Maria Carmela Agodi

L'estrazione di dati dalla Rete: una nota introduttiva

1. Nelle società contemporanee, i dispositivi tecnologici che mediano tanta parte del nostro agire e delle nostre interazioni sociali alimentano un flusso continuo di informazioni che descrivono quell'agire dal punto di vista dei sistemi cui quei dispositivi le inviano[1]. Tracce opportunamente codificate di quel che facciamo[2] vengono immagazzinate in qualche *data base* da cui risulteranno potenzialmente estraibili ed interrogabili ogni volta che, ad esempio: acquistiamo un prodotto etichettato con un codice a barre; presentiamo una carta fedeltà in un punto vendita; paghiamo con il bancomat o la carta di credito; consultiamo un sito Internet; inseriamo un post su un blog o un social network; usiamo il cellulare; attiviamo il navigatore satellitare della nostra auto; passiamo davanti a una delle tante videocamere che riprendono quel che accade in certi punti della città.

Gli sviluppi delle tecnologie informatiche e della comunicazione – e la loro incorporazione entro pratiche sociali che ne vengono più o meno profondamente rimodellate – rendono dunque potenzialmente disponibili[3] basi di dati con caratteristiche di novità, sotto almeno tre aspetti: in termini di scala (con scarti di ordine di grandezza tale per cui il mutamento da quantitativo diventa anche qualitativo); di velocità di incremento nel tempo (non si tratta di stock di dati, ma di flussi che si alimentano incessantemente); di integrabilità (è possibile mettere in relazione tra loro, all'interno di un unico *data warehouse*[4], basi di dati generate da processi di diversa natura). Crescono in misura altrettanto significativa anche le

[1] Su quanto quelle descrizioni retroagiscano poi sulla organizzazione dell'agire sociale, si veda G. Bowker, L. Star, *Sorting things out: Classification and Its Consequences*, Boston, MIT Press, 1999.

[2] In termini di "tracce" lasciate dall'agire di attori sociali e dalle attività di sistemi e organizzazioni sociali, piuttosto che di "variabili", si ragiona, in generale, sull'uso dei dati per la ricerca sociale in D. Byrne, *Interpreting Quantitative Data*, London, Sage, 2002.

[3] L'effettiva disponibilità di tali basi di dati è, in generale, subordinata al possesso di specifiche "chiavi di accesso".

[4] Un *data warehouse* è un archivio elettronico che organizza, in relazione a una medesima unità di analisi, informazioni provenienti da diverse basi di dati, relative a fonti e/o occasioni di rilevazione eterogenee.

capacità computazionali dei sistemi di elaborazione che rendono possibile esplorare e analizzare quelle basi di dati.

Dagli esempi sopra citati appare chiaro che per lasciare tracce in questo flusso informativo non occorre essere utilizzatori diretti di tecnologie più o meno avanzate: è sufficiente fare la spesa al supermercato[5]. L'emblema e il più importante supporto tecnologico di questa nuova tracciabilità dei comportamenti individuali e della comunicazione interpersonale è tuttavia ben identificato: si tratta del Web e dei suoi sviluppi. Com'è noto, i mutamenti tecnologici che lo riguardano hanno progressivamente investito: la capacità d'immagazzinamento e di reperimento dell'informazione (WEB 1.0)[6]; l'immediatezza della connessione spazio-temporale e quindi della comunicazione di qualunque tipo di informazione tra i singoli punti della rete (WEB 2.0)[7]; la trasmissibilità e l'interconnettività attraverso dispositivi e sensori, distribuiti nello spazio e/o mobili (WEB 3.0 o WEB degli oggetti)[8].

L'opportunità aperta dalle nuove tecnologie di raccogliere, conservare ed elaborare dati su larga scala ha indotto trasformazioni significative nelle scienze della natura, come la fisica o la biologia, ma anche in un

[5] Il punto che merita di essere sottolineato sociologicamente (ricordandosi del famoso taglialegna di Weber) è che l'agire che l'attore sociale definisce come "andare a far la spesa al supermercato", dal punto di vista dei sistemi che attraverso i diversi dispositivi tecnologici acquisiscono le informazioni per essi rilevanti, si ridefinisce come l'insieme di "azioni" sopra riportato: acquistare prodotti con un dato codice a barre; presentare la carta fedeltà; pagare con il bancomat o la carta di credito; consultare un sito Internet (dal punto di vista dell'attore: controllare il saldo dopo il pagamento, via *smartphone*?); inserire un post su *twitter* (per l'attore: fare un po' d'ironia sul caro-vita?); usare ancora il cellulare (avvertire casa che si sta per rientrare?); attivare il navigatore satellitare dell'auto (scegliere il tragitto meno trafficato?); passare davanti a una delle tante videocamere che riprendono quel che accade in certi punti della città (ritrovarsi una bella multa perché passava davanti al semaforo mentre stava scattando il rosso!).

