito liste di URL classificati e consigliati agli utenti. Quando alle variabili si aggiunge la posizione geografica dell'utente, la lista degli URL può essere specifica per un paese o una regione, e lo stesso vale per le lingue. In questo modo si può parlare di web *country specific / language-specific*.

La personalizzazione può avere influenza sugli URL, che vengono consigliati in un paese specifico o in una determinata lingua.

Le culture dei dispositivi organizzano i diversi ambiti dei web nazionali. La pratica di ricerca fa uso degli strumenti del web che agiscono nella dimensione locale, ma forniscono il web su base territoriale. Un motore di ricerca può fornire siti in una specifica lingua, che possono avere origine nel paese o fuori di esso.

Nel dibattito sulla decadenza del cyberspazio e sull'affermarsi di un web consapevole della posizione geografica, si instaura quindi una tensione tra due nuovi modi di interpretare l'oggetto di studio: web nazionale contrapposto a web linguistico.

Si può ad esempio analizzare se gli URL classificati come *top blog* dagli aggregatori di blog assomiglino agli URL classificati come interessanti dalle piattaforme di *cloudsourcing*.

Lo studio dei web nazionali non implica soltanto una critica del web come spazio senza luogo e universale, ma invita anche a sviluppare ulteriormente l'analisi della relazione tra misurazione del web e indicatori del territorio.

2. "blocked yet blogging": il caso particolare dell'Iran

Quello dell'Iran è un caso particolare sotto vari aspetti. L'esperienza del web di chi vive in Iran è diversa da quella di chi scrive fuori dei suoi confini. Internet appare scritto in modo diverso fuori e dentro i confini del paese. Di conseguenza, molti iraniani che scrivono sul web devono fare i conti con la censura. Fare i conti con la censura significa tenersi informati su quali siano le parole vietate ed evitare di avvalersene. L'aspetto più interessante è il grado in cui gli iraniani si esprimono online nonostante le

restrizioni. Ci si può proteggere attraverso l'attenta scelta di un software o di una piattaforma rispetto a un'altra, a seconda di quale fornisca la migliore protezione e le migliori forme di anonimato.

Per questo tipo di analisi si rivela particolarmente utile considerare separatamente i diversi web nel corso della procedura di campionamento. I blogger iraniani che vengono letti grazie a Google Reader e indicizzati da Likekhor, pur essendo completamente bloccati dallo Stato, continuano tuttavia a scrivere sui loro blog: *blocked yet blogging*.

Esiste molta letteratura scientifica a cui attingere per studiare i web nazionali, a partire dal pionieristico studio etnografico sull'uso del web a Trinidad e Tobago, con cui si esprime la cultura locale più che la cultura globale.

Anche nelle scienze politiche i web nazionali sono sempre più spesso cartografati per alimentare dibattiti sul ruolo del web. È interessante il tentativo di costruire strumenti per aggirare la censura in modo che queste voci possano essere udite.

3. La definizione di un web nazionale e le sue conseguenze sull'archiviazione

La definizione di un sito come «olandese» si basa sui seguenti criteri: il sito è in lingua olandese ed è registrato in Olanda; è in qualsiasi lingua, ma è registrato in Olanda; è in olandese, ma registrato in un altro paese; non è in lingua olandese, né registrato in Olanda, ma tratta argomenti riguardanti l'Olanda.

Si può cominciare con i siti del dominio nazionale .nl, sia in olandese che in altre lingue, che possono essere individuati automaticamente da un software. Da questa lista si devono poi rimuovere i siti con dominio .be, corrispondente al Belgio.

Quello della Biblioteca nazionale olandese potrebbe essere definito un metodo relazionale, dal momento che i siti che parlano dell'olandese o quelli in lingua olandese sono registrati in un altro paese. Nella pratica molti paesi usano URL esterni al loro dominio nazionale.

In una ricerca preliminare sulla nozione di web iraniano, uno studente ha svolto un'indagine tra i blogger iraniani usando Google Reader nella sua rete Gooder. Agli intervistati sono state proposte varie definizioni di web nazionale. All'inizio la domanda veniva accolta con sospetto, perché l'espressione stessa *web nazionale* era interpretata come un possibile espediente del governo iraniano per creare la propria rete Internet.

Sulla base dei risultati di questo sondaggio, si è concluso che il web nazionale iraniano possa essere definito come scritto e prodotto da iraniani, indipendentemente dalla lingua che usano, dal luogo da cui scrivono e dal soggetto che trattano. Questa definizione di web nazionale comprende siti con contenuti prodotti da iraniani, che vivono all'estero e scrivono in una lingua diversa dal persiano, su argomenti che possono anche non riguardare l'Iran. Questa definizione rende praticamente impossibile delimitare un web iraniano.

La definizione adottata dalla Biblioteca nazionale olandese richiede una selezione manuale, ma non estende la sua definizione a siti scritti da olandesi che vivono fuori dall'Olanda.

4. Delimitare il web iraniano attraverso le culture dei dispositivi

Con «web nominale» si intende un web definito in base agli strumenti, con i quali è organizzato tramite i dispositivi e le piattaforme online, da cui viene visitato sia dall'utente che dal ricercatore.

Il web fornito dalle tre piattaforme *crowd-sourcing* per gli utenti iraniani si differenzia da quello raccolto da uno strumento di marketing per gli inserzionisti in lingua persiana.

Per studiare i web nazionali si dovrebbe innanzi tutto conoscere la popolazione di un web ed essere in grado di estrarne un campione.

Per questo progetto di ricerca è stata intrapresa un'attività di *scraping* e *querying* di media entità, evitando l'arena tecnico-amministrativa cercando di utilizzare quanto è disponibile ai normali utenti del web. Utilizzare gli strumenti del web per la ricerca presenta due vantaggi metodologici:

1. gli strumenti più usati possono essere considerati come mediatori e quantificatori di un uso specifico.
2. la definizione del web iraniano così ottenuta evita di ricorrere alla metodologia che si avvale di grandi masse di dati per ordinare il contenuto e che combina tecniche algoritmiche con la partecipazione su larga scala degli utenti.

La rifinalizzazione dei dispositivi del web è sia una strategia per ricavare da grandi masse di dati piccole serie di dati da usare come campione per la ricerca, sia un mezzo per ottenere campioni.

