Università degli Studi di Pisa

FACOLTÀ DI FILOLOGIA LETTERATURA E LINGUISTICA
Corso di Laurea Magistrale in Informatica Umanistica

Tesi di laurea magistrale

Tutela della reputazione personale sui motori di ricerca

Applicazione di strategie di correzione dell'identità digitale e realizzazione di un crawler per il monitoraggio della SERP di Google

Candidato:
Martino Bartalesi
Matricola 512444

Relatore:
Nicoletta Salvatori

Correlatore:
Maria Simi

Anno Accademico 2016-2017

Indice

Introduzione

Quando Google parla di noi

Poco tempo fa mia nonna, ottantatré anni, non particolarmente in sintonia con le nuove tecnologie ma abbastanza aggiornata da saper maneggiare un *tablet*, mi ha detto: «Quando incontro qualcuno per la prima volta, dopo lo vado a cercare su Google, così poi so tutto su di lui».

Era già diverso tempo che mi interessavo al tema della reputazione *online*, conscio della sua estrema attualità e soprattutto dell'influenza, ancora poco chiara ai più, che esercita sulla cosiddetta "generazione Internet" di cui faccio parte. Tuttavia, non posso nascondere lo stupore che mi ha colto quando ho realizzato che la questione è chiara, a quanto pare, persino alle nostre nonne: la pagina dei risultati di Google per la ricerca del nostro nome e cognome è una vera e propria carta d'identità digitale.

Il problema è che quella carta d'identità non l'abbiamo compilata noi e nemmeno l'impiegato del comune: è opera di Google e ciò che ne salta fuori potrebbe non piacerci, in particolar modo se in Rete è rimasta qualche vecchia notizia sul nostro conto di cui non andiamo tanto fieri.

Un *collage* di contenuti testuali, immagini, video: questa è la nostra identità digitale. È il riflesso frammentario e approssimativo della nostra identità *offline* ricostruita e interpretata attraverso le informazioni che ci riguardano sparpagliate per il Web. Pertinenza, aggiornamento, contesto non sono criteri coinvolti nella selezione dei contenuti che vanno a fare parte di questo *collage*, tutto ciò che conta è la visibilità e accessibilità delle informazioni.

Per comprendere quale sia il rapporto tra identità digitale e vita *offline*, si può citare l'aneddoto della "piratessa ubriaca"[1]. Si tratta della storia di

[1] J. Rosen, *The Web Means the End of Forgetting*, "New York Times", 20 luglio 2010, http://www.nytimes.com/2010/07/25/magazine/25privacy-t2.html?pagewanted=all, 17/4/2017.

Stacy Snyder, una venticinquenne di Lancaster in Pennsylvania, che stava facendo un tirocinio come insegnante in una scuola superiore. Un giorno pubblicò, sulla sua pagina My Space, una foto che la ritraeva ad una festa, mentre beveva da un bicchiere di plastica ed indossava un cappello da pirata. La foto era intitolata «Drunken Pirate».

L'immagine della piratessa ubriaca finì sotto gli occhi del preside della scuola, il quale accusò Stacy di promuovere il consumo di alcol tra gli alunni minorenni. L'Università le negò così la specializzazione e di conseguenza l'abilitazione all'insegnamento.

Questo episodio insegna che mentre una festa e i venticinque anni sono il luogo e il momento giusto per bere e indossare cappelli da pirata, il Web è il luogo sbagliato e, in un certo senso, anche il momento sbagliato. È il luogo sbagliato perché, mentre alla festa ci sono amici e altre persone con la stessa intenzione di bere e indossare cappelli carnevaleschi, sul Web ci sono tutti, compreso il preside della scuola in cui si lavora e il rettore dell'università nella quale si studia. È il momento sbagliato perché sul Web il tempo non esiste, tutto ciò che avviene rimane cristallizzato e persiste per sempre, ogni momento che passa non passa realmente, ma si accumula ai momenti precedenti.