[6] I sistemi di archiviazione informatica rendono accessibili ed esplorabili i testi di volumi e riviste che nessuna biblioteca, per quanto capiente, potrebbe fisicamente contenere. L'informatizzazione di tanta parte degli archivi e delle pratiche di imprese private e di organizzazioni modifica, almeno potenzialmente, l'esplorabilità di informazioni relative a processi decisionali di rilevanza pubblica per raccogliere le quali sarebbe stata un tempo necessaria una lunga consultazione di pratiche e documenti cartacei.

[7] Nel momento in cui i dati sono caricati direttamente dai singoli individui l'operazione del caricamento non è più centralizzata e affidata a operatori *ad hoc*, ma è distribuita tra tutti coloro che hanno accesso alla rete. Il principale vincolo al caricamento dei dati risulta così superato e la portata del flusso dei dati immessi può crescere praticamente senza limiti.

[8] Il web degli oggetti consente, ad esempio, alle telecamere mobili munite di GPS (Geographical Positioning System), di inviare in tempo reale le immagini che registrano per strada e renderle visibili, in modalità *street view*, localizzandole su *Google maps*. Si veda la veloce descrizione che ne fa l'inventore stesso del web in T. Berners-Lee, *The Web of Things* (*European Research Consortium for informatics and Mathematics*) ERCIM News, http://ercim-news/ercim.eu/the-web-of-things.

ambito particolare dell'economia, come quello della finanza[9]. Si tratta di sviluppi dell'indagine resi possibili da una nuova strumentazione tecnica, come spesso accade nelle scienze (almeno dai tempi del cannocchiale di Galileo). Anche in questo caso, la strumentazione rende "osservabile" un livello di realtà che non era accessibile con gli strumenti prima disponibili: ne sono esempi i flussi di dati che arrivano alla NASA dalle rilevazioni satellitari della superficie terrestre e dell'atmosfera[10]; la decifrazione del genoma umano, che si è avvalsa delle tecnologie informatiche procedendo a una velocità superiore a quella prevista dagli stessi ricercatori; i dati sul funzionamento dei mercati finanziari, dai quali sono rette le sorti dell'economia mondiale. Nessuno di tali sviluppi è tuttavia meramente riducibile all'applicazione di quelle tecnologie: è il risultato di scelte e decisioni cognitive di volta in volta negoziate all'interno dello specifico campo d'indagine e che richiedono a loro volta un riadattamento degli apparati cognitivi elaborati dalle stesse scienze[11]. Le tipologie di dati cui effettivamente accedere ed i sistemi di rilevanza secondo cui orientarne l'esplorazione sistematica sono solo co-determinati dai dispositivi tecnologici: si definiscono entro i rispettivi dominî d'indagine delle diverse scienze o, per meglio dirla con gli studiosi di tecno-scienza, entro le rispettive arene trans-epistemiche di azione[12].

Nell'ambito proprio delle scienze sociali e culturali, l'interesse alle grandi masse di dati provenienti dall'inserimento delle tecnologie informatiche e della comunicazione entro le pratiche sociali appare molto diversificato, concentrandosi soprattutto in alcune nicchie specifiche,

[9] Su quanto gli strumenti avanzati della finanza siano dipendenti dal funzionamento dei software e dei dispositivi di calcolo più sofisticati, sono estremamente istruttivi i lavori di D. MacKenzie, a partire da *Physics and Finance: S-Terms and Modern Finance as a Topic for Science Studies*, «Science, Technology and Human Values», XXVI, 2, 2001, 115-44. Si veda anche D. MacKenzie, F. Munesia, L. Siu (a cura di), *Do Economists Make Markets?,* Princeton University Press, 2007.

[10] La trasmissione in rete, in tempo reale, delle immagini provenienti dai satelliti ha trasformato il sistema in un vero e proprio osservatorio distribuito sul web (Sloan Digital Sky Survey/Skyserver), in cui ciascuno può entrare, visualizzare, per osservarla sistematicamente, una porzione di spazio e quindi segnalare eventuali osservazioni "inattese" o comunque interessanti. Si veda all'URL: http://cas.sdss.org/dr7/en/.

[11] La percezione della complessità, oltre che della bidirezionalità, dello stretto legame tra scienza e tecnologia è uno dei motivi per cui negli *Science and Technology Studies* è entrato in uso il termine "tecno-scienza".