Lo scopo delle analisi è registrare la lingua e altre proprietà formali in ogni settore del web iraniano.

Un altro aspetto di interesse è l'ampiezza con cui ciascun settore del web è censurato o filtrato dallo Stato, e se ci sia una relazione tra i siti web reattivi e aggiornati e quelli filtrati. Per valutare il successo della censura dal punto di vista dei censori sono stati usati dei dati temporali tratti da Balatarin, una delle principali piattaforme *crowd-sourced* iraniane, che è stata analizzata con uno *scraper* per paragonare gli URL significativi, con quelli dello stesso periodo.

5. Le culture dei dispositivi: valutare e classificare i siti web

Il *crowdsourcing* → deriva dalla pratica dell'*outsourcing*, ovvero dell'esternalizzazione, per indicare un processo in cui non solo la cosiddetta «saggezza della folla», ma anche il lavoro della collettività tornano utili a un beneficiario. Il *geoweb* o web localizzante indica il mezzo con cui vengono forniti i siti web. L'economia delle visite classifica i siti in base al numero di contatti o visualizzazioni.

Anche se l'Iran non è tra i paesi contenuti nella lista di Google Ad Planner, probabilmente a causa dell'assenza del dominio *google.ir* e delle sanzioni economiche statunitensi contro l'Iran, una delle categorie di siti disponibili nelle statistiche sul pubblico è «sito in lingua persiana».

L'espressione *economia dei link* registra anche lo slittamento delle logiche di classificazione degli URL da un modello pubblicitario, basato sul conteggio delle visite, a un criterio più bibliografico e scientometrico, basato sul conteggio dei link.

I siti *crowd-sourced* come Balatarin richiedono di registrarsi prima di poter segnalare un link, che viene poi votato dagli utenti registrati: gli URL che ricevono più voti vengono spinti ai primi posti. Lanciato nel

2006, Balatarin è considerato il primo sito web 2.0 in lingua persiana, ed è stato riconosciuto come uno dei più popolari siti in persiano nel 2007 e 2008.

L'introduzione del pulsante «mi piace» e altre misure simili nei social media hanno dato vita a quella che può essere definita «economia dei like», che valuta i contenuti basandosi sull'attività dei pulsanti sui social media.

Likekhor si concentra in particolare sui blog, facendo emergere una relazione tra gli utenti di Google Reader, i blogger e i lettori di blog. Da Likekhor abbiamo estratto una lista di 2.600 host, che sono stati raccolti da una pagina dove sono elencati tutti i blog contenuti su questa piattaforma.

6. Analizzare le caratteristiche del web iraniano; lingue e reattività

Un'area di ricerca su cui basarsi per studiare i web nazionali è costituita dagli studi di caratterizzazione del web. Una delle maggiori difficoltà è come ottenere un campione rappresentativo di un web nazionale o di altri tipi di web. La lista dei web nazionali è composta da siti con lo stesso ccTLD. Il metodo adottato per la ricerca cerca di prendere in considerazione i siti con dominio .com, .net, .org ecc., quando vengono considerati rilevanti per il pubblico iraniano o di lingua persiana dagli strumenti e dalle piattaforme sotto osservazione. Alle tecniche di campionamento si aggiunge lo *scraping*.

A eccezione dei web di ricerca, la percentuale di siti .ir tra gli host significativi forniti dai dispositivi si è rivelata relativamente bassa.

Dopo aver passato in rassegna in che modo vengano costruiti i campioni, [Baeza-Yates](#) e colleghi hanno confrontato dieci web nazionali per arrivare a una serie di misure fondamentali condivise dalla maggior parte di essi.

I codici producono la cosiddetta “reattività”, che è considerata un indice fondamentale della salute di un sito, insieme all'età della pagina che ne misura la freschezza. La discontinuità può essere rilevata attraverso i validatori di link. Stabilire se un sito sia parcheggiato o se abbia subito l'attacco di un hacker può servire per verificare se sia stato abbandonato dai precedenti proprietari

6.1 il web iraniano e le sue lingue

Per individuare le lingue dei siti web, il gruppo di ricerca ha utilizzato uno strumento che fa uso della funzionalità di riconoscimento linguistico di alchemyAPI. In una seconda fase, i risultati sono verificati manualmente da un ricercatore. Circa due terzi dei siti del web iraniano sono in persiano, mentre un quinto è in inglese.

La blogosfera, e in grado leggermente minore il geoweb, sono i più legati all'idea di un web nazionale unicamente in lingua persiana, sebbene tra di essi sussista una media del 10% di siti in altre lingue, una percentuale da non sottovalutare. I web iraniani con le più vaste percentuali di siti in altre lingue sono quelli degli inserzionisti e il web regionale.

6.2 la reattività del web iraniano

Per analizzare la reattività dei web iraniani i codici di stato http degli host unici sono stati recuperati dall'Olanda grazie a uno strumento costruito *ad hoc*. I dati inseriti in questo strumento sono costituiti dalle liste di siti raccolte in precedenza in ciascun web. Analizzando i risultati si è scoperto che esistono dei codici abitualmente forniti negli spazi web iraniani.

Il reindirizzamento non è necessariamente un'indicazione di mancanza di reattività e può essere dovuto a una serie di ragioni, tra cui il trasferimento di vari nomi di dominio nello stesso luogo. Tuttavia, i reindirizzamenti possono anche essere messaggi. I risultati di questa parte della ricerca rivelano in primo luogo che nel complesso i web iraniani sono relativamente in buona salute.

7. *Il web iraniano e la censura di internet*

I dispositivi del web sono probabilmente gli strumenti meglio informati sulla censura in Internet e quindi i più adatti a monitorarla. I motori di ricerca e le piattaforme ricevono richieste di cancellazione di contenuti e in questo modo favoriscono la creazione di una lista nera in divenire e di un indice della censura. Nel caso del web iraniano non ci sono testimonianze di richieste di rimuovere contenuti rivolte a Google dal governo.

L'accessibilità all'interno del paese dei vari web iraniani raccolti è stata sottoposta a verifica utilizzando dei proxy. Censorship Explorer compila elenchi aggiornati di proxy nel paese e può essere usato per verificare quasi siti siano bloccati.

