Le scienze sociali a caccia di dati sul Web:
il progetto HCI

Tommaso Venturini (Sciences Po, médialab)

Il médialab di Sciences Po (médialab.sciences-po.fr) è un laboratorio multimediale digitale, fondato nel maggio 2009 per portare la rivoluzione digitale nel cuore di una delle più antiche e prestigiose comunità delle scienze sociali francesi. Tuttavia, un ostacolo impedisce ancora alla squadra del médialab di sfruttare appieno la più importante fonte di dati dell'epoca digitale: il World Wide Web. Fino a oggi non esiste alcuno strumento che permetta ai ricercatori in scienze sociali di selezionare, estrarre e archiviare un corpus d'informazioni online. L'obiettivo del progetto HCI è sviluppare questo strumento per metterlo al servizio non solo della squadra del médialab e dei ricercatori di Sciences Po, ma anche di tutti i ricercatori interessati a studiare il Web.

Da qualche anno, le scienze sociali si trovano in una situazione completamente nuova. Queste scienze, ancora relativamente giovani, erano lontane dal dotarsi delle enormi macchine per dati a disposizione delle scienze naturali. A differenza dei fisici che si destreggiano con miliardi di particelle dentro i loro acceleratori o dei biologi che coltivano milioni di microbi sotto i loro microscopi, i sociologi potevano seguire solo qualche centinaio di esseri umani ed erano condannati a indovinare la forma dei fenomeni collettivi tramite queste parziali intuizioni. Tuttavia, negli ultimi anni, questa situazione è stata interrotta dall'arrivo dei media digitali.

I media digitali possiedono una caratteristica interessante: tutte le interazioni che li attraversano lasciano *tracce* che possono essere registrate facilmente, conservate e ritrasmesse. Questa caratteristica ha conseguenze capitali per le scienze sociali (Lazer *et al*., 2009). Mano a mano che il digitale s'infiltra nelle società moderne, la vita collettiva diventa sempre più *tracciabile* (Mitchell, 2009). Mano a mano che gli archivi pubblici e privati sono inghiottiti dalla memoria dei computer, che le transazioni economiche migrano online, che le relazioni sociali si radicano nel Web, la quantità di tracce accessibili ai ricercatori cresce esponenzialmente.

Improvvisamente, le scienze sociali si trovano a confrontarsi con altrettanti dati delle scienze naturali, ma con una differenza cruciale: le scienze sociali non hanno fatto nulla per meritarli. Non hanno costruito i loro radiotelescopi, i loro microscopi, i loro sequenziatori. I dati digitali sono stati raccolti per finalità diverse dalla ricerca scientifica. Si tratta d'informazioni raccolte per scopi di marketing (come nel caso delle carte fedeltà), di sorveglianza (come nel caso dei viaggi aerei), di tecniche di manutenzione (come nel caso delle reti di telecomunicazione), di trasparenza (come nel caso di Wikipedia). Si tratta, in altre parole, di *dati d'occasione*, che investono le scienze sociali senza che i ricercatori possano controllare la loro produzione e soprattutto senza che i ricercatori possano prepararvisi. La situazione delle scienze sociali somiglia a quella di certi paesi rurali spinti a una brusca industrializzazione dalla pressione dell'economia internazionale. Nate in un'epoca di penuria, le scienze sociali hanno accesso a un'età d'abbondanza troppo in fretta e senza preparazione.

Questo è particolarmente vero per i ricercatori che si interessano al *World Wide Web* come fonte di dati per le scienze sociali. Se ai suoi esordi il Web poteva essere considerato come un dominio a parte (la cybercultura di Negroponte, 1996; le comunità virtuali di Rheingold, 2000; le identità online di Turkle, 1995), la sua

diffusione capillare trasforma questo medium in una sorta di carta carbone di una parte crescente della vita collettiva (Rogers, 2009). Dentro i *blog*, i *wiki*, i *social media*, i fenomeni collettivi lasciano una quantità crescente di tracce che aspettano solo di essere sfruttate dai ricercatori.