[12] Per arena trans-epistemica d'azione s'intende un ambito di ricerca che si estende oltre quello della comunità disciplinare sino a includere un'ampia varietà di *stakeholders* ed entro il quale si realizzano transazioni basate su risorse di vario tipo (da quelle strettamente cognitive a quelle finanziarie e di legittimazione) che a loro volta orientano le scelte cognitive che vi si realizzano, le direzioni d'indagine che la ricerca esplora e quelle che tralascia. Il concetto è introdotto per la prima volta in K. Knorr Cetina, *Scientific Communities or Transepistemic Arenas of Research? A Critique of Quasi-Economic Models of Science*, «Social Studies of Science», XII, 1982, 101-30.

prima tra tutte quella degli studiosi della comunicazione. In un articolo collettivo apparso nel 2009 su «Science»[13], una quindicina di studiosi richiamava l'attenzione delle scienze sociali sulla enorme mole di dati e sui flussi, ininterrottamente generati, di informazioni di cui potrebbero fare tesoro l'economia, la sociologia, la scienza politica. Nel frattempo, a tali fonti – si faceva notare – avevano da tempo iniziato ad attingere le agenzie di marketing, i colossi dell'informatica come Google e Yahoo ed i servizi di *intelligence*. Gli autori auspicavano dunque, anche per le scienze sociali, lo sviluppo in ambito accademico – in un contesto, cioè, aperto alla pubblica discussione e alla disamina critica – di programmi di ricerca basati sulla esplorazione sistematica di questa massiccia produzione di micro-dati riferibili ai più diversi ambiti del sociale ed orientati all'incremento della conoscenza dell'agire individuale e collettivo[14].

Almeno tre sono gli ordini di motivi per condividere questo auspicio. Vi sono, innanzitutto, domande di conoscenza cui un'opportuna interrogazione sistematica di quei dati potrebbe fornire risposte rilevanti per gli scienziati sociali e la società nel suo insieme, ma che potrebbero non avere significato alcuno dal punto di vista delle agenzie di marketing, dei gestori dei servizi informatici, delle compagnie telefoniche o delle diverse agenzie di *intelligence* degli stati nazionali. In secondo luogo, è più che opportuno che la conoscenza prodotta a partire da queste fonti sia soggetta a uno scrutinio il più aperto possibile[15], controllata e validata intersogget-

[13] D. Lazer *et al.*, *Computational Social Science*, «Science», CCCXXIII, 5915, 2009, 721-3. L'articolo è accompagnato da una bibliografia, dai curricula dei firmatari e da riferimenti a risorse *on line* sull'argomento.

[14] Gli autori invocano, in effetti, lo sviluppo di ciò cui si riferiscono come "scienza sociale computazionale". Con questa locuzione si intende un ambito di ricerca interdisciplinare relativamente recente, in cui sociologi, informatici, studiosi di scienze cognitive si dedicano alla ricerca sui fenomeni sociali con una specifica attenzione alla elaborazione dell'informazione e con gli strumenti del calcolo avanzato. Le principali aree di competenza sono i sistemi di estrazione automatica dell'informazione, l'analisi dei reticoli sociali, i sistemi d'informazione geografica, i modelli di sistemi complessi e i modelli di simulazione. Per una rassegna recente sulle aree comprese in questo ambito di ricerca, si veda C. Cioffi-Revilla, *Computational Social Science*, «Wiley International Reviews: Computational Statistics», 2, 3, 2010, 259-71. Nel 2009 è nata, dalla preesistente *North American Association for Computational, Social and Organizational Science*, la *Computational Social Science Society*, con sede a Washington. Nel Novembre 2010, questa ha tenuto il suo primo Convegno. Una raccolta in quattro volumi delle ricerche più significative in questo campo di studi è: N. Gilbert (a cura di), *Computational Social Science*, London, Sage, 2010. Particolarmente attivo è, anche dal punto di vista della regolazione del settore e dei suoi standard di qualità, lo *Special Interest Group on Knowledge Discovery and Data Mining* della *Association for Computing Machinery* (http://www.sigkdd.org/).

[15] Tale scrutino dovrà essere tanto più aperto in quanto non potrà che essere comunque "esperto", cioè attrezzato dal punto di vista concettuale e tecnico-strumentale. La conoscenza prodotta dai "centri di calcolo" non può essere confutata che da altri "centri di calcolo", come si rilevava già in B. Latour, *Science in Action*, Boston, Harvard University Press, 1987 (cfr. trad. it., Milano, Edizioni di Comunità, 1998, 120-24).

tivamente dal punto di vista metodologico e alla luce del patrimonio di conoscenze e teorie sin qui prodotte dalle scienze sociali. In terzo luogo, le preoccupazioni sul rispetto della privacy[16], che un accesso non controllato a quei dati può generare, potrebbero trovare risposte più adeguate ove a quelle fonti fosse interessata anche la comunità accademica[17].