Sebbene effettuare i test utilizzando i proxy non garantisca di replicare fedelmente l'esperienza dell'utente medio, la verifica dei codici risposta attraverso i proxy fornisce indicazioni su specifiche modalità di censura di Internet, per esempio il blocco degli URL e degli indirizzi IP attraverso tecniche come il filtraggio delle intestazioni TCP/IP.

Per questa richiesta sono stati utilizzati 12 proxy situati in sei diverse città iraniane e utilizzati da diversi proprietari. Tra i vari web sotto osservazione, i siti *crowd-sourced* e quelli elencati da Likekhor sono quelli più bloccati, sollevando non solo la questione della sostanza di questi spazi ma anche dell'utilità di utilizzare le piattaforme per ottenere liste di URL da monitorare.

8. *Il web iraniano e la "freschezza" del contenuto*

La domanda posta è se la censura sopprima i contenuti, o se invece, nonostante i siti vengano bloccati, i blogger continuino a scrivere sui propri blog. Bisogna controllare anche se questi siti siano attivi.

L'oggetto di studio è un sottoinsieme composto sia dai siti bloccati del web *crowd-sourced* e di quello dei blogger. In questo caso perché un blog sia considerato fresco, il suo autore deve avere pubblicato almeno un post attraverso un feed nell'ultimo mese, a partire dall'ultima volta in cui il sito è stato controllato per verificare se fosse censurato. Nello studio sull'Iran il 65% dei siti presenti contenuti freschi. Questi risultati confermano che difficilmente la censura riesce a reprimere il contenuto.

9. *Il web iraniano: voci critiche e libertà di espressione*

L'utilizzo del web per ottenere informazioni sulle condizioni del campo misura la forza delle voci dell'opposizione e il grado di libertà di espressione in periodi difficili. Negli ultimi anni le voci critiche sono state soppresse e la libertà di espressione soffocata? Come misurare questi fenomeni? Prima delle elezioni del 2009 e del movimento conosciuto come «Movimento verde», i due ricercatori avevano ipotizzato che la blogosfera iraniana organizzasse le voci critiche in un modo particolare.

Tra i dati che detiene, Balatarin conserva la data in cui ogni URL è stato postato sul suo sito fin dal 2006. Per la ricerca, il database di Balatarin è stato percorso da uno *scraper* in modo da ottenere gli URL più consultati. La lista era costituita da parole che in persiano sono considerate sensibili, critiche o provocatorie. La lista è stata suddivisa in tre categorie, ogni termine potendo appartenere a più categorie:

1. «linguaggio provocatorio» → linguaggio incendiario che, se utilizzato, porta alla censura di un blog o di un sito.
2. «linguaggio schierato» → fa uso di termini che esprimono esplicitamente una certa posizione politica.
3. «linguaggio in codice» → parole che vengono utilizzate in modo da evitare la censura.

Tutte le parole sono state selezionate in quanto importanti forme di espressione. La distinzione fra le categorie era dettata dalla volontà di trovare un senso ai cambiamenti di comportamento.

In generale l'uso di parole sensibili non è diminuito, ma è rimasto stabile, e anzi è cresciuto nel corso delle tre estati in cui si è svolta la ricerca. In quello stesso periodo l'uso delle parole provocatorie è aumentato il linguaggio provocatorio non si è autocensurato, né ha fatto maggiore ricorso a parole in codice, e il linguaggio schierato non ha assunto toni più moderati. La lista di parole sensibili è stata mantenuta costante.

Nell'estate del 2009, dopo le elezioni presidenziali, ci fu un significativo incremento del linguaggio provocatorio e schierato, e anche di quello in codice. Negli anni successivi, quando ci si sarebbe potuti aspettare una diminuzione del loro uso ci verificò invece un ulteriore aumento. Da questi risultati si può desumere l'esistenza di un robusto pubblico per questo tipo di linguaggio, e probabilmente di un'abituale circumnavigazione della censura, se si assume che la maggior parte dei lettori di questi siti vivano in Iran.

La ricerca non ha rilevato tendenze specifiche nella censura dei siti in questione, perché i dati sulla censura risalgono solo all'ultimo periodo, ovvero all'estate del 2011. Tuttavia, quello iraniano è un web reattivo, in cui i blog vengono aggiornati con regolarità nonostante la censura, e molto probabilmente con un pubblico reattivo di lettori, che risiedono sia all'interno che all'esterno del paese.

10. L'indice di salute di un web nazionale

Le culture dei dispositivi possono essere definite in modo più preciso come l'interazione fra l'utente e il motore di ricerca.

I web nazionali vengono delimitati grazie a dispositivi che agiscono a livello locale, mentre la posizione geografica o la lingua vengono aggiunte come un valore che definisce gli URL. Nell'esaminare le serie di dati si è scoperto che la maggior parte degli host raccolti nei vari web iraniani non appartengono al dominio .ir ma .com.

Il contributo di questa ricerca agli studi sulla caratterizzazione del web nazionale è duplice:

- a. Concettuale → nel senso che i criteri di caratterizzazione di un web nazionale sono utilizzati anche come indici del suo stato di salute.
- b. generalizzabile a tutti i paesi che sono esposti alla censura da parte dello Stato, e sono stati applicati al caso di studio dell'Iran.

Il metodo ha permesso di trovare un buon numero di blog che inviavano un codice risposta positivo, nonostante fossero censurati. Questo risultato indica che esiste un pubblico che legge questi contenuti, sia al di fuori dell'Iran che al suo interno.

Lo studio delle condizioni di un web nazionale fornisce una serie di informazioni aggiuntive che riguardano lo stato attuale delle università, dei ministeri e delle altre istituzioni di un paese. Verificando quali siti siano attivi e quali siano invece trascurati, questo metodo può diventare fonte di uno studio comparativo e stimolare l'analisi delle condizioni di uno o più web.

7. Social media e studi postdemografici

1. Studi postdemografici?

Il metodo postdemografico si differenzia radicalmente dai procedimenti usati dalla demografia tradizionale per analizzare sociologicamente gruppi, target di mercato ed elettori. Questo metodo marca uno scarto teorico dall'uso «biopolitico» della demografia a un uso «infopolitico». La definizione degli studi postdemografici apre la strada a nuovi modi di analizzare i social network, che abbandonano le tradizionali categorie demografiche, per interessarsi ai gusti, interessi e preferenze.

