Filosofi(e)Semiotiche, Vol. 8, N. 1, 2021
ISSN 2531-9434

Dati mancanti, dati mancati. *Cancel culture*, *data bias* e *data gap* nell'era dei *big data*

Valentina Manchia
CROSS; Politecnico di Milano; Università di Bologna
valentina.manchia@polimi.it

Abstract
In a media landscape increasingly impregnated with cancel culture, it is extremely urgent to deal with the way in which sign destructure and erasure is itself the establishment of a regime of signification.
The specific examples of erasure that we would like to discuss in this article, however, propose to broaden the reflection from the practices and strategies of erasure, replacement, obliteration and substitution of already given cultural units to the practices that make possible the emergence of these cultural units as given, in fact preventing the emergence of other cultural units, as shown by recent studies on data gaps and data bias.
More specifically, what we would like to point out in these pages is the complicated relationship between data (statistical data but also big data) and cancel culture. As we shall see, the discourses that, implicitly or explicitly, open a reflection on how data are not given at all but, on the contrary, are in fact objects of continuous negotiation, on the one hand inaugurate a new way of looking at data, very different from the classic functionalist and positivist approach, and on the other hand offer new insights into the dynamics of cancellation itself, when they are no longer exercised over types and tokens, but over the conditions of possibility of such types and tokens.

Keywords: semiotics, big data, Wikileaks, data gap, data bias

In un panorama mediatico sempre più impregnato di *cancel culture*, sia dei discorsi intorno a questo tema che dei discorsi che *fanno cancel culture*, di fatto rendendo ancora più visibili strategie di "competizione semiotica" (MAZZUCCHELLI 2017) sempre esistite, dall'iconoclastia alla censura politica, risulta estremamente urgente occuparsi del modo in cui la cancellazione segnica è essa stessa instaurazione di un regime di significazione.
Gli specifici esempi di cancellazione di cui vorremmo discutere in quest'articolo, tuttavia, propongono di allargare la riflessione dalle pratiche e dalle strategie di cancellazione, sostituzione, obliterazione e sostituzione di unità culturali già date alle pratiche che rendono possibili l'emersione di tali unità culturali come *date*, di fatto impedendo l'emersione ad altre unità culturali.
Più nel dettaglio, quello che vorremmo rilevare in queste pagine è il complicato rapporto, così come sta emergendo nei discorsi mediatici sul tema, tra dati (dati statistici ma anche *big data*) e *cancel culture*. Come vedremo, i discorsi che, implicitamente o esplicitamente, aprono una riflessione su come i dati non siano affatto *dati* ma al contrario siano di fatto oggetti di negoziazione continua da un lato inaugurano un modo inedito di guardare ai dati, molto diverso dall'approccio funzionalista e positivista classico, dall'altro offrono nuovi spunti di riflessione sulle dinamiche di cancellazione stessa, quando non si esercitino più su occorrenze segniche concrete, ma sulle condizioni di possibilità di tali occorrenze.

1. Cancellazione come occultamento. L'esempio di WikiLeaks, o dei dati "nudi e crudi" contro le pratiche di cancellazione governativa

Per esaminare meglio i casi in cui diventa rilevante parlare di pratiche e di strategie di cancellazione nella costituzione stessa dei dati, può essere utile concentrarsi, in prima battuta, sul caso emblematico di WikiLeaks.

È stato già notato, da più parti, come, la sete di trasparenza di Julian Assange non possa che ricorrere a delle strategie di visibilità (e dunque di opacità) mediatica (GOMEZ 2015, FRANCESCUTTI 2015) e, nello specifico, come l'effetto di presa diretta sul reale cui Wikileaks aspira non possa che essere il risultato di una complessa costruzione discorsiva a opera dei grandi quotidiani cui i *leaks* sono stati affidati, dall'interpretazione dei dati fino alla messa a punto di specifici dispositivi grafici capaci di prenderli in carico e di restituirli visivamente (MANCHIA 2020).

Quello che invece ci interessa riepilogare qui non è tanto come WikiLeaks abbia veicolato, anche attraverso la collaborazione con la stampa, dati e informazioni, ma quale sia lo statuto dei dati che WikiLeaks rende pubblici in quanto "verità nude e crude", "*unvarnished truth*".[1]

Per riepilogare in breve il funzionamento di Wikileaks, potremmo dire che la definizione dell'organizzazione internazionale che fa capo a Julian Assange è implicita nella scelta del nome, che nasce come fusione tra il prefisso *Wiki-* che, a partire da Wikipedia, sta per diffusione libera e open source di un sapere aperto, e il verbo *to leak*, al contempo intransitivo (*trapelare*, detto appunto delle informazioni, su estensione del significato letterale del verbo, che è quello di perdere contenuto, come di un vaso forato) e transitivo (con il senso di *rivelare*). In entrambi i casi, si tratta di qualcosa che passa *oltre* un segno dato, che *deborda* dai limiti che le sono stati assegnati.

In un certo senso, potremmo, prima di tutto, definire questi limiti come parte dello statuto di *documento-monumento* assegnato dal potere a determinati dati e informazioni nel momento in cui vengono comunicati, e, all'opposto, la strategia di Wikileaks come una strategia di abolizione di questi limiti e al contempo di ridefinizione dell'oggetto stesso della questione, i dati.

Tornando a *documento* e *monumento*, secondo l'accezione di Zumthor e di Foucault (1969) – e di Le Goff (1978) che nella sua definizione per l'*Enciclopedia Einaudi* fa il punto sul campo semantico che circonda e stringe insieme questi due concetti chiave della prospettiva documentaria e della ricerca storica – il documento che scaturisce dalle procedure della lettura storica e che quindi passa per selezione, connessione di dati ad altri dati e di fatti ad altri fatti è già un *monumento*. Il *documento-monumento* scaturisce, infatti, da un processo di interpretazione che finisce per costruirlo a partire dal punto di vista di chi lo individua, in un processo che non può che stringere insieme chi vede e quello che è visto.[2]

L'impressione è che Le Goff voglia eliminare – o quantomeno indebolire – l'associazione diretta tra documento e verità, rimarcando che ogni documento è costruzione, e al contempo sciogliere la falsa equivalenza tra monumento e mistificazione, mostrando piuttosto, con Foucault, che occorre rivolgere la propria attenzione critica "al discorso inteso nel suo proprio spessore, come *monumento*" (FOUCAULT 1969, trad. it.: p. 184).