L'analisi secondaria di dati non generati all'interno del processo di ricerca è stata da sempre una componente essenziale dell'indagine sociologica, a partire dalle ricerche di Durkheim sul suicidio, basate su dati istituzionali. Le analisi dei dati di fonte istituzionale[18] hanno prodotto significativi contributi di conoscenza e utili spunti di riflessione per i responsabili delle politiche pubbliche. Se gli scienziati sociali non dedicheranno la opportuna attenzione a queste nuove fonti per l'analisi secondaria[19], la domanda di esplorazione del potenziale di conoscenza che da essi potrebbe ricavarsi rimarrà ancorata alla logica sistemica delle organizzazioni che li generano e/o che sono in grado di appropriarsene.

2. L'estrazione automatica delle informazioni ha in effetti dato luogo a un campo di ricerca estremamente vasto e specializzato, il cui processo

[16] Queste preoccupazioni sono generate non solo dai timori legati all'uso da parte dei privati (aziende, gruppi d'interesse politici o economici, ecc.), ma anche da organismi statali, organi di polizia o agenzie di *intelligence*: si veda, in proposito, S. Harris, *The Watchers. The Rise of America's Surveillance State*, New York, Penguin, 2010; da un punto di vista tecnico, M. Gertz, S. Jajodia (a cura di), *Handbook of Database Security. Applications and Trends*, New York, Springer, 2008.

[17] Gli appartenenti a una comunità scientifica hanno infatti non solo *l'expertise* ma anche la motivazione per suggerire regole di circolazione e utilizzo di quelle fonti che consentano di ottenere conoscenze significative dal punto di vista generale – perché interessati esclusivamente a queste – garantendo, al tempo stesso, la massima tutela per la *privacy* individuale.

[18] L'esigenza di una profonda revisione del concetto di "istituzionalità" delle fonti sarebbe, alla luce delle trasformazioni in atto, quanto mai urgente. Significativa di quanto queste trasformazioni stiano incidendo anche sulle fonti ufficiali è la decisione, da parte dell'ISTAT, di modificare radicalmente le modalità di rilevazione per il Censimento 2011, riducendo la portata dell'indagine su tutta la popolazione (con la differenza prevista tra *long* e *short form*). Si veda, sul nuovo Censimento, http://www.istat.it/censimenti/popolazione2011. Sulla tematica delle fonti, tre recenti convegni segnalano che la percezione della rilevanza del tema va comunque facendosi strada tra gli scienziati sociali italiani: *Qualità del dato e rispetto della persona nella ricerca sociale e di marketing*, Milano, 2008; *Interrogare le fonti: un confronto interdisciplinare tra domande conoscitive e basi di dati*, Napoli, 2009; *2060: Con quali fonti si farà la storia del nostro presente? Tecniche, pratiche e scienze sociali a confronto*, Milano, 2010 (documentazione sul sito http://www.ais-sociologia.it/)

[19] Una prima attenzione da parte della sociologia italiana verso tali fonti è comunque già documentata, a partire dal decennio scorso, in un volume a cura dell'AIS: si veda il contributo di G. Delli Zotti, *Le nuove fonti di dati per la ricerca sociale: opportunità e limiti*, in G. Amendola (a cura di), *Anni in Salita. Speranze e paure degli italiani*, Milano, Franco Angeli, 2002, 46-51.

di formazione risale almeno alla metà degli anni Ottanta[20]. L'esplorazione sistematica di queste grandi basi di dati – vere e proprie miniere di informazioni – va, per l'appunto, sotto il nome di *data and text mining*[21]. Essa si inserisce all'interno di un processo più complesso di recupero, selezione, preparazione dei dati e, infine, di interpretazione dei risultati, con cui spesso viene identificata, ma che può essere meglio designato, con una espressione più generale, come *Knowledge discovery from databases* (KDD). A monte della fase di *data mining* vero e proprio, infatti, i dati vanno recuperati da una fonte primaria (un *data warehouse* o altra fonte); va selezionato il sotto-insieme di informazioni che interessa ai fini dell'esplorazione; i dati così selezionati vanno quindi depurati dagli errori riconoscibili. A questo punto si procede al *data mining* vero e proprio: alla ricerca, cioè, in essi, di regolarità e strutture potenzialmente significative, da cui vanno poi selezionate quelle ritenute più rilevanti e conoscitivamente interessanti. Anche se il *data mining* è, nel suo insieme, un campo di specializzazione piuttosto recente[22], le tecniche di cui si avvale sono molteplici e non tutte nuove: hanno le loro radici nell'informatica, in particolare nel riconoscimento di strutture, e nella statistica[23]. La scelta tra di esse dipende sia dal tipo di dati analizzati che dagli obiettivi dell'analisi esplorativa. Tra le più utilizzate, vi sono le tecniche di raggruppamento (*clustering*), le reti neurali (supervisionate e non), gli alberi di decisione, le tecniche per l'individuazione di associazioni, la *network analysis*. Ciascuna di queste famiglie di tecniche è ampiamente nota agli scienziati sociali:

[20] Nel 1996 esce, per la MIT Press, un volume collettaneo, a cura di M. Fayyad *et al.*, dal titolo *Advances in Data Mining and Knowledge Discovery*, che si propone come una rassegna degli strumenti e delle tecniche più avanzate, risultato degli sviluppi di quell'ambito di studi nel decennio precedente. Già nel 1997 viene edita da Springer la prima rivista specializzata, dal titolo «Data Mining and Knowledge Discovery». Seguiranno «Data Mining Research», «International Journal on data Mining», «International Journal of Data Warehousing and Mining». Un manuale italiano recente è S. Dulli, S. Furini, E. Peron, *Data Mining. Metodi e Strategie*, New York, Springer, 2009.

[21] L'esplorazione di corpi testuali non strutturati può essere considerata un ambito di ricerca a sé, anche se strettamente collegato al *data mining* (cfr. S. Bolasco, *Statistica testuale e text mining: alcuni paradigmi applicativi*, «Quaderni di Statistica», 2005, 17-53). Anche su di esso, la letteratura è ormai vastissima. Si veda, da ultimo, un'utile e aggiornata rassegna in A. Brien, B. Hopp, *Computer Assisted Text Analysis in the Social Sciences*, «Quality & Quantity», 45, 2011, 103-28. Un manuale italiano relativamente recente è S. Dulli, P. Polpettini, M. Trotta (a cura di), *Text Mining. Teoria e applicazioni*, Franco Angeli, Milano, 2004. Si veda anche S. Bolasco, I. Chiari, L. Giuliano, *Statistical analysis of textual data*, Milano, Led, 2010.

[22] Per quanto si tratti di sviluppi relativamente recenti, le loro origini risalgono comunque già agli anni Ottanta. Una sintetica introduzione al *data mining* si può ormai trovare anche in un manuale italiano di metodologia della ricerca sociale: L. Bocci, *Il data mining*, in L. Cannavò, L. Frudà (a cura di), *Ricerca sociale. Dall'analisi esplorativa al data mining*, Roma, Carocci, 2007.

[23] Si veda in proposito D. J. Hand, *Data Mining. New Challenges for Statisticians*, «Social Science Computer Review», , XVIII, 4, 2000, 442-9.

quel che è meno familiare è il loro uso combinato nell'ambito delle procedure di *data mining*.

Le applicazioni, nel dominio dei fenomeni sociali, sono spesso molto specifiche e orientate al perseguimento di obiettivi pragmatici. Le principali sono orientate all'analisi dei consumi e della competizione sui mercati: segmentazione della clientela di un bene o servizio, al fine di individuarvi gruppi omogenei nei comportamenti di acquisto; analisi delle associazioni, applicata generalmente ai dati di vendita, per scoprire quali prodotti sono più spesso acquistati congiuntamente; raggruppamento di testi, per argomento trattato, e di argomenti, in quanto più spesso trattati insieme; monitoraggio dell'innovazione tecnologica, per individuare (attraverso le registrazioni dei brevetti) quali sono le tecnologie che crescono più velocemente, le persone e le organizzazioni coinvolte nelle invenzioni e nella loro acquisizione proprietaria. Insieme a queste si sono sviluppate anche applicazioni aventi per oggetto fenomeni di tipo organizzativo; da tale campo di studi, ha poi preso l'avvio tutto l'ambito applicativo degli strumenti della *network analysis* dinamica, che è stato utilizzato per lo studio di fenomeni tra loro eterogenei come il terrorismo, la diffusione sociale delle epidemie, la dinamica delle reti di collaborazione scientifica, dei cluster disciplinari, delle controversie scientifiche, delle reti amicali, ecc. Più di recente, la gamma dei fenomeni indagati si è ulteriormente allargata, comprendendo l'identificazione di profili devianti, l'ambito dell'istruzione[24], la diffusione dei farmaci, delle tecnologie informatiche e della comunicazione, ma concentrandosi, per lo più, in centri di ricerca specializzati in *computational social science*, che si propongono come centri di *expertis*e anche per il sostegno a politiche *evidence-based*[25].