Normalmente i demografi analizzano i documenti ufficiali dove sono registrati nascite, decessi e matrimoni, e intervistano le popolazioni: il censimento è il più conosciuto di tali procedimenti. I profilatori invece registrano e utilizzano i dati inseriti dagli utenti delle piattaforme online per creare e mantenere le relazioni sociali su quelle stesse piattaforme.

L'approccio demografico tradizionale studierebbe i nativi digitali come una categoria, per analizzare cosa li differenzi dalle generazioni precedenti; una ricerca postdemografica invece non si interessa tanto al nuovo *digital divide* quanto piuttosto al modo in cui i profilati consigliano informazioni, prodotti culturali, eventi o altre persone.

2. I social network come oggetto di studio postdemografico

«Definiamo i social network come siti dove gli utenti possono creare un profilo e connetterlo con altri profili allo scopo di creare una propria rete personale esplicita»; i non utilizzatori sono coloro che non gestiscono relazioni amicali o sentimentali, ma comunque visitano i siti e leggono i profili degli altri.

Con gli studi postdemografici si intende contribuire all'analisi dei non utilizzatori, condotta dai ricercatori e dai profilatori, che raccolgono i dati dai social network per studiarli o per costruire software particolari, come i *mashups*.

Altre critiche che venivano rivolte alle prime pratiche di profilatura sottolineavano come il risultato più significativo delle analisi fosse l'«anomalia». Alcune persone, intese come l'insieme di dati che ne delineavano il profilo, spiccavano tra le altre perché non rientravano nella normalità statistica. Con le più recenti piattaforme online non sussistono più limitazioni al numero di caratteri; si mettono a disposizione più campi e una maggiore possibilità di rappresentare sé stessi o, come ha acutamente scritto uno studioso, di «digitare la propria essenza».

3. “tu sei il medium”

Se prima i link collegavano i siti formando un enorme spazio ipertestuale, ora la sfera dei social media è considerata come un “iperoggetto”; il web qui è visto come uno spazio sociale che come uno spazio di informazioni.

Per creare il proprio profilo, l'utente è invitato a fornire alcune informazioni personali e una lista di preferenze. Oltre a richiedere le classiche informazioni demografiche. Una volta completato il profilo comincia l'attività social. La socialità fa aumentare le attività: più l'attività social di una persona è

intensa, più la sua presenza in questo spazio diventa rilevante. I legami che si formano e le attività che vengono registrate offrono agli studiosi di social media e ai profilatori infinite possibilità di analisi.

Rendere anonimi i dati può essere una risposta ad alcuni dibattiti etici sull'analisi dei social network. Esistono delle regole per l'utilizzo dei dati, la più basilare delle quali è il consenso dell'interessato. Nei social network l'utente si assume sempre più la responsabilità della propria privacy attraverso la scelta delle impostazioni. Mentre i servizi hanno pensato il sistema attraverso le impostazioni di default. La privacy è «personalizzabile» nel senso che le azioni di qualcuno possono essere rese visibili ad alcune persone specifiche. Un'altra serie di pratiche di profilatura non si interessa ai dati personali, ma piuttosto ai gusti e soprattutto alle relazioni che intercorrono tra i gusti.

4. Macchine postdemografiche

Le macchine postdemografiche più interessanti sono proprio i social network stessi, dal momento che raccolgono i gusti degli utenti e li mostrano agli altri. es. → Elfriendo.com, che utilizza i profili di Myspace. Inserendo un singolo interesse su Elfriendo, il dispositivo crea un nuovo profilo basato su quelli delle persone che esprimono lo stesso interesse. Il software è anche in grado di verificare se uno o più interessi siano compatibili tra loro.

Myspace appare meno “recintato” di Facebook, nel senso che permette ai profilatori di vedere gli amici di un utente e i loro profili anche senza aver chiesto loro l'amicizia.

5. Il progetto Leaky Garden

«Per funzionare correttamente, i social network richiedono un certo grado di esclusività». Comunemente associata ad alcuni social network, l'espressione «giardino recintato». Uno degli argomenti a favore di questi sistemi chiusi è che gli hardware dedicati assicurano il corretto funzionamento delle relative tecnologie. Nel caso dei social network il concetto di giardino recintato non può essere applicato agevolmente. Questi siti incoraggiano le applicazioni prodotte da terzi, perché si sono resi conto che il contenuto prodotto dagli utenti, anche le loro applicazioni aumentano il valore e il livello di partecipazione degli utenti.

Nella ricerca, usernamecheck.com è stato utilizzato come macchina per costruire profili: digitando un nome utente si può verificare quali servizi sono utilizzati da una persona. Generalmente le persone ricorrono a due nomi utente, un alias e il vero nome in una sola parola. In seguito, leakygarden.net controlla le occorrenze di un nome utente.

6. Cosa farebbe la Nielsen?

Gli studi sui dati di ascolto dei vecchi media utilizzano principalmente due metodi di ricerca:

1. il diario → lo spettatore annota quali programmi segue
2. la misurazione → registrata automaticamente di quanto a lungo un nucleo familiare o un membro rimangono sintonizzati.

Si potrebbe trasferire il metodo di conteggio utilizzato nelle ricerche sul pubblico televisivo ai social network, utilizzando i campi di interesse disponibili e i classici dati demografici come età, sesso e posizione geografica.

8. Wikipedia come riferimento culturale

1. Punti di vista nazionali o punti di vista neutrali?

Wikipedia → l'enorme popolarità dell'enciclopedia online viene spesso spiegata anche con la sua capacità di trasformare gli utenti in redattori e collaboratori, diffondendo una cultura che incoraggi un continuo impegno e investimento di sé. A oggi esistono circa 290 edizioni in altrettante lingue o sottodomini di wikipedia.org, ognuna delle quali condivide tre principi fondamentali:

1. punto di vista neutrale → articoli scritti in modo da «(rappresentare) tutti i punti di vista rilevanti pubblicati da fonti affidabili in maniera equa.
2. verificabilità delle informazioni → basati su fonti affidabili esterne da wikipedia
3. assenza di ricerche originali → reperire informazioni già esistenti e riconosciute.

Il fine è ottenere l'accordo di tutti i collaboratori.

Caso studio su uno studio comparativo di una selezione di articoli pubblicati su Wikipedia, che riguardano il massacro in Srebrenica nel 2005. Le lingue degli articoli sono in olandese, bosniaco e serbo.