In questo senso, allora, quella tra documento e monumento, più che un'opposizione categorica, diventa una tensione tra due estremi ideali, tra il polo della trasparenza e dell'immediatezza e il polo del montaggio, della costruzione, della messa in forma a partire da una specifica prospettiva di visione.

Impossibile, dunque, che si dia documento senza monumento, trasparenza senza opacità – e lo stesso insegna la semiotica a prescindere da dove posi il suo sguardo e dalla tradizione con la quale scelga di dialogare (il pensiero corre, qui, tanto a Louis Marin, all'impossibilità di una

[1] "We are fearless in our efforts to get the unvarnished truth out to the public": così nella presentazione ufficiale del sito (<https://wikileaks.org/About.html>).

[2] "Il documento non è innocuo. È il risultato prima di tutto di un montaggio, conscio o inconscio, della storia, dell'epoca che lo hanno prodotto" (Le Goff 1978: p. 46).

rappresentazione che non re-*présente*,[3] quanto a Bruno Latour e a Françoise Bastide, alla costruzione dell'oggetto da laboratorio sotto lo sguardo dello scienziato[4]).

Partendo da queste categorie, e guardando alle pratiche di acquisizione, validazione e diffusione dei documenti protagonisti delle fughe di notizie da parte dell'organizzazione internazionale di Assange, appare chiaro come la retorica di WikiLeaks sia volta a descrivere il potere come un costruttore di documenti che sono in realtà *monumenti*, operazioni di montaggio e di menzogna, e WikiLeaks stessa come l'operatore critico che riporta alla luce i *documenti* originari, portando alla luce l'autentico che le barriere e le costrizioni delle interpretazioni governative tendono a omettere.

Prima di passare agli esempi, può essere utile dare qualche ulteriore coordinata sull'operato di WikiLeaks in generale. L'organizzazione, infatti, opera ricevendo, in modo anonimo, grazie a un contenitore (detto *drop box*) protetto da un potente sistema di cifratura, documenti coperti da segreto (di vario tipo: di Stato, militare, industriale, bancario) e rendendoli poi accessibili e disponibili sul suo sito Web per chiunque li voglia leggere. Secondo dichiarazioni rese nel 2010, i documenti ricevuti sono controllati da un gruppo di cinque revisori con competenze in campi diversi, ma la decisione finale circa la valutazione di un documento spetta invece allo stesso Assange.

In casi particolarmente rilevanti, come quello degli *Afghan War Logs*, ovvero la pubblicazione nel 2010 di una raccolta di 91.731 documenti militari relativi alla guerra in Afghanistan dal 2004 al 2009, i documenti sono stati rilasciati e pubblicati anche da *Guardian*, *New York Times* e *Der Spiegel*. Lo stesso si è verificato con le rivelazioni di Bradley Manning sulla guerra in Iraq (e che sono costate al soldato americano ben 35 anni di reclusione).

Al di là, però, e prima del rapporto di WikiLeaks con i giornali che diffondono i *logs*, è interessante vedere come l'organizzazione costruisce discorsivamente la sua immagine sull'unico fronte attraverso il quale comunica direttamente con l'esterno: il proprio sito web. È infatti su <https://www.wikileaks.org/> che WikiLeaks pubblica e rende pubblici i documenti riservati ai quali decide di dare rilevanza.

La piattaforma, costruita sul modello di Wikipedia, con una ricerca interna e una suddivisione interna alle varie voci, pur trattando documenti e dati molto diversi lo fa in modo sistematico, su alcune costanti strutturali.

Tre sono, in particolare, gli esempi che prenderemo in considerazione: l'accordo TPP (Trans-Pacific Partnership), ovvero un accordo segreto e internazionale sul libero scambio, il caso FinFisher, società tedesca accusata di malware per via di un software spyware, SpyFiles, e i *Syria files.*[5]

Questi tre casi, diversi in quanto a contenuti, condividono un'unica struttura:

- l'*atto* con cui WikiLeaks istituisce i documenti in quanto tali, secondo una formula del tipo "oggi, [segue data] WikiLeaks pubblica i dati relativi a…", che ha una funzione performativa perché battezza, nominandoli e descrivendoli, eventi prima fuori dalla nostra portata di lettori/utenti;
- il *giudizio interpretativo* sull'evento appena costituito, dato da uno dei membri del board di WikiLeaks (identificato con nome e cognome *vs* l'anonima formula "WikiLeaks pubblica…" di cui sopra) o più spesso dello stesso Julian Assange.

[3] Come diceva Marin, ogni mappa è al contempo *dessin* (la rappresentazione grafica) e *dessein* (l'intenzione che la struttura): ogni mappa è al contempo ciò che vi è rappresentato e lo sguardo che la *re-présente*: "In quanto rappresentazione, una mappa significa […] e allo stesso tempo mostra che significa" (Marin 1983, trad. it.: p. 77). Una mappa, suggerisce Marin, non può non essere opaca, proprio perché non può non portare la traccia del punto di vista che la struttura.

[4] Il riferimento, qui, oltre che alla concezione del laboratorio come soggetto di produzione di senso e ai *laboratory studies* legati a Latour, è alla formalizzazione, decisamente semiotica, di questo tema fatta da Bastide (si veda, in particolare, Bastide 1985).