3. L'interrogativo sulle potenzialità euristiche associate all'esplorazione sistematica delle basi di dati rese accessibili dalle nuove tecnologie va, a nostra avviso, orientato in una duplice direzione: quella di esplorare le possibilità di accesso della ricerca sociologica a queste nuove fonti, da una parte; dall'altra, quella di mettere a frutto le risorse metodologiche e teoriche elaborate nell'ambito delle scienze sociali, e della sociologia in particolare, per ricavare da tali fonti, perlopiù generate in relazione a

[24] Il campo degli *Educational Studies* ha anche una specifica nicchia dedicata a ricerche su questo tipo di fonti: esiste un *International Working Group on Educational Data Mining* che ha, dal 2009, anche una propria rivista on line, «The Journal of Educational Data Mining» (http://www.educationaldatamining.org/JEDM/).

[25] Tra gli studiosi che, in ambito internazionale, hanno contribuito a dare impulso a questo tipo di studi e Centri di ricerca, Kathleen Carley (Carnegie Mellon University, http://www.casos.cs.cmu.edu/bios/carley/carley.html), David Lazer (Harvard University, http://www.hks.harvard.edu/davidlazer/html/), Claudio Cioffi-Revilla (George Mason University, http://socialcomplexity.gmu.edu/director.php) e V. S. Subrahmanian (University of Maryland, http://www.cs.umd.edu/~vs/).

obiettivi pragmaticamente circoscritti, conoscenze di tipo propriamente sociologico e risposte alle domande di riflessività che provengono dalle collettività sociali.

La prima direzione di esplorazione richiede che la comunità scientifica, oltre a contribuire a diffondere la consapevolezza sulla mole di dati già disponibili soprattutto sul *web*, produca una pressione verso l'ampliamento delle condizioni di accesso a fonti attualmente spesso di tipo proprietario – in quanto gestite direttamente dalle aziende che le ricavano dal controllo su dispositivi tecnologici che mediano le transazioni con la loro clientela. Spesso tali fonti non sono altrimenti accessibili, per gli scienziati sociali, che attraverso la collaborazione alla ricerca orientata al mercato.

Nel già ricordato articolo di Lazer e colleghi su «Science», tra i rischi intravisti nello sviluppo di una scienza sociale computazionale, si prefigurava quello che essa potesse diventare dominio esclusivo di compagnie private e di agenzie governative; o, in alternativa, una riserva per èlite di ricercatori che, grazie all'accesso a dati secretati, producessero ricerche i cui risultati non potessero essere né controllati né replicati. La pubblicità delle fonti, delle tecniche e degli strumenti di analisi sembra irrinunciabile per uno sviluppo scientificamente fecondo e, insieme, socialmente responsabile di questo campo di ricerche.

Non si intende sottovalutare la ricchezza conoscitiva che può scaturire, da collaborazioni che presuppongono una committenza esterna, per la conoscenza sociologica. E, reciprocamente, la collaborazione con quella accademica può essere di stimolo, per la ricerca orientata al mercato, alla esplorazione di nuovi percorsi cognitivi, ulteriori e potenzialmente innovativi rispetto a quelli trainati dalla domanda della committenza. Non si tratta certo di una novità per la sociologia. È questa la lezione che, per citare un caso paradigmatico, deriva dai risultati delle ricerche di Lazarsfeld sulla mediazione esercitata dall'interazione sociale e dalla comunicazione interpersonale, sull'efficacia della comunicazione veicolata da quelli che allora erano i nuovi *media* (la radio, in particolare)[26]. Quelle ricerche ci ricordano quanto domande sociologicamente intelligenti – sia pur collegate a interessi conoscitivi di rilevanza immediatamente pragmatica – siano suscettive di dar luogo a conoscenze sociologicamente significative e, opportunamente contestualizzate, capaci di generare nuovi interrogativi euristicamente utili. Le nuove tecnologie dell'informazione e della comunicazione offrono alle aziende l'opportunità di ottenere in tempo reale informazioni dettagliate su comportamenti espliciti, consumi

[26] Ci si riferisce ovviamente alle ricerche sugli "effetti" della propaganda politica e della pubblicità, nell'orientare, rispettivamente, le scelte di voto e di consumo: P. Lazarsfeld *et al.*, *The People's Choice*, New York, Columbia University Press, 1948 e E. Katz, P. F. Lazarsfeld, *Personal Influence*, Glencoe, The Free Press, 1955.