Il disaccordo sugli articoli controversi è durato almeno cinque anni. Le rispettive angolazioni degli articoli in bosniaco, serbo e olandese sono attribuibili a gruppi specifici di redattori, che contribuiscono alla versione di Wikipedia nella propria lingua, e alle fonti che citano. I redattori delle varie versioni linguistiche collaborano anche a quella in inglese: l'articolo collettivo che ne risulta è oggetto di continua contestazione e spesso viene definito come «occidentale».

I confronti tra le diverse versioni linguistiche di Wikipedia si basano su un tipo di analisi del contenuto che si concentra su determinati elementi alla base di un articolo: il titolo, gli autori, l'indice, alcuni dettagli del contenuto, le immagini e le citazioni. Considerati in aggiunta la posizione geografica dei redattori anonimi, la lettura delle pagine di discussione e segnalazioni di *templates*.

Pozderac è autore degli articoli tradotti dall'inglese in bosniaco, croato e serbo, a ciascuno dei quali ha apportato particolari cambiamenti per quanto riguarda il titolo e alcuni dettagli del contenuto. Dado ha smesso di prendere parte alle discussioni e all'editing a causa delle continue liti scoppiate nelle pagine di discussione tra redattori bosniaci e serbi.

I principi fondamentali di Wikipedia hanno lo scopo di produrre come risultato finale un articolo di buona qualità. Nei periodi di conflitto su articolo e in quelli di relativa calma dopo che le modifiche sono state bloccate, i redattori possono occuparsi del coordinamento di altre attività redazionali. È il lavoro dei redattori che porta l'incremento della qualità nel tempo.

Gli articoli visualizzati avevano più di cinque anni al momento dell'analisi; venivano modificati, quando emergevano nuove prove e quando venivano rilasciate nuove dichiarazioni, anche all'avvicinarsi di ogni anniversario degli avvenimenti del luglio 1995, ma, si differenziano in modo drastico o in dettagli.

2. Studiare la qualità e l'accuratezza di Wikipedia

Se chiunque può scrivere o modificare un articolo, allora chiunque può inserire errori o vandalizzare il contenuto, nonostante la vigilanza dei robot.

Un confronto tra il numero di biografie della Wikipedia in lingua inglese con quelle contenute nella *American National Biography Online* e in *Encarta* ha rivelato che Wikipedia era meno accurata, ma conteneva un maggior numero di biografie nell'ambito storico, Wikipedia è stata confrontata di nuovo con l'*Encyclopaedia Britannica*, risultando ancora una volta meno accurata e in alcuni casi fonte di errori. L'accuratezza di Wikipedia è diseguale, data la vastità di argomenti trattati.

La maggior parte degli articoli su Wikipedia sono incompleti, in alcuni periodi alcuni di essi, i *featured articles*, sono stati valutati degni di pubblicazione su un supporto cartaceo.

Un altro metodo per studiare l'accuratezza di Wikipedia consisteva nell'inserire volontariamente degli errori, valutarne gli effetti e la velocità con cui venivano corretti. In un certo senso, questi test di accuratezza sono anche test di vigilanza dei bot e della capacità di Wikipedia.

La qualità è stata studiata in relazione ai meccanismi burocratici di controllo, al coordinamento dei redattori, all'editing e agli argomenti trattati. Più elevati sono l'interesse e la rilevanza dell'argomento, migliore è la qualità dell'articolo; c'è quindi una relazione tra l'attualità e l'attività redazionale. Si è notato che il numero di modifiche corrisponde a un miglioramento dell'articolo. Un articolo ha più probabilità di raggiungere un buon livello di qualità quando è il risultato dell'attività coordinata di un piccolo gruppo di redattori piuttosto che di un gruppo altrettanto ristretto ma privo di coordinamento, o di un gruppo troppo numeroso.

I redattori principali sono anche i più normativi e giustificano le loro modifiche in discussioni che si attengono ai principi di Wikipedia.

3. Confronti interculturali nello studio di Wikipedia

La ricerca ha rivelato che gli articoli della versione inglese contenevano più informazioni sulle vite personali dei polacchi famosi di quelle contenute negli articoli della versione polacca sugli americani famosi. La versione inglese, che gli studiosi considerano una sorta di versione globale dell'enciclopedia online, «riflette i valori culturali e la storia degli Stati Uniti».

In questo campo d'indagine, gli studiosi sono giunti alla conclusione che gli articoli non dovrebbero essere considerati come «fonti imparziali di informazione». La questione è come le persone scelgono di scrivere o modificare un articolo specialmente nel caso di articoli su argomenti controversi o sensibili.

Per gli argomenti non problematici, l'autoselezione è considerata positiva; per gli argomenti controversi l'autoselezione può avere altri effetti, come le «guerre di editing» e il blocco dell'articolo. In questo caso, molti dei redattori più attivi dell'articolo in inglese sul massacro di Srebrenica sono stati bloccati per non aver aderito alle regole di Wikipedia.

Tenendo conto degli effetti che gli argomenti controversi possono avere sul processo di autorialità collettiva, la ricerca esamina le somiglianze e le differenze nei resoconti di un evento nelle diverse versioni linguistiche. La domanda è se le narrazioni presentate da articoli sullo stesso argomento, si somigliassero o si differenziassero in base alle varie versioni linguistiche.

Attraverso un confronto degli articoli sul massacro di Srebrenica del luglio 1995 in sei versioni linguistiche: olandese, inglese, bosniaca, croata, serba e serbo-croata. La ragione dell'esistenza di tre diverse Wikipedia in bosniaco, croato e serbo è illustrata dall'incidente del blocco e sblocco della versione precedente in serbo-croato.

Il redattore Pokrajak, nel 2005, ha convinto la commissione linguistica di Wikipedia a sbloccare la versione serbo-croata.

4. *Un confronto tra articoli di Wikipedia: caduta, massacro o genocidio di Srebrenica?*

Tra tutti gli articoli contemplati nella ricerca, quello in olandese è stato il primo a essere creato. Questa data precede di poco la pubblicazione del rapporto dell'Istituto olandese per la documentazione bellica. L'articolo in olandese, creato da un ex soldato del Dutchbat, era intitolato semplicemente *Srebrenica*. Nella pagina di discussione, dichiarava di aver creato l'articolo per chiarire come fosse avvenuta la caduta della città.