[5] I documenti e i testi cui si fa riferimento sono visionabili ai rispettivi link: https://wikileaks.org/tpp-ip2/pressrelease/; https://wikileaks.org/spyfiles4/index.html; https://wikileaks.org/Syria-Files.html.

Ecco, per esempio, il testo che introduce la pubblicazione del documento dell'accordo TPP:

> Today, Thursday 16 October 2014, WikiLeaks released a second updated version of the Trans-Pacific Partnership (TPP) Intellectual Property Rights Chapter. The TPP is the world's largest economic trade agreement that will, if it comes into force, encompass more than 40 per cent of the world's GDP. […]
> Despite the wide-ranging effects on the global population, the TPP is currently being negotiated **in total secrecy** by 12 countries.[6]

Il testo prosegue ricordando le ristrette categorie di persone che hanno finora avuto accesso al testo, evidenziando la disparità tra i pochi che hanno letto il testo dell'accordo per intero, i grandi gruppi che hanno avuto accesso a larghe porzioni del testo, e il grande pubblico che è stato totalmente tenuto all'oscuro ("in total secrecy").

Segue, virgolettata e opportunamente separata dal resto del testo, l'interpretazione di Assange:

> **Few people**, even within the negotiating countries' governments, have access to the **full text** of the draft agreement and the public, who it will affect most, none at all. Large corporations, however, are able to see portions of the text, generating a powerful lobby to effect changes on behalf of these groups and bringing developing country members reduced force, while the public at large gets no say.
> **Julian Assange, WikiLeaks' Editor-in-Chief, said:**
> "The **selective secrecy** surrounding the TPP negotiations, which has let in a few cashed-up megacorps but excluded everyone else, reveals a telling fear of public scrutiny. By publishing this text we allow the public to engage in issues that will have such a fundamental impact on their lives."

Un altro elemento importante cui fare attenzione è l'insistenza, polemica e dalla forte carica retorica, su termini come "segreto" ("*in total secrecy*", "*selective secrecy*"), "mai visto" e simili nella comunicazione di WikiLeaks. Un'insistenza che risulta accentuata dalla scomposizione della notizia in un documento presentato come evento, per così dire, e in un commento al documento-evento.
Importante anche notare che, sia che si ponga come rivelazione di un evento prima non noto ai più, e quindi del tutto *segreto*, sia che si ponga come rivelazione di dettagli ulteriori in merito a eventi già noti (come il caso di FinFisher e di FinSpy), il documento è sempre esibito come documento in quanto tale, ovvero come *documento-verità*, non come documento-monumento forgiato e diffuso dal potere di turno.
Questo perché, nella cornice interpretativa del discorso di Assange e dei suoi, i documenti collegati alle fughe di notizie di WikiLeaks ripristinano il materiale originario e trasparente da cui il potere è partito per edificare i propri documenti a guisa di monumento, per dirla con Zumthor. In altre parole, i documenti di WikiLeaks sono descritti come documenti *liberati*, strappati alle maglie del potere, discorsi che debordano oltre i confini che decidono di quello che deve essere taciuto e occultato.
Nel primo caso, per esempio, quello dell'accordo TPP, si tratta di materiale mai divulgato, tenuto segreto, e poi riportato alla luce. Nel secondo, il caso FinSpy, dell'aggiunta di elementi nuovi al quadro dell'operato dell'azienda tedesca, e della conseguente rivelazione degli accordi con i paesi sospettati di perseguire illegalmente giornalisti e oppositori del regime, come in Bahrein. Nel terzo caso, quello dei *Syria files*, della rivelazione di molti elementi inediti (raccolti in un database di

[6] I bold delle citazioni, qui e più avanti, sono di chi scrive.

oltre 2 milioni di dati, con e-mail e documenti di aziende, capi di stato, uffici diplomatici) sulla situazione in Siria sotto il regime del dittatore Bashar Assad.
Occorre specificare, infatti, che i documenti con cui WikiLeaks ha a che fare, e che mette in disposizione sul sito, pur essendo completamente accessibili non sono facilmente maneggevoli: sono liberamente consultabili, ma data la mole di dati sono organizzati in grandi dataset che devono a loro volta essere interrogati, spesso per mezzo di appositi algoritmi e di figure professionali addestrate alla caccia dei dati (i data journalist), per poter dare risposte.
Sempre per riprendere le categorie di Le Goff, quelli liberamente scaricabili dal sito di WikiLeaks sono documenti in cui il montaggio (e quella parte di menzogna che il montaggio include in sé, come specifica lo storico) sembra non essere ancora intervenuto – o meglio, sono presentati come i documenti originali, prima che la fase di montaggio e di costruzione li trasformi a uso dei governi o degli altri centri di potere.
In altre parole, quelli che WikiLeaks presenta come documenti sono in realtà dei documenti *prima* dei documenti: la massa informe di dati che occorre saper processare prima di poterne ricavare delle informazioni strutturate, o, nel caso dei dati che completano montaggi e documenti-monumenti parziali (e di parte), il ricongiungimento degli elementi *tagliati fuori* dall'inquadratura costrittiva (e restrittiva) del potere.
Dallo specifico punto di vista di quest'articolo, le strutture governative operano per cancellazione, mettendo in circolo, tramite i mezzi di comunicazione, soltanto una selezione dei dati a disposizione. Più nel dettaglio, alle pratiche di cancellazione applicate dalle strutture di potere, che operano per selezione e per montaggio della totalità originale di dati e informazioni, corrisponde anche un'ostensione dei dati scelti per essere comunicati come ufficiali, esibiti come unici dati disponibili e pertanto come unica verità immessa nel circuito comunicativo e mediatico.
Questa esibizione "ufficiale" dei cosiddetti "dati di fatto" da parte delle istituzioni è dunque comunicata come un'*ostensione* per occorrenza, nel senso in cui Eco utilizza questo termine, accennando per esempio al "linguaggio puramente ostensivo […] dei saggi dell'isola di Laputa che recavano in un sacco tutti gli oggetti di cui dovevano parlare" (ECO 1975: p. 294), ma sarà allo stesso tempo da intendersi anche come un'*invenzione*, ovvero come invenzione di un nuovo codice di significazione.
Al contrario – e un'analisi ancora più estesa dei formulari utilizzati nella comunicazione ufficiale di Wikileaks lo mostrerebbe bene – Assange e i suoi operano, all'inverso, per ripristinare ciò che è stato espunto dall'operazione di cancellazione governativa, dunque per restaurare quella totalità originaria che si suppone però, una volta acquisita, completamente trasparente e bastante a sé stessa: "One of our most important activities is to publish original source materials alongside our news stories so readers and historians alike can see evidence of the truth" (Cfr. <https://wikileaks.org/About.html>).
L'evidenza della verità, tuttavia, che è lasciata ai "lettori" e agli "storici", non è soltanto esibita. Ovvero, nemmeno Wikileaks procede attraverso un'ostensione totale di tutte le occorrenze, come ci si potrebbe aspettare.
Ogni *leak* è infatti accompagnato, come abbiamo visto, da una specifica chiave di lettura: il virgolettato di Assange, o del rappresentante di WikiLeaks di turno, che sottolinea e accompagna "l'evidenza della verità" a partire da un taglio ben preciso, che introduce di fatto un'interpretazione dei dati. Assange parla da testimone, proprio perché non c'è dato sul portale che egli non abbia vagliato, ed è da questa posizione di osservazione privilegiata che dalle masse di dati maneggiati da WikiLeaks prendono forma, loro malgrado, dei nuovi *documenti-monumenti*. Anche la cosiddetta "evidenza della verità", insomma, se *raccontata*, non può sfuggire a questa condanna.