pregressi, trend sociali e modelli socioculturali di riferimento dei loro clienti. Su questi flussi di informazione sono stati sperimentati, a partire dalla seconda metà degli anni Novanta, strumenti di analisi specifici, orientati soprattutto al monitoraggio dei dati che passano attraverso la rete (*web-analytics*) e che consentono di studiare con continuità i "profili" di consumo dei clienti. Il percorso che conduce all'acquisto sembrava, con l'avvento della società di massa, riconducibile a un processo lineare nel quale la comunicazione di mercato, saltate le mediazioni interpersonali, assorbiva al proprio interno il ruolo degli *opinion leaders*, affidandolo a figure mediatiche. Le osservazioni compiute sulla rete mostrano la rilevanza della mediazione delle relazioni interpersonali – sia pure quelle che si costituiscono all'interno del web 2.0, attraverso i blog ed i social networks – ed al consumatore viene riconosciuto nuovamente a pieno titolo lo status di attore sociale, come ben evidenzia il contributo di Molteni in questa sezione. La rete assume la funzione di luogo di raccolta e redistribuzione (*Hub*) di informazioni, conoscenze e relazioni in cui gli attori sociali (interni ed esterni alle aziende produttrici di beni e servizi) sono immersi e che trovano nell'agire di consumo solo una delle tante modalità espressive. Per gli esperti di marketing, l'innovazione che ne consegue consiste nella possibilità di monitorare le conversazioni online ed eventualmente introdursi in esse – direttamente attraverso il marchio rappresentato o indirettamente attraverso i cosiddetti Heroes[27] – ma anche di identificare bloggers influenti (i nuovi opinion leaders) da invogliare a divenire volontari testimonials del marchio[28]. La propaganda politica – il caso della campagna presidenziale di Obama ne è stato la manifestazione più eclatante[29] – ha recepito queste stesse strategie di costruzione di relazioni con i propri interlocutori, gli elettori, e la scienza politica si è adeguata, ricorrendo agli strumenti di analisi della comunicazione mediata dal web[30].

[27] Acronimo che sta per *Highly Empowered and Resourceful Operatives*, dipendenti delle aziende che si rivelano capaci di usare al meglio le tecnologie *on line* per entrare e rimanere in contatto con i loro clienti 2.0 e che, dalle stesse aziende, "aspettano solo di essere scoperti e valorizzati": cfr. A. Jaconi, *I nuovi eroi d'azienda*, «Il Sole 24 ore», 26 gennaio 2011. Questo concetto è introdotto in J. Bernoff e T. Schadler, *Empowered*, Boston, Harvard Business School Press, 2010.

[28] Le aziende europee produttrici di beni e servizi di lusso – su suggerimento degli esperti di marketing – stanno ormai puntando su strategie di questo tipo anche per la conquista del pubblico di quello che oggi si presenta come il più promettente mercato mondiale, quello cinese. Cfr. Y. Atsmon, V. Dixit, C. Wu, *Tapping China's Luxury-Good Market*, «McKinsey Quarterly», aprile, 2011 (2779).

[29] K. Wallsten, "*Yes We Can": How Online Viewership, Blog Discussion, Campaign Statesment and Mainstream Media Coverage Produced a Viral Video Phenomenon*, «The Journal of Information, Technology and Politics», VII, 2-3, 2010, 163-181.

[30] Nel 2004 nasce «The Journal of Information, Technology and Politics», che diventerà rivista ufficiale della Sezione *Information, Technology and Politics* dell'*American Po-*

Per il ricercatore sociale, la novità è costituita, oltre che dalla vastità delle basi di dati potenzialmente disponibile, dal fatto che i comportamenti degli attori *on line* diventano, attraverso questo particolare tipo di fonti, direttamente *osservabili*, piuttosto che accessibili solo attraverso i resoconti raccolti, ad esempio, in risposta a un questionario (come accadeva dai tempi di Lazarsfeld sino a l'altro ieri)[31]. Ne è un'esemplificazione il saggio di Fragapane, Giuffrida e Zarba, in questa sezione monografica, sui dati di accesso a un quotidiano *on line*. I risultati che si ricavano dai dati disponibili sono descrittivi di diverse *pratiche* di accesso, lettura e consultazione del sito del quotidiano che aprono interessanti prospettive di ricerca e approfondimento sul significato della differenziazione delle attività che si organizzano intorno alla relazione tra utenti e testata giornalistica, con l'incorporazione, nella pratica della lettura del quotidiano, della realtà della sua versione *on line*. Da tale incorporazione, quella pratica risulta in parte modificata, in parte appunto "differenziata", dando luogo a un ventaglio di attività (lettura del quotidiano cartaceo; consultazione *on line* della edizione del giorno; ricerche nell'archivio elettronico; accesso a servizi dedicati) che si diversificano pur essendo organizzate tutte intorno a una medesima riconoscibile entità: la testata giornalistica di riferimento[32].