Poco dopo la denominazione della voce fu modificata in *Dramma di Srebrenica*. Il titolo fu nuovamente trasformato in *Caduta di Srebrenica*, utilizzando lo stesso termine militare che compariva nel rapporto dell'Istituto olandese e diventò quello definitivo. La modifica fu apportata in seguito a una serie di discussioni riguardo alla neutralità della parola *dramma* e al fatto che l'articolo corrispondente nella versione inglese di Wikipedia era intitolato *Massacro di Srebrenica*.

Più recentemente l'utente Bacchus ha riassunto le argomentazioni della scelta in questa frase: «Una buona ragione per cui la parola *caduta* deve essere utilizzata nella versione olandese di Wikipedia è che dal punto di vista olandese la caduta in sé, e il ruolo dell'UNPROFOR, sono molto più interessanti».

La creazione degli articoli sui fatti di Srebrenica nelle Wikipedia in inglese, serbo, croato, bosniaco e serbo-croato avvenne due o tre anni dopo la creazione della voce in olandese.

Dopo la sentenza della Corte internazionale di giustizia nel caso *Bosnia ed Erzegovina contro Serbia e Montenegro*, sentenziò che la Serbia non avesse fatto tutto il possibile per prevenire il genocidio di Srebrenica e inoltre non avesse consegnato al tribunale i sospetti incriminati; tuttavia, il titolo non fu modificato.

Nella versione bosniaca furono aggiunti dettagli che mancavano in quella inglese, tra cui i primi ritrovamenti di fosse comuni, il resoconto del decimo anniversario dell'evento con i nomi degli oratori, il famoso video girato da un membro degli Scorpions che documenta l'esecuzione di alcuni giovani bosniaci musulmani nel luglio del 2005. L'articolo rimodula l'inquadramento generale degli eventi in un attacco da parte dei serbi per eseguire un'operazione di pulizia etnica ai danni dei bosniaci musulmani.

Nella versione croata permane il titolo *Genocidio di Srebrenica*. Con il passare del tempo i wikipediani croati hanno modificato la sintassi e editato l'articolo per renderlo più croato dal punto di vista linguistico.

Contrariamente a quello in croato, l'articolo in serbo non resistette a lungo e «fu subito attaccato come propaganda».

Nella pagina di discussione della versione bosniaca, l'immediato cambiamento del titolo nella versione serba venne commentato dai redattori degli articoli in bosniaco e in serbo-croato. Questo cambiamento era la prova flagrante che i serbi negavano si fosse trattato di un genocidio.

Dalla sua creazione nel settembre del 2005 l'articolo in serbo-croato subì una serie di modifiche, ma è difficile ricostruire esattamente il corso degli eventi. In un primo momento quella in serbo-croato venne reputata una versione unificatrice, secondo una visione derivante inizialmente dall'argomentazione

usata per sbloccare la Wikipedia in serbo-croato. La parola *evento* «è stata scelta proprio per evitare la politicizzazione dell'articolo, dal momento che le diverse parti non hanno raggiunto un accordo su come caratterizzare questi fatti.

5. *L'editing degli articoli su Srebrenica*

Sono stati oggetto di confronto anche i redattori, specialmente i *power editors* o coloro che hanno contribuito maggiormente all'articolo nelle varie versioni linguistiche. Uno degli obiettivi del confronto era quello di prendere nota di tutti i redattori che contribuivano a diverse versioni linguistiche: tendevano a dedicarsi a un'unica versione. La mancanza di editing incrociati è un segnale per verificare la peculiarità di un articolo nelle rispettive versioni linguistiche.

I *power editors* non collaboravano in maniera significativa alle versioni nelle altre lingue balcaniche, mentre quelli delle versioni serba e bosniaca contribuivano alla versione inglese, il cui consenso non venne raggiunto facilmente. Per i redattori delle versioni in serbo, croato, bosniaco e serbo-croato, la versione in inglese costituiva il riferimento principale e in molti casi la base di partenza: essa è servita come articolo comune sull'argomento.

La migrazione dell'articolo da una versione linguistica all'altra ebbe alcuni effetti sul contenuto, e comportò alcune aggiunte ed eliminazioni. La pubblicazione della voce in serbo fu accolta da un'attività frenetica. Nel momento in cui fu pubblicato nella versione serba di Wikipedia infatti, l'articolo non era più accettabile nemmeno per i redattori serbi che avevano contribuito alla versione originale e gli fu apposto subito il *template* di avviso che lo definiva come fonte di un «conflitto di editing».

I contributi degli articoli altrui avvengono quindi per lo più in maniera anonima.

Si riscontrano divergenze tra le varie versioni linguistiche per quanto riguarda il numero di vittime, le responsabilità, le colpe e le controversie intorno a questi e altri punti fondamentali. Nella discussione *Milos* (*power editor*) solleva la questione dell'accuratezza dei dati anche riguardo alla mancanza di informazioni precise sugli abitanti di Srebrenica all'epoca dei fatti, dal momento che non era mai stato condotto un censimento della popolazione né prima della guerra né dopo il suo inizio.

Per quanto concerne l'attribuzione delle responsabilità, un confronto tra le pagine di discussione e gli indici mostra con certezza punti in comune tra gli articoli in bosniaco, croato e inglese, ma anche elementi peculiari negli articoli in olandese e in serbo.

Le versioni in bosniaco e in croato hanno in comune l'accusa esplicita che i serbi abbiano eseguito metodicamente un piano prestabilito, invadendo la città, separando gli uomini dalle donne e dai bambini, evacuando donne e bambini e uccidendo gli uomini.

L'articolo in serbo si concentrava inizialmente sulle operazioni militari. A differenza di quanto accadde per l'articolo in olandese, discussioni molto accese portarono al cambiamento dell'impostazione generale dell'articolo.

Vale la pena soffermarsi su un punto toccato da un utente nel corso della discussione. Gli articoli sui conflitti militari, e il *template* corrispondente, riguardano eventi «strettamente militari» e non «accessori». Anche l'attribuzione delle responsabilità è stata motivo di discussione. L'articolo in serbo evita di utilizzare «forze serbe» ed «esercito serbo di Bosnia», preferendo «esercito della Republika Srpska» o il suo acronimo VRS.