2. Cancellazione come mancata produzione di dati. Pratiche di costruzione e pratiche di cancellazione tra *data gap* e *data bias*

Nel discorso di Wikileaks, dunque, il ripristino dei puri dati, riportati "così come sono", si dice sufficiente a colmare la cancellazione originaria che deriva dal monumentale e irreggimentante agire delle strutture governative, ed è allo stesso tempo capace di arricchire di nuove possibilità quello che Eco (1975) chiamerebbe il continuum del contenuto, reso disponibile per nuove e inedite pertinentizzazioni a partire dalla massa dei dati forniti.

Corre sotterranea, in questa presa di posizione, la certezza dell'evidenza puramente trasparente dei dati stessi che, pur nella loro molteplicità e nella loro grande complessità (si è poco sopra evidenziata la grande mole di informazioni e documenti relativi agli *Afghan War Logs*), sono pur sempre l'unica e prima base di partenza di ogni interpretazione successiva.

Questa certezza sfocia in quella che è una vera e propria retorica della trasparenza e dell'oggettività dei dati in quanto tali: una posizione che è stata imperante nella tradizione funzionalista, dal *graphic design* e dalla tipografia (basti pensare a Max Bill ma anche a Beatrice Warde)[7] fino all'*information design* di Edward R. Tufte e alla *sémiologie graphique* di Jacques Bertin, e di cui si trova traccia fin dagli scritti di Otto Neurath, sociologo e filosofo del Circolo di Vienna e padre del metodo Isotype, uno dei primi metodi di visualizzazione di dati e informazioni, dedicati al modo in cui le scienze del pensiero possano fare da supporto alle scienze matematiche e statistiche.[8]

Secondo Neurath, infatti, il principale compito della nuova filosofia diventa quello di aiutare la scienza a descrivere, maneggiare e strutturare i "dati" sensibili, evidenti e primari in quanto tali:

> What marks the modern scientific world-view is this: each statement that does not fit without contradiction into the total structure of laws must disappear; each statement that does not rely on formulations that relate to "data" is empty, it is metaphysics. (NEURATH 1931: p. 326).

In questo quadro, in cui i dati – come suggerisce l'etimologia della parola che ha a che vedere con il loro essere "dati", dal verbo latino (e poi italiano) *dare* – ci sono appunto dati, ovvero sono qualcosa che arriva a noi, ai nostri sensi, e che ci immaginiamo sia perfettamente autonomo da noi e completamente oggettivo, eventuali pratiche di cancellazione sono supposte agire sui dati come su oggetti già concreti, tangibili, formati.

Tali pratiche, volte a dire il falso, a nascondere il vero, o frutto dell'incapacità tecnica di un «vedere chiaro» al servizio di un «pensare chiaro», per dirla con Tufte (1997)[9], possono essere di fatto descritte, seguendo Mazzucchelli (2017), come pratiche di alterazione, di contraffazione, di nascondimento, di falsificazione di quello che i dati già naturalmente "direbbero".

Tufte (1997: pp. 55-71), per esempio, ha dedicato molte pagine all'analisi del *disinformation design*, da intendersi come incapacità di rispecchiare quello che i dati offrirebbero quasi spontaneamente alla vista, e a esempi classici di inadempienza (fino alla *defigurazione* estrema, per dirla con Mazzucchelli 2017) della visualizzazione dei dati come nel drammatico caso del lancio dello space shuttle Challenger (1997: pp. 38-53).

Tuttavia, all'aumentare esponenziale sia dei dati disponibili sia della loro complessità, è risultato sempre più evidente l'apporto fondamentale e fortemente strutturante delle procedure sia di messa in forma e di interpretazione che di raccolta dei dati – procedure già messe in evidenza, a proposito

[7] Sul rapporto tra trasparenza e opacità nel graphic design e nella progettazione tipografica, in una prospettiva semiotica, ci permettiamo di rimandare a Manchia 2012).

[8] Il linguaggio visuale Isotype (International System of TYpographic Picture Education) avrebbe dovuto costituire, nelle intenzioni del suo inventore, un linguaggio semplice, immediato, capace di visualizzare anche concetti complessi al di là delle barriere linguistiche e di fare da base a una vera e propria *visual education* per le classi economiche meno scolarizzate. Cfr. in particolare Neurath (1945) e (1946).