Tuttavia, le promesse dei metodi digitali rischiano d'essere vane senza gli strumenti concettuali e le tecniche necessarie per trasformare le tracce Web in dati di ricerca. La distinzione è cruciale, poiché l'abbondanza delle tracce Web non si traduce automaticamente in un aumento di dati a disposizione degli scienziati sociali. I dati non sono mai semplicemente *dati*, ma sempre costruiti dal lavoro dei ricercatori (Latour, 1993). Per mettere la tracciabilità del Web al servizio della ricerca, non basta raccogliere le informazioni disponibili in Rete: bisogna sapere estrarle, pulirle, indicizzarle, prepararle per l'analisi. In altri termini, bisogna costituirle in un corpus e, ad oggi, questo resta un serio ostacolo per le scienze sociali digitali.

Esistono tuttora due soluzioni per la costituzione di corpora ipertestuali, ma nessuna corrisponde alle condizioni della ricerca accademica. Da un lato, le biblioteche pubbliche riunite nell'International Internet Preservation Consortium (IIPC netpreserve.org) hanno lanciato importanti iniziative di conservazione del Web. Eppure, benché meritorie, queste iniziative riescono a conservare solo una parte infinitesimale del Web. L'Internet Archive (archive.org), oggi la più ampia raccolta pubblica di documenti Web, contiene solo 150 miliardi di pagine, mentre, secondo le stime degli ingegneri di Google, la Rete supera il trilione di pagine (Google, 2008). Il Web è semplicemente troppo grande e troppo dinamico perché si possa sperare di archiviarlo in modo esaustivo. Di conseguenza, tutte le iniziative di archiviazione centralizzata si concentrano su regioni ben delimitate. Le raccolte pubbliche, indispensabili per la conservazione storica, non sono dunque capaci di rispondere alle esigenze della ricerca, poiché i loro campioni corrispondono solo raramente agli interessi dei ricercatori.

Dall'altro lato, le tracce Web sono oggetto di un'attenzione crescente da parte d'aziende di marketing commerciale e politico. Queste aziende dispongono di archivi molti più limitati rispetto alle raccolte pubbliche, ma godono di una flessibilità che permette loro di adattarsi ai bisogni dei loro clienti (si veda, per esempio, l'approccio innovativo di Linkfluence – linkfluence.fr – azienda pioniera dell'archiviazione e dell'analisi del 'Web sociale' e partner di ricerca del médialab). Questa soluzione, benché piuttosto soddisfacente, ha il serio svantaggio di subordinare la ricerca accademica alle strategie private di ricerca e sviluppo.

Per superare l'ostacolo dell'archiviazione Web, il médialab di Sciences Po ha lanciato nel 2010 la Hypertext Corpus Initiative (HCI). L'obiettivo di questa iniziativa è offrire un'alternativa ai grandi progetti d'archiviazione pubblici o privati. HCI cerca di incoraggiare la costituzione di una moltitudine di micro-corpora altamente specializzati, fornendo a ciascun ricercatore la possibilità di raccogliere e conservare i campioni del Web che gli interessano.

Tale strategia si basa sull'osservazione che la Rete, lungi dall'essere uno spazio uniforme, è invece un paesaggio variegato, fatto di discontinuità, di limiti, di lacune. Il Web è un tessuto di collegamenti ipertestuali, ma non un tessuto senza cuciture. I nodi e gli strappi della Rete non sono difetti, ma sono le ragioni del suo successo. Proprio perché il Web non è omogeneo è potuto diventare l'habitat di una molteplicità di gruppi che vi hanno trovato la loro nicchia. Ammettere la natura intrinsecamente discontinua della comunicazione online implica riconoscere che l'esaustività dell'archiviazione è probabilmente meno importante della capacità d'identificare le frontiere che separano le differenti regioni della Rete. Affinché il

Web possa essere messo al servizio delle scienze sociali, bisogna rimpiazzare la strategia di stoccaggio massiva e centralizzata con un approccio che miri a fornire ai ricercatori i mezzi per delimitare i territori che desiderano indagare. La costituzione di un corpus Web non è soltanto una questione di archiviazione, ma anche – e principalmente – una questione di cartografia.