Man mano che l'articolo cresceva, e con esso le pagine di discussione, ogni paragrafo diventava fonte di disaccordo. Fin dall'inizio si osserva spesso che i *power editors* serbi e bosniaci discutono cosa dovrebbe comparire nell'articolo, e che sono i *power editors* occidentali a prendere la decisione finale, assumendo il ruolo di mediatori e *peacekeepers*.

L'articolo in serbo-croato usa la combinazione «Visioni alternative degli eventi, revisionismo e teorie della cospirazione», mentre l'indice dell'articolo in olandese non ha alcun contenuto che faccia riferimento a questo tipo di controversie.

6. Citazioni e immagini negli articoli su Srebrenica

Chiunque può contribuire a un articolo su Wikipedia, sebbene ci siano degli ostacoli da superare. Il fatto che alcune fonti siano considerate accettabili dai redattori di una versione linguistica e non dagli altri ci ha portato a chiederci quali fossero le fonti uniche e quali quelle condivise negli articoli analizzati, e a interrogarci sulla questione più ampia della distribuzione delle attribuzioni o della diffusione delle citazioni attraverso gli articoli. La questione di fondo è quindi se gli articoli siano basati su fonti simili o al contrario molto diverse.

Gli articoli di Wikipedia contengono spesso sia citazioni in forma di note a piè di pagina, sia suggerimenti bibliografici per approfondire l'argomento. Sia le citazioni che i suggerimenti sono link, che offrono la possibilità di ricavare e confrontare i collegamenti attraverso gli articoli, a livello della pagina e a quello dell'host.

In linea di principio non dovrebbe essere strano che le citazioni siano condivise, se si tiene conto del fatto che i redattori importano potenzialmente le proprie citazioni nell'articolo in inglese o le esportano da quella inglese alla propria versione, per così dire. Tuttavia, la scoperta più interessante è che la maggior parte delle citazioni di tutti gli articoli, a eccezione di quello in serbo-croato, sono citazioni uniche. Confrontare le citazioni è un altro modo di guardare oltre i resoconti offerti negli articoli.

L'altro documento importante dal punto di vista cronologico è il rapporto del segretario generale all'Assemblea dell'ONU sulla caduta di Srebrenica, citato negli articoli in olandese e in serbo, ma non in quello in bosniaco.

Alcune citazioni sono condivise dagli articoli in serbo e bosniaco. Dal momento che è la più specifica, la selettività delle citazioni dell'articolo in olandese merita un'ulteriore breve analisi, per chiarire meglio la peculiarità del suo resoconto dei fatti. L'articolo contiene dodici link. Oltre ad articoli su un processo tenutosi recentemente in Olanda, le altre citazioni comprendono due articoli critici. Un'altra citazione fa da contrappeso alle accuse di razzismo e collaborazionismo con l'esercito serbo di Bosnia.

L'analisi delle immagini segue un percorso simile e prende in considerazione il loro numero totale, quante se ne trovino in ogni articolo, quasi siano condivise e quali invece risultino presenti in una sola versione linguistica.

Queste immagini si ritrovano nella maggior parte degli articoli, a volte identiche, a volte simili. L'articolo serbo-croato usa le stesse immagini di quello serbo. L'articolo in bosniaco contiene il più alto numero di immagini non presenti negli altri articoli. Tre delle quindici immagini non ricorrono negli altri articoli. L'analisi delle immagini porta ad affermare che in generale l'articolo in bosniaco contiene un numero

maggior di prove riguardanti gli eventi in sé, mentre gli altri, tendono a focalizzarsi soprattutto sulle indagini.

7. Conclusioni

Gli articoli sarebbero uguali o molto simili se fossero semplicemente tradotti da una lingua all'altra. Possono quindi sussistere delle versioni madri. Qualsiasi articolo su Wikipedia può nascere come traduzione o prodotto di copia-incolla. Qualunque sia la loro origine, questa ricerca studia se essi diventino più universalistici o al contrario più particolaristici, man mano che vengono modificati. L'appello alla specificità culturale può essere interpretato anche come una critica ai valori dal contenuto "americano".

Gli articoli in inglese, bosniaco e croato, seguendo le sentenze del Tribunale penale internazionale per la ex Jugoslavia e della Corte internazionale di giustizia, considerano le uccisioni pianificate di un gruppo di bosniaci musulmani come una parte della conquista della città. La caduta della città costituisce invece l'argomento principale dell'articolo in olandese.

I benefici dell'autoattribuzione degli argomenti da parte dei redattori possono non essere tali nel caso degli articoli controversi, in cui le diverse versioni degli eventi vengono contestate in maniera emotiva, come i *power editors*; i redattori sono attratti da questo tipo di articoli, ma li abbandonano dopo discussioni animate.

Cinque dei principali redattori della versione inglese dell'articolo su Srebrenica sono stati bloccati per un periodo indeterminato o sono stati sospettati di aver utilizzato diversi nomi utente. Dopo che uno o più nomi utente sono stati bloccati, il redattore può rientrare come utente anonimo e verificare se anche il suo indirizzo IP sia stato bloccato. Sarebbe utile che i ricercatori avessero accesso anche agli indirizzi IP dei redattori registrati per poter procedere a un'analisi automatica della posizione geografica di tutti coloro che contribuiscono agli articoli.

È emerso che la maggior parte degli articoli condividono raramente titolo, indice, redattori, citazioni e immagini, e si differenziano anche nei contenuti, a partire dall'argomento dell'articolo. La prospettiva adottata non influenza solo il titolo, ma anche il tipo di riquadro informativo scelto.

Il diverso conteggio delle vittime costituisce un caso speciale nel nostro studio, non soltanto perché mostra le differenze in sé, ma anche perché spesso si basa su fonti non condivise dai diversi articoli. Una delle questioni più delicate è se le vittime fossero o meno in età di combattere, e di conseguenza come interpretarne le uccisioni.

Attraverso le scelte iconografiche, per esempio l'inserimento della fotografia della tomba di un ragazzo tredicenne, l'articolo in bosniaco attira l'attenzione sull'eccessiva giovinezza dei combattenti. Altri articoli condividono immagini di crimini di guerra. Queste fotografie non compaiono negli articoli in olandese e serbo.

I *power editors* delle Wikipedia bosniaca, serba, olandese e serbo-croata modificano continuamente i propri articoli per ottenere testi accettabili, i quali rispettino i tre principi fondamentali dell'enciclopedia e le relative linee guida, che spiegano come raggiungerli.