[9] «Visual representations of evidence should be governed by principles of reasoning about quantitative evidence. For information displays, design reasoning must correspond to scientific reasoning. Clear and precise seeing becomes as one with clear and precise thinking» (Tufte 1997: p. 53).

della costruzione del discorso scientifico, dalla sociologia della scienza, dalla semiotica e dagli STS,[10] e rilevate e analizzate anche nel funzionamento dei cosiddetti *big data* in studi all'incrocio tra scienze sociali, *media studies*, e *computer science*.[11]
Di conseguenza sta rapidamente emergendo, nel vasto panorama di discorsi attorno ai dati, panorama che si allarga e si estende di giorno in giorno sia tra gli addetti ai lavori che nella saggistica di settore, un interesse del tutto nuovo per le forme di costruzione dei dati e per le parallele forme di cancellazione che si esercitano non su dati esistenti (*token*, per riprendere la terminologia cara a Eco 1975 e ripresa da Mazzucchelli 2017) o su *type* codificati ma sulla possibilità stessa dell'emersione di nuovi dati e allo stesso tempo di nuovi filtri interpretativi.
Nello specifico, c'è chi ha messo in luce i *bias* impliciti nell'esistenza stessa del fenomeno *big data* in quanto tale, come fenomeno tecnologico-sociale che implica dinamiche di costruzione complessa che spesso non vengono considerate (BOYD e CRAWFORD 2012) o ha messo in guardia contro un «naîve usage of social data» da parte dei ricercatori (OLTEANU et al 2019).
Inoltre, non manca chi, parlando di *data gap*, ha mostrato l'incidenza, spesso non rilevata, delle pratiche di "messa in ombra" dei dati (LERMAN 2013; HAND 2020). Esistono, ovvero, dati non raccolti e non registrati, e pertanto fenomeni che di fatto restano "invisibili", pur se esistenti, perché su di essi non si possono (o non si vogliono) raccogliere dati. Lerman, a questo proposito, parla della «nonrandom, systemic omission of people who live on big data's margins», e dunque di «big data's exclusions» (LERMAN 2013: p. 57).
In altre parole, parlare di *data gap* e di *data bias* implica prima di tutto riconoscere che ogni set di dati è appunto *un* set di dati, non una fotografia perfetta e completa di quello che esiste; e, in seconda istanza, porre l'attenzione al fatto che le categorie che danno forma ai dati istituiscono nuove unità di senso ma allo stesso tempo impediscono l'applicazione di altre categorie e l'emersione di altri dati possibili.
Alcuni esempi interessanti di questa nuova prospettiva sui dati sono per esempio contenuti nei saggi che compongono *Invisibili. Come il nostro mondo ignora le donne in ogni campo* (2019), della scrittrice, giornalista e attivista Caroline Criado Perez. Nelle parole dell'autrice:

> *Invisibili* racconta quel che succede quando ci si dimentica di prendere in considerazione metà del genere umano. Denuncia i danni provocati dall'assenza di dati di genere lungo il corso piú o meno normale della vita di ogni donna. Pianificazione urbana, politica, lavoro: questi sono alcuni dei settori dove il danno è piú evidente. Per non parlare della sorte che, in un mondo costruito su dati maschili, attende le donne a cui le cose vanno male. (CAROLINE CRIADO PEREZ 2019, trad. it.: ebook)

L'"assenza di dati di genere" (dunque la raccolta di dati genericamente attribuiti a individui senza distinzioni di genere piuttosto che a uomini e donne) è di fatto, per tutto il libro, dipinta come una cancellazione. In precedenza, da Neurath ad Assange, erano i dati i fatti da cui partire; detto altrimenti, erano i nudi, puri e crudi dati a costituire la verità cui fare riferimento, il documento di grado zero su cui esercitare un'interpretazione.
Nei saggi di Criado Peres (2019), invece, così come più in generale nella letteratura già citata intorno ai *data bias*, emerge una considerazione del tutto nuova: i dati sono solo il punto di arrivo di una costruzione che inizia nel momento in cui si decide di porre l'attenzione su un fenomeno a partire da una determinata prospettiva.
Un primo caso interessante, tra i tanti, è il caso che ha coinvolto il bisfenolo A, un composto chimico utilizzato fin dagli anni Cinquanta nella fabbricazione di oggetti di plastica, «presente in

[10] Cfr. in particolare Latour (1987, 1996, 2011), ma anche Callon (1986), Latour, Woolgar (1979) e Latour, Fabbri (1977).
[11] Cfr. in particolare Boyd, Crawford (2012), Lerman (2013).

milioni di oggetti di consumo, dai biberon ai contenitori di latta per generi alimentari, fino alle condutture dell'acqua».
Nel 2008 fu annunciato che tale composto poteva in realtà essere cancerogeno, causare alterazioni genetiche e altri gravi disturbi e squilibri anche con livelli di esposizione inferiori ai limiti di legge. Tuttavia, la pericolosa capacità del bisfenolo A di "mimare" l'azione degli ormoni estrogeni femminili era nota sin dalla metà degli anni Trenta, così come, sin dagli anni Settanta, l'effetto cancerogeno degli estrogeni sintetici sulle donne. Nonostante questo, dopo l'allarme del 2008 la sostanza fu eliminata solo dalle bottiglie in plastica e dai biberon per neonati, e non ci si attivò in alcun modo per raccogliere dati più specifici sull'impatto del bisfenolo A sulla salute delle donne:

> Quella del Bpa non è solo una storia di genere: è anche una storia di stratificazioni sociali, o almeno di stratificazioni sociali di genere. Per paura di un boicottaggio da parte dei consumatori, la maggior parte dei fabbricanti di bottiglie in plastica per neonati eliminò subito il Bpa dai prodotti; e se negli Stati Uniti la sostanza è ufficialmente dichiarata non tossica, l'Unione europea e il Canada sono sul punto di metterla per sempre al bando. Il problema è che le attuali disposizioni in materia proteggono soltanto i consumatori, e non c'è nessuna norma che limiti l'esposizione alla sostanza sui luoghi di lavoro. «Mi è sempre sembrato paradossale, – dice Jim Brophy, ricercatore nel campo della medicina del lavoro, – che si parlasse tanto dei pericoli per le gestanti e le donne che avevano appena partorito, e mai per chi fabbricava quelle bottiglie. Eppure nelle fabbriche di prodotti plastici i livelli di esposizione al Bpa erano di gran lunga superiori. Ma dell'operaia incinta che azionava la macchina che produceva quelle bottiglie non si è mai parlato». (CAROLINE CRIADO PEREZ 2019, trad. it.: ebook)

Stando al modo in cui Criado Perez espone la questione, la pericolosità del bisfenolo A è dapprima trattata come una questione marginale (senza disamine e indagini più accurate), o al limite ristretta ai consumatori più fragili (neonati e mamme) senza che si abbia alcun particolare riguardo né per le lavoratrici delle industrie chimiche né per altre fasce di consumatrici. Di fatto, è la mancata costruzione di una struttura di rilevazione dei dati – e il *data gap* che ne consegue – a cancellare il problema dell'agenda, o se non altro a minimizzarlo.
Tuttavia, non è solo la mancata attivazione di procedure di rilevazione dei dati a cancellare un fenomeno dal panorama sociale, culturale e in definitiva politico: anche la mancata applicazione di categorie o sottocategorie (capaci di "tagliare" il continuum dei dati secondo nuove pertinenti "linee di tendenza", per rifarsi a Eco 1997) può di fatto cancellare la possibilità dell'emersione di nuovi dati e di nuovi approcci di analisi.
Ne è un esempio l'analisi dell'apporto femminile all'agricoltura, per la quale spesso non sono disponibili dati disaggregati per sesso:

> In un documento della Fao, l'organizzazione delle Nazioni unite per l'alimentazione e l'agricoltura, l'economista Cheryl Doss sostiene che molto dipende anche dalla definizione e dal valore che attribuiamo al termine «raccolto alimentare»: stiamo parlando di apporto calorico (e quindi diamo la precedenza agli alimenti base) o di apporto monetario (e quindi mettiamo al primo posto, per esempio, il caffè)? Poiché le donne «tendono a essere più coinvolte nella produzione degli alimenti base», una comparazione in termini di apporto calorico «potrebbe innalzare di parecchio la quota di produzione agricola attribuibile alle donne».
> La parola chiave, in questo caso, è «potrebbe», poiché le statistiche nazionali spesso non dicono se gli addetti all'agricoltura sono donne o uomini. E anche quando i dati disaggregati per sesso ci sono, l'impianto lacunoso delle ricerche rischia talora di sottostimare l'apporto della manodopera femminile: quando alle donne si chiede di indicare se durante il giorno si occupano dei «lavori domestici» oppure «lavorano» come se le due cose si escludessero a vicenda (o come se i lavori domestici non fossero lavoro), le donne tendono a rispondere «lavori domestici», perché è quello il termine che descrive la maggior parte dei loro impegni quotidiani. Se poi si considera anche la tendenza a «dare più risalto alle attività che generano reddito», il

> risultato finale è una frequente sottovalutazione delle produzioni agricole di sussistenza, spesso affidate alla manodopera femminile. (CAROLINE CRIADO PEREZ 2019, trad. it.: ebook)

Se la ricerca quantitativa adopera dati numerici, siamo dunque invitati ad accorgerci che tali numeri e tali evidenze sono di fatto il diretto prodotto del sistema di categorie adottato per costruire tali dati, come il set di domande nei questionari per la raccolta dati, o il trattamento statistico dei dati raccolti.
Per quanto riguarda quest'ultimo punto, anche le procedure di trattamento dei dati dopo la raccolta, persino le più semplici come l'aggregazione o la disaggregazione di categorie di analisi (nel caso appena esaminato la disaggregazione dei dati per sesso) sono di fatto procedure di selezione di determinati filtri interpretativi che comportano una cancellazione della possibilità stessa di rilevare i dati che nelle maglie di quei filtri non rientrano.
Non includere, pertanto, nei questionari di rilevazione per gli addetti all'agricoltura l'appartenenza di genere cancella di fatto ogni possibilità di ottenere dei dati che siano capaci di fotografare la situazione relativa agli uomini e alle donne nel settore e alle differenze tra loro.