Sotto il nome di 'cartografia del Web' si raccoglie una serie di nozioni, metodi e strumenti che peremttono l'esplorazione dei territori digitali (Boullier, 2009). Alla base di questa cartografia sta l'osservazione che, nonostante il costo necessario a stabilire collegamenti ipertestuali sia minimo, le persone si mostrano piuttosto prudenti a creare connessioni. Ne risulta che il Web non è uno spazio caotico. Scegliendo a quali discorsi legare i propri discorsi online, gli utenti del Web stabiliscono gerarchie e cluster (Gibson, Kleinberg and Raghavan, 1998). Esattamente come Internet, il Web non è una rete regolare, ma una 'rete di reti': un grafico dove zone densamente connesse sono separate da zone caratterizzate da una rarefazione di collegamenti. La cartografia del Web permette di delineare queste zone e di caratterizzare i fenomeni collettivi che esse designano. Sviluppati per rappresentare la topografia dei territori online, i metodi della cartografia del Web offrono una base ideale per la definizione dei limiti di un corpus digitale. L'obiettivo della Hypertexte Corpus Initiative è sviluppare questi metodi e implementarli in uno strumento che permetta ai ricercatori di scienze sociali di costituire, tracciare e archiviare i loro corpora Web.

Sono stati ottenuti molti risultati significativi grazie al médialab di Sciences Po, che rappresenta già un centro d'innovazione riconosciuto nel dominio della cartografia del Web. Ancora prima della fondazione del laboratorio, i membri della squadra del médialab erano già attivi nell'esplorazione dei territori digitali, sia nel quadro del web-mining (webatlas.fr) sia della cartografia delle controversie (medialab.sciences-po.fr/controversies, mappingcontroversies.net). Dalla sua fondazione, il médialab è dunque al centro della ricerca sulla cartografia del Web e partecipa attivamente allo sviluppo dei suoi strumenti (in particolare Navicrawler e Gephi).

Le basi concettuali e tecniche di HCI sono dunque già posate. Su tali basi resta da costruire una metodologia di creazione di corpus Web a un tempo semplice e solida, capace di rivaleggiare con i metodi quantitativi e qualitativi classici. Attorno a questa idea, il médialab ha riunito i principali attori attivi nel dominio dell'archiviazione e dell'analisi del Web. L'Hypertext Corpus Initiative raccoglie molteplici attori accademici (Sciences Po, Institut de Système Complexe di Parigi), istituzionali (BNF, INA) e privati (Linkfluence, Web Atlas, il Consortium Gephi), non soltanto in Francia ma anche all'estero (Digital Méthodes Initiative d'Amsterdam, Density Design Lab di Milano).

Il raggiungimento degli obiettivi del progetto HCI passa attraverso due fasi, una metodologica, l'altra tecnica.

A) Perfezionamento dei metodi della cartografia del Web. La cartografia è una disciplina molto giovane, che ha prodotto risultati molto incoraggianti, ma che merita di essere sviluppata di più. In particolare, il progetto affronterà due punti metodologici.

In primo luogo, la cartografia del Web si è concentrata finora esclusivamente sui collegamenti ipertestuali. Se questo approccio ha il vantaggio di semplificare le attriti di analisi, esso trascura completamente il contenuto delle pagine Web. L'approccio topografico deve essere completato da un approccio lessicometrico

capace di prendere in considerazione almeno il testo delle pagine analizzate (i contenuti multimediali restano al di fuori della portata di questo progetto). Le difficoltà implicate da questa espansione del metodo cartografico sono metodologiche, prima ancora che tecniche: come dar conto del flusso di idee, rumors, opinioni che circolano sul Web (Leskovec, 2009)? Come concettualizzare una citazione non formalizzata con l'aggiunta di un collegamento ipertestuale?

In secondo luogo, i metodi di cui disponiamo oggi sono capaci di fotografare lo stato corrente del Web, ma incapaci sul tema delle dinamiche della comunicazione online. Questo è, più in generale, un problema riconosciuto della *network analysis* ancora relativamente incapace di descrivere la trasformazione delle reti e la circolazione dei flussi di traffico sulle loro connessioni. Qui si tratta di una seria difficoltà che non ha ancora trovato una soluzione standard. Uno degli obiettivi del progetto sarà dunque di testare le differenti soluzioni proposte e identificare quelle che sembrano più efficaci per rendere conto della dinamica del Web.

B) Sviluppo di uno strumento per la costituzione di corpora Web. La ricerca digitale nella scienza sociale è impossibile senza un equipaggiamento all'altezza delle quantità di dati che vorrebbe trattare. Questo è vero in particolare per la cartografia del Web, che implica l'esplorazione, la salvaguardia e l'analisi di una molteplicità di siti. Per quanto sia specializzato, un corpus Web (soprattutto se è archiviato e aggiornato nei tempi) può facilmente prevedere diverse migliaia (anche diverse decine di migliaia) di pagine. La metodologia della costituzione dei corpora Web necessita dunque di essere *implementata* in uno strumento capace di automatizzare le operazioni ripetitive e cronofaghe, lasciando ai ricercatori il massimo controllo sulla definizione dei dati.