Sarebbe difficile definire uno degli articoli come universale, mentre esistono cosiddetti «articoli ombrello», alcuni dei quali frutto del lavoro di vari redattori, altri di pochi contributi.

9. Dopo il cyberspazio: “bis data” e “small data”

Ad oggi bisogna abbandonare i concetti di cyberspazio e virtualità come punti di partenza per studiare Internet, o, riposizionarli in modo che riflettano il loro orizzonte concettuale attuale. Il cyberspazio è diventato un campo di studi specifico nell'ambito della sicurezza in Internet: nel 2009, per esempio, l'esercito americano ha creato un *cybercommando*. Analogamente il «virtuale» si riferisce oggi più ai cosiddetti «mondi virtuali» come *Second Life* e agli ambienti di gioco come *World* o *Warcraft*.

Gli studi sul virtuale che si occupano di questi mondi sono diventati un sottogruppo della ricerca su Internet e anche i *cyberspace studies*. Con il declino della navigazione e della teoria letteraria dell'ipertesto come base dello spazio del navigatore, la rete ha perso una parte della produttività ermeneutica che aveva all'inizio.

1. Il web come fonte di dati

I dubbi riguardo a questi dati derivano ancora dall'essere essi storicamente associati all'idea di un cyberspazio libero per tutti e da un'epistemologia basata su un medium di autopubblicazione fai-da-te. Per sostanziare le opinioni che fluttuano in Internet, bisognerebbe disconnettersi dal medium.

Nella nuova storia del web 1.0, seguito dal web 2.0, la rete è concepita come un *continuum* di due versioni stabili di software, a ognuna delle quali corrisponde un particolare dibattito sulla qualità.

Considerare il web come una fonte di dati per la ricerca sociale e culturale significa confrontarsi con una grande varietà di argomentazioni sul disordine di questi dati. Questa situazione fa sì che molti ricercatori rinuncino a utilizzarlo come fonte. Se bisogna ricorrere ai dati online, Thelwall conclude che si debbano confrontare con quelli offline.

2. Fare ordine nel caos online

I ricercatori stanno lavorando alla risoluzione di vari problemi che circondano i dati del web; finora queste difficoltà sono state gestite piuttosto bene da Google e da altre aziende.

Google ha affrontato le critiche che sono state mosse ai dati del web ed è andato ancora oltre, invitando gli studiosi a lavorare con i dati di log del motore di ricerca in modi che si differenzino dalla diffusione dei dati di AOL sia nella forma che nel formato. L'azienda segue lo spirito da piattaforma dei nuovi media.

3. Dati digitalizzati e dati nativi digitali

Nel dibattito sui dati digitali e sulle loro proprietà rispetto a quelle di altri dati, la studiosa Christine Borgman elenca i dati classici, come quelli ricavati dall'osservazione, i dati sperimentali, quelli computazionali e quelli ricavati dai registri, e spiega perché essi siano considerati di buona qualità. In base a questi criteri alcuni dati del web risulterebbero miseramente fallimentari; tuttavia, alcune serie di dati digitalizzati potrebbero superare il test.

Una ricerca culturale basata su dati digitalizzati, *cultural analytics*, propone di considerare «la cultura come un insieme di dati che è possibile estrarre e visualizzare».

Il gruppo di Manovich ha analizzato diacronicamente i cambiamenti delle proprietà delle copertine delle riviste «Time», «Science» e «Popular Science», e dei quadri di Mark Rothko. Alle copertine e ai quadri digitalizzati vengono applicate tecniche informatiche di visione «per generare descrizioni numeriche della loro struttura e del loro contenuto».

La *culturomics* persegue un'analisi quantitativa della cultura, utilizzando come *corpus* iniziale Google books. Le *culturomics* condividono con i metodi digitali il progetto di “fare ricerche su Internet come forma di ricerca sociale”. Questo metodo ha portato ad alcune scoperte interessanti a livello lessicografico, nonché dei trend culturali più vasti.

È il metodo incorporato nei servizi web che vale la pena studiare per la sua capacità di dare senso ai dati: forse malgrado tutto il web è davvero in grado di fornire dati strutturati. In questa prospettiva, i servizi web filtrano, puliscono e ordinano i dati per l'uso finale e magari anche per la ricerca.

Mentre i dati possono essere raccolti senza un apparato metodologico esplicito per l'utente, o senza la presenza di un ricercatore che ascolti casualmente, qualsiasi utilizzo di informazioni desunte dal web deve essere compreso alla luce dello scandalo causato nel 2006 dalla divulgazione di dati da parte di AOL. → Rese disponibili ai ricercatori le *queries* effettuate nell'arco di tre mesi sul suo motore di ricerca da circa 650.000 utenti. Le *queries* erano associate a utenti anonimi identificati da un numero, e inoltre erano disponibili anche altri dati, come gli URL su cui essi avevano cliccato. Osservando l'elenco delle *queries* un giornalista era riuscito a identificare l'utente.

In risposta a particolari «costruttori di *queries*», gli studiosi potrebbero cercare di progettare elementi dell'algoritmo in grado di aiutare il motore di ricerca a fornire agli utenti le informazioni che cercano.

La serie di dati di molti utenti individuali suggerisce l'idea che i log del motore di ricerca forniscano un «database delle intenzioni», rispetto al «corpo formato da dati», generato dalla cronologia delle ricerche effettuate da un utente, gli utenti di AOL sono vittime di atti spregevoli, dal momento che hanno cercato dei rimedi a essi.

La principale lezione tratta dalla diffusione dei dati di log da parte di AOL è la difficoltà di garantire l'anonimato degli utenti dei motori di ricerca. Le aziende che gestiscono motori di ricerca si sono impegnate a proteggere la privacy degli utenti, rendendo anonimi i dati, oltre che attraverso delle direttive di distruzione dei dati stessi, deteriorandoli e accorciandone la vita; tuttavia, l'indirizzo IP è anche un mezzo per geolocalizzare l'utilizzatore, e qui sta la seconda lezione tratta dalla diffusione dei dati di AOL: invece di concentrarsi sull'utente individuale, reso anonimo sostituendo il nome con un numero o rimuovendo una parte o tutto l'indirizzo IP, la ricerca dei log può rivolgersi direttamente ai luoghi da cui sono partite le *queries*.