Una foto scattata e pubblicata nel pieno diritto di libera manifestazione di pensiero, che ritrae un'azione del tutto legale e commessa al di fuori dell'orario di servizio ha compromesso per sempre la carriera di Stacy. Tutto ciò perché il Web era il posto sbagliato dove pubblicarla e lei, come tantissimi altri, non l'aveva ancora capito.

Le informazioni che si trovano sulla Rete hanno un'influenza fortissima sulla nostra vita reale, maggiore di quella che avrebbero se fossero rimaste semplicemente *offline*. Il grado di influenza che noi possiamo avere sulle informazioni *online* invece non è sempre garantito: a volte rimuoverle o modificarle è un attimo, a volte impossibile e altrettanto impossibile impedire che si diffondano nella Rete a macchia d'olio.

È bastata una foto su Internet per far perdere il lavoro a Stacy, non sono bastati quattro anni di procedure di ricorso per farglielo riavere. Anzi, l'eco di questa vicenda è risuonata sulla stampa, la saggistica e, a distanza di undici anni dall'episodio, anche su queste pagine.

In realtà, siamo tutti pienamente consapevoli di quanto il Web racconti di noi. Tuttavia, proprio come mia nonna, vediamo le potenzialità del Web

sempre sotto l'ottica del ficcanaso e mai dell'osservato. In altre parole, cerchiamo amici, conoscenti e colleghi su Google e su Facebook, ma non sempre pensiamo che loro potrebbero fare altrettanto con noi.

Produrre e condividere informazioni oggi è diventato economico e banale: questo ci ha spinti ad usarle con superficialità. Così, molti fanno un uso avventato e poco ponderato della Rete, scambiando un profilo Facebook con un diario personale o pubblicando *tweet* come se fossero battute dette all'interno del salotto di casa propria. Pensieri, foto, video vengono lasciati, a volte in dosi massicce, sui propri profili *social* credendo che siano di natura tanto effimera quanto le chiacchiere e le risate fatte al bar con gli amici, senza pensare che il Web non è scritto a matita, ma con il pennarello indelebile.

Tanto è facile produrre e condividere informazioni, tanto è difficile rimuoverle da dove sono state archiviate e rese pubbliche. Da un lato diamo scarsissimo valore a ciò che si pubblica in rete, dall'altro attribuiamo un altissimo grado di credibilità a ciò che vi troviamo sopra. Questi sono i paradossi che costituiscono un grande pericolo per l'identità digitale.

L'inconsapevole tentativo di compromettere la propria immagine sul Web attraverso la pubblicazione impropria, o talvolta semplicemente poco lungimirante, di contenuti personali è un fenomeno diffusissimo, ma in un certo modo arginabile. Basta un poco di buon senso e magari l'assimilazione di un "galateo" dell'internauta per evitare di autoarrecarsi danni permanenti. Dopotutto, i canali attraverso i quali vengono condivisi pensieri e foto personali sono in buona parte sotto il controllo diretto di chi li pubblica. Basta fare un uso responsabile e consapevole delle piattaforme sociali e stare attenti a ciò che si condivide per potersi tutelare: per intendersi, niente pirati ubriachi sul proprio profilo Facebook. Tutto ciò che si è pubblicato, si può nella maggior parte dei casi rimuovere, purché lo si faccia in tempo e che il contenuto non sia così sconcertante da scatenare fenomeni virali e perderne per sempre il controllo.

Se già il danno che possiamo crearci da soli costituisce un pericolo concreto e per nulla trascurabile, ciò che deve destare più preoccupazione è ciò che gli altri possono dire di noi.

L'identità digitale non si compone di sole informazioni emesse da noi stessi, ma anche di quelle prodotte e condivise da terzi. Se ciò che ha detto qualcuno sul nostro conto non ci piace o, peggio, lede la nostra immagine pubblica, bisogna far fronte a un complesso e fumoso equilibrio tra il nostro

diritto alla *privacy* e il suo diritto di libera manifestazione di pensiero e di cronaca, prima di poter rivendicare la rimozione di quel contenuto.