Ma la comprensibilità del significato sociologico di quella differenziazione è soggetta a una condizione: che le informazioni che, analiticamente, "dissezionano"[33] l'agire codificandone le "tracce" sui diversi dispositivi di rilevazione dei dati, secondo criteri di classificazione e standard che traggono significato dai sistemi cui quei dispositivi li inviano (in questo caso, quelli che regolano la gestione delle testate giornalistiche che raccolgono quei flussi di dati), possano essere ricondotte a contesti concreti d'azione

litical Science Association. Nel maggio 2011 la rivista terrà a Washington la terza edizione del proprio Convegno annuale dal titolo "The Future of Computational Social Science".

[31] Ancora sino a qualche anno fa le indagini sull'accesso dei lettori ai quotidiani *on line* erano basate su dati raccolti tramite questionario: si veda M. B. Salwen, B. Garrison, P. D. Driscol, *On line news and the public*, New York, Routledge, 2004.

[32] Un punto di vista interno alla produzione delle testate giornalistiche è assunto, invece, in P. J. Boczkowski, *Digitizing the news: innovation in on line newpapers*, Boston, The MIT Press, 2004: un'analisi, localmente situata, del lento processo di appropriazione delle tecnologie informatiche entro le pratiche redazionali, sino alla pubblicazione, intorno alla metà degli anni Novanta, dei primi quotidiani *on line*. Un'indagine, condotta con metodo etnografico, sulle pratiche di produzione dei quotidiani *on line* si trova in C. Paterson, *Making on line news. The Ethnography of New Media Production*, New York, Peter Lang Publishing, 2008.

[33] Il termine fa riferimento al titolo del volume di P. Hedstrom, *Dissecting the Social. On the Principles of Analytical Sociology*, Cambridge, CUP, 2005, tradotto in italiano con il titolo *Anatomia del Sociale* (Milano, Bruno Mondadori, 2006), che della sociologia computazionale recupera la dimensione generativa della spiegazione, non come connessione causale tra eventi ma come esito congiunto, generato dalla interconnessione delle azioni di una pluralità di agenti.

ulteriori dai quali quell'agire deriva il suo senso e ai quali contribuisce a conferirne. La complessità di questo processo di "ri-assemblaggio"[34] delle informazioni, condizione per la ricomposizione della dimensione socio-culturale ed autoregolativa dell'agire sociale, è esemplificata dal resoconto della fase di "costruzione" dei dati, nel secondo contributo di questa sezione monografica, di De Felice, Giuffrida, Giura e Zarba. Avvalendosi della conoscenza diretta dello spessore semantico della fonte, costituita dalle sentenze penali su reati di mafia, tale fase passa per una preliminare codifica ed esplorazione statistica di un campione di sentenze, su cui prima progettare e quindi controllare gli algoritmi di estrazione automatica del materiale testuale, da "addestrare", poi, secondo i canoni del *machine learning*. Il confronto con le classificazioni ottenute utilizzando come fonte le statistiche giudiziarie mostra cosa s'intenda quando si propone l'obiettivo di andare oltre le finalità proprie dei sistemi di produzione dei dati: emerge la differenza tra conoscenza finalizzata al monitoraggio della "macchina" giudiziaria e conoscenza finalizzata, come si propongono gli autori, alla ricostruzione dei rapporti di forze e delle gerarchie di valore fatte valere socialmente, nella risposta istituzionale ai reati di mafia.

È nella capacità di recuperare un significato sociologico ai dati – attraverso il riferimento ai meccanismi che strutturano i contesti d'azione della cui riproduzione e del cui mutamento l'agire sociale degli attori di cui sono "traccia" è costitutivo – che si gioca la sfida principale che la disponibilità delle fonti accessibili attraverso *data, web* e *text mining* pone alla conoscenza sociologica, superando l'auto-referenzialità tipica dei singoli sistemi di riferimento[35] da cui queste fonti perlopiù originano e aprendo spazi di ulteriore riflessività per le collettività sociali.

Dipartimento di Sociologia
Università di Napoli «Federico II»

[34] In questo caso, si fa riferimento al titolo del volume di B. Latour, *Reassembling the Social. An Introduction to Actor Network Theory*, New York, Oxford University Press, 2005.

[35] Per superare quella auto-referenzialità "è necessario introdurre un livello di osservazione di secondo ordine", come ben si argomenta in N. Luhmann, *Organizzazione e decisione*, Milano, Bruno Mondadori, 2005.