3. Dalla cancellazione "invisibile" al *data gap* come questione politica (e semiotica)

In conclusione, nell'ambito dei *big data*, dell'*information design* e della visualizzazione delle informazioni, proponiamo di guardare non soltanto alle pratiche e alle strategie di cancellazione che creano senso dalla distruzione segnica di *token* e *type* già definiti (come già dicevamo, le strategie per nascondere o minimizzare informazioni invece importanti ma anche le astuzie usate per "mentire con i dati", tra *defigurazione* e *disinvenzione*)[12] ma anche alle pratiche e alle strategie di cancellazione implicite nell'attivazione o nella magnificazione di determinate pertinenze all'interno dei dati raccolti a scapito di altre pertinenze possibili, come chi si occupa di *data bias* e di *data gap* non manca di rilevare.
Già Eco (1975), occorre ricordarlo, rubricava diagrammi e grafi tra le invenzioni, accanto alle immagini e ai testi estetici (sulla scia di Peirce 1931-1935, che tra le icone inseriva anche immagini e diagrammi), sottolineando come parte integrante dell'istituzione di codice in regime di invenzione sia la scelta di «un nuovo continuum materiale non ancora segmentato», su cui applicare una nuova forma «per TRASFORMARE in esso gli elementi pertinenti di un tipo di contenuto».
Nei casi analizzati da Criado Perez (2019) – ma anche in quelli citati in Lerman (2013) – appare ancora più evidente, a monte di ogni espressione possibile, quanto cruciale sia l'individuazione delle pertinenze a livello del contenuto, ovvero, sempre per seguire ancora Eco (1975), quanto sia delicata l'esistenza "culturale" di quelle unità che chiamiamo (e consideriamo) "dati".
È a questo proposito che sarebbe interessante esaminare, da un punto di vista semiotico, quei fenomeni legati ai dati mancanti e ai dati mancati che da più parti si è tentato di categorizzare.
Sarah Giest e Annemarie Samuels (2020), per esempio, interessate al modo i cui i *policymakers* adoperano i dati, rilevando come la qualità dei dati in combinazione con potenziali lacune di dati noti o sconosciuti limiti la capacità dei governi di creare politiche inclusive, hanno proposto una categorizzazione dei *data gap* che distingue tra *primary data gap*, *secondary data gap* e *hidden data gap*.
Il *primary gap* è la vera e propria mancanza di dati, conclamata e nota ai *policymakers*, con possibilità molto limitate di essere colmata; il *secondary data gap* si verifica quando i dati sono sì disponibili, ma non sono nelle mani di chi prende le decisioni per la comunità (per esempio sono di privati, o derivano dai social media); il terzo caso, quello degli *hidden data gap*, è il più insidioso, perché è quello dei *bias* strutturali ma invisibili, se non esplicitamente dichiarati, come è per le

[12] Su questo tema cfr., per esempio, Cairo (2019).

distorsioni, possibili e già note, nel *data mining* e a opera del *machine learning* (cfr. per esempio Crawford, Paglen 2019).
Da un punto di vista semiotico, potrebbe essere interessante ragionare, a partire da qui, sulle diverse modalità di esistenza dei dati, e sui diversi regimi di prassi enunciativa (FONTANILLE, ZILBERBERG 1998) grazie a cui possono emergere nel discorso: il *primary data gap* identifica i dati che potremmo definire "mancati", esistenti solo virtualmente ma comunque *marcati*, ovvero dotati di un peso, proprio per la loro assenza, nel discorso politico e sociale; nel *secondary data gap* ricadono dati "mancanti" ma comunque esistenti sotto altre forme, dunque potenziali e suscettibili di poter essere ottenuti e messi in circolo; nel caso dell'*hidden data gap*, o dei *data bias*, sono in gioco le procedure, tra attualizzazione e realizzazione, che consentono l'esistenza stessa dei dati in quanto tali, in quanto dati all'interno di uno specifico intorno discorsivo.
Quello che abbiamo tentato di suggerire, dunque, è che diversamente da quanto ci si potrebbe aspettare non sono solo l'interpretazione e la conseguente trasposizione visiva dei dati a convocare procedure di invenzione, capaci di guidare una complessa «attività tracciante» (FABBRI 2001: p. 14) attraverso specifiche regole di trasformazione diagrammatica, ma anche l'individuazione, volta per volta differente, delle linee di tendenza lungo le quali segmentare grandi masse di dati per poi poterli farli emergere sulla superficie discorsiva, per così dire, come unità culturali.
Procedure di invenzione e di pertinentizzazione che, come abbiamo tentato di mostrare, non possono che correre di pari passo a procedure di cancellazione, a loro volta strettamente intrecciate alle strategie discorsive e interpretative che rendono *dati* i dati e assegnano loro un senso (e prima ancora una direzione di lettura).

Bibliografia

BASTIDE, F. (1985), «Essai d'épistemologie à partir d'un texte technique sans prétention: une invention peu connue des frères Lumière», *Fundamenta scientiae*, VI, 2, pp. 127-150; trad. it.: «Saggio di epistemologia nato da un testo senza pretese», in LATOUR, Bruno e FABBRI, Paolo [a cura di], *Una notte con Saturno. Scritti semiotici sul discorso scientifico*, Roma, Meltemi, 2001, pp. 215-250.
BERTIN, Jacques (1967), *Sémiologie graphique. Les diagrammes, les réseaux, les cartes*, Paris-La Haye, Gauthier-Villars/Mouton & Cie.
BOYD, Danah, CRAWFORD, Kate (2012), «Critical questions for big data. Provocations for a cultural, technological, and scholarly phenomenon», in *Information, Communication & Society*, 15:5, pp. 662-679, DOI: 10.1080/1369118X.2012.678878
CALLON, Michel (1986), «Elements of a Sociology of Translation: the Domestication of the Scallops and the Fishermen of StBrieuc Bay", in LAW, John [a cura di], *Power, action and belief. A new sociology of knowledge*, London, Routledge & Kegan Paul.
CAIRO, Alberto (2019), *How Charts Lie: Getting Smarter about Visual Information*, London, W.W. Norton & Company.
CRIADO PEREZ, Caroline (2019), *Invisible Women: Exposing Data Bias in a World Designed for Men*, Vintage; tr. it. *Invisibili. Come il nostro mondo ignora le donne in ogni campo. Dati alla mano*, Torino, Einaudi 2020.
CRAWFORD, Kate, PAGLEN, Trevor (2019), «Excavating AI: The Politics of Training Sets for Machine Learning», https://excavating.ai.
ECO, Umberto (1975), *Trattato di semiotica generale*, Milano, Bompiani.
ECO, Umberto (1997), *Kant e l'ornitorinco*, Milano, Bompiani.
FABBRI, Paolo (2001), introduzione a LATOUR, Bruno e FABBRI, Paolo [a cura di], *Una notte con Saturno. Scritti semiotici sul discorso scientifico*, a cura di B. Latour e P. Fabbri, Roma, Meltemi, 2001, pp. 9-23.