In particolare, ognuno di noi ha una pagina che non abbiamo creato noi e nessun altro: è la pagina che compare cercando il nostro nome e cognome su Google. Si tratta di un vero e proprio profilo personale, alla stregua di Facebook e LinkedIn: contiene foto, video, informazioni sulla vita privata, pubblica e professionale. Questa pagina può essere vista da chiunque ed è probabilmente il primo profilo in cui si imbatterà colui vorrà cercare qualche informazione sul nostro conto.

Il motore di ricerca restituisce all'utente che esegue la ricerca un elenco di risultati associati al nome di una persona fornendo in questo modo una visione complessiva strutturata delle informazione relative al soggetto sul Web. Informazioni che coprono una vasta molteplicità di aspetti legati alla vita privata e pubblica di quella persona che, senza l'azione di un motore di ricerca, potrebbero non risultare così correlate agli occhi dell'utente.

La scelta dei contenuti più in vista, la loro disposizione nel *layout* e la loro gerarchia non si basano su criteri di pertinenza, cronologia, distinzione tra sfera pubblica e privata, ma sui criteri di *ranking* del motore di ricerca che tratta il nostro nome e cognome, e di conseguenza la nostra identità, come una *keyword* qualunque.

Quando qualcuno decide di fare una ricerca su di noi nel Web, Google è il primo a prendere la parola, a scegliere i contenuti che dovranno essere visti per primi e in quale ordine. Google parla al posto nostro e potremmo anche non essere d'accordo. C'è un modo di reagire, c'è un modo per assumere il controllo di ciò che Google dice di noi? Da questo spunto nasce e si sviluppa la seguente ricerca.

L'Economia della Reputazione

Oggi la reputazione è denaro. Il fatto che le compagnie assicurative abbiano iniziato a calcolare le polizze in base ai profili Facebook[2] dei propri clienti dimostra che quest'affermazione è un dato di fatto e non un modo di dire.

La reputazione è esattamente come una moneta, il suo valore nel tempo è determinato dai termini con i quali può essere scambiata: una buona repu-

[2]Ric Romero, *Are Insurance Companies Spying on Your Facebook Page?*, abc/.com, 7 novembre 2010, http://abc7.com/archive/8422388/, 19/4/2017.

tazione ci consente di accedere a relazioni di valore, professionali e non; con una reputazione cattiva non potremo "permetterci" determinate opportunità; una reputazione falsa durerà giusto il tempo prima di essere scoperta e invalidata; una reputazione nulla ci escluderà semplicemente dal mercato.

Siamo all'alba della Reputation Economy, termine coniato da Michael Fertik nell'omonimo saggio per indicare

> un mondo in cui la reputazione di ognuno è istantaneamente analizzata, archiviata e utilizzata come passaporto per speciali trattamenti e benefici. Nell'Economia della Reputazione, potremo usare la nostra reputazione come contante, come garanzia dei debiti e per effettuare operazioni altrimenti impossibili.[3]

Nel futuro immaginato dall'autore canadese Cory Doctorow nel suo best seller *Down and Out in the Magic Kingdom*[4], l'economia gira intorno ad un nuovo tipo di valuta: il Whuffie. Il Whuffie è l'indicatore del capitale sociale di una persona e chiunque, quando incontra qualcuno, può immediatamente saperne il suo valore. Ciascuno può incrementare il suo Whuffie in tre modi: compiendo buone azioni, avendo una buona rete di relazioni oppure diventando popolare e famoso. Nel momento in cui però compie una cattiva azione, si lamenta troppo o abbandona la vita sociale, il valore del Whuffie cala.

L'idea di poter visualizzare in tempo reale la reputazione di chiunque non è poi così lontana dalla realtà: basta avere a portata di mano uno *smartphone* e conoscere il nome della persona che si ha davanti per fare una rapida ricerca su Google. Secondo un breve saggio del 2003 della società Risk2Reputation [5], nel mondo degli