Hyphen, lo strumento sviluppato per HCI, mira a dotare i ricercatori di uno dispositivo in grado di:

1. accompagnare i ricercatori nell'esplorazione del Web, tenendo traccia dei siti che visitano (siti-visitati) e dei siti citati da questi siti (siti-vicini);
2. assistere la selezione delle risorse, offrendo ai ricercatori una serie di indicazioni topografiche e lessicografiche sui siti-visitati e sui siti-vicini;
3. facilitare l'estensione del corpus grazie a tecniche d'esplorazione semi-automatiche (crawling);
4. permettere la qualificazione delle ricerche del corpus secondo un codebook individuale o collettivo;
5. estrarre, stoccare, indicizzare, mettere a disposizione i contenuti dei siti del corpus;
6. favorire l'aggiornamento del corpus tramite la possibilità di pianificare sessioni di crawling automatico;
7. consentire l'esportazione delle informazioni raccolte in una pluralità di formati allo scopo di facilitare il trattamento tramite software di analisi statistica, i CAQDAS (Computer Assisted Qualitative Data Analysis Software), i software per la manipolazione delle reti.

Il progetto HCI mira a due risultati principali. Da un lato, HCI intende equipaggiare i ricercatori in scienze sociali che desiderano utilizzare il Web come fonte di dati. Molto concretamente, si tratta di mettere a punto le tecniche di selezione, estrazione e archiviazione delle tracce digitali e di implementarle in uno strumento diffuso con una licenza *open source*. La scelta di rinunciare allo sfruttamento commerciale delle tecniche sviluppate dal progetto e di iscrivere HCI nel quadro giuridico e ideale del software libero non è banale. Questa scelta non mira soltanto a facilitare la diffusione dei risultati del progetto, ma anche ad assicurare la

trasparenza dei metodi elencati nello strumento (chiunque potrà verificare fino all'ultima linea del codice d'Hyphen) e soprattutto a incoraggiare la creazione di una comunità di utenti-partner.

Lo sviluppo di uno strumento per la costituzione di corpus Web è solo l'obiettivo a breve termine del progetto. Le tracce digitali non sono interessanti in sé, ma per il potenziale di rinnovo che offrono alle scienze sociali. Una volta rimosso l'ostacolo alla costituzione dei dati digitali, è questo potenziale che il progetto mira a esplorare. Il secondo obiettivo di HCI è l'istituzione della squadra di médialab come fulcro della ricerca sui metodi digitali nel cuore di Parigi. Questa rete, mobilitata attorno allo sviluppo di uno strumento, tende a stabilizzarsi negli anni a venire. Concepito come un progetto sperimentale, HCI ambisce a rinnovare a lungo termine l'uso dei dati digitali nelle scienze sociali.

Bibliografia

Boullier, D. (2009), "Au-delà des territoires numèriques en dix theses", in F. Rowe (ed.), *Sociétés de la connaissance et prospective. Hommes, organisations et territoires*, Nantes: Lemna.

Gibson, D., Kleinberg, J., & Raghavan, P. (1998), "Inferring Web communities from link topology", in *Proceedings of the ninth ACM conference on Hypertext and hypermedia* HYPERTEXT '98, New York, USA: ACM Press, pp. 225-234.

Google (2008), *We Knew that the Web was Big…*, http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html

Latour, B. (1993), "Le topofil de Boa Vista ou la référence scientifique–montage photo-philosophique", in *Raison Pratique*, Vol. 4, pp. 187–216.

Lazer, D. *et al*. (2009), "Computational Social Science", in *Science*, Vol. 323, pp. 721-723.

Leskovec, J., Backstrom, L., & Kleinberg, J. (2009), "Meme-tracking and the dynamics of the news cycle", in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 497–506.

Mitchell, T. M. (2009), "Mining our reality", in *Science*, Vol. 326, pp. 1644-1645.

Negroponte, N. (1996), *Being digital*, New York: Vintage Books.

Rheingold, H., (2000), *The Virtual Community*, Cambridge, Mass.: MIT Press.

Rogers, R. (2009), *The End of the Virtual: Digital Methods* (p. 36), Amsterdam: Amsterdam University Press.

Turkle, S. (1995), *Life on the screen: Identity in the age of the Internet*, New York: Simon & Schuster.