FONTANILLE, Jacques, ZILBERBERG, Claude (1998), *Tension et signification*, Mardaga, Liège.
FOUCAULT, Michel (1969), *L'archéologie du savoir*, Paris, Gallimard; trad. it., *L'archeologia del sapere*, Milano, Rizzoli, 1971.
FRANCESCUTTI, Pablo (2015), «La trasparenza mediatizzata: il Cablegate e l'agenda del quotidiano El Pais», in ALBERGAMO, Maria [a cura di], *La trasparenza inganna*, Roma, Luca Sossella, 2015, pp. 97-110.
GIEST, Sarah, SAMUELS, Annemarie (2020), «'For good measure': data gaps in a big data world», in *Policy Sci* 53, pp. 559-569. https://doi.org/10.1007/s11077-020-09384-1
GOMEZ, Oscar (2015), «Il fenomeno Wikileaks e la trasparenza», in ALBERGAMO, Maria [a cura di], *La trasparenza inganna*, Roma, Luca Sossella, 2015, pp. 86-96.
GOMEZ, Oscar, SERRA, Marcello (2015) [a cura di], *Transparencia y secreto*, atti del II Congresso internazionale GESC (Grupo de Estudios de Semiótica de la Cultura), Madrid, 20-22 novembre 2013, Madrid, Visor Libros, pp. 33-46.
HAND, David J. (2020), *Dark data. Why what you don't know matters*, Princeton, Princeton University Press.
LE GOFF, Jacques (1978), voce "Documento/monumento", *Enciclopedia Einaudi*, Torino, Einaudi, vol. V, pp. 38-48, 1978.
LATOUR, Bruno (2011), «La semiotica dei testi scientifici dopo il lavoro di Françoise Bastide», in E|C, pp. 1-7. Web. 30 maggio 2013.
LATOUR, Bruno (1996), *Petite réflexion sur le culte moderne des dieux faitiches*, Le Plessis-Robinson, Synthélabo; trad. it.: *Il culto moderno dei fatticci*, Roma: Meltemi, 2005.
LATOUR, Bruno (1987), *Science in Action. How to Follow Scientists and Engineers through Society*, Harvard University Press; trad. it.: *La scienza in azione. Introduzione alla sociologia della scienza*, Edizioni di Comunità, Torino, 1998.
LATOUR, Bruno, FABBRI, Paolo (1977) «La rhétorique de la science. Pouvoir et devoir dans un article de science exacte», in *Actes de la Recherche en sciences sociales*, Paris, Minuit (tr. it. in FABBRI, Paolo, MARRONE, Gianfranco [a cura di], *Semiotica in nuce I. I fondamenti e l'epistemologia strutturale*, Roma, Meltemi, 1999, pp. 260-279).
LATOUR, Bruno, WOOLGAR, Steve (1979) *Laboratory Life. The Construction of Scientific Facts*, Los Angeles, Sage Publications (ed. rivista, *Laboratory Life: the Construction of Scientific Facts*, Princeton, Princeton University Press, 1986.
LERMAN, J. (2013), «Big data and its exclusions», in *Stanford Law Review Online*, 66, 55–63.
LOZANO, J. (1987), *El discurso histórico*, Alianza Editorial, Madrid; trad. it.: *Il discorso storico*, Sellerio, Palermo, 1991.
MARIN, Louis (1983), «La ville dans sa carte et son portrait : proposition de recherche», in *De la représentation*, Paris, Seuil-Gallimard, pp. 204-218; trad. it.: «La mappa della città e il suo ritratto. Proposte di ricerca», in CORRAIN, Lucia [a cura di], *Della rappresentazione*, Milano, Mimesis, 2014, pp. 75-137.
MANCHIA, Valentina (2012), *Il calice d'oro e il calice di cristallo. Per un'analisi delle forme di figurazione nella scrittura: la tipografia espressiva e le interprétations typographiques di Massin*, Tesi di dottorato, Università degli Studi di Siena.
MANCHIA, Valentina (2020), *Il discorso dei dati. Note semiotiche sulla visualizzazione delle informazioni*, Milano, FrancoAngeli.
MAZZUCCHELLI, Francesco (2017), «Modi di distruzione segnica. Come si arresta la semiosi?», in Versus, n. 124, Fascicolo 1, gennaio-giugno 2017, pp. 105-128.
NEURATH, Otto (1931), «Empirical Sociology. The Scientific Content of History and Political Economy», in NEURATH, Marie, Cohen, Robert S. [a cura di] (1973), *Empiricism and Sociology*, Dordrecht-Boston, Reidel, pp. 319-421.

NEURATH, Otto (1945), «Visual education: humanisation versus popularisation» [unfinished manuscript]. In M. Neurath & R.S. Cohen (eds.) (1973), *Empiricism and Sociology* (pp. 227-248). Dordrecht-Boston: Reidel.
NEURATH, Otto (1946), *From Hieroglyphics to Isotype*, London, Future Books; M. Eve & C. Burke (eds.) (2010) *From hieroglyphics to Isotype: a visual autobiography*. London: Hyphen Press.
OLTEANU, Alexandra, CASTILLO, Carlos, DIAZ, Fernando, KICIMAN, Emre (2019), «Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries», in *Front. Big Data* 2:13. doi: 10.3389/fdata.2019.00013
PEIRCE, Charles Sanders (1931-1935), *Collected Papers*, Cambridge, Mass., Harvard University Press (tr. it. parziale in *Opere*, a cura di M.A. Bonfantini, Milano, Bompiani, 2003).
TUFTE, Edward R. (1983) *The visual display of quantitative information*, Cheshire, Conn., Graphics Press.
TUFTE, Edward R. (1990), *Envisioning Information*, Cheshire, Conn., Graphics Press.
TUFTE, Edward R.(1997) *Visual Explanations. Images and Quantities, Evidence and Narrative*, Cheshire, Conn., Graphics Press.
ZUMTHOR, Paul (1960), «Document et monument: À propos des plus anciens textes de langue française», in *Revue des sciences humaines*, 97, pp. 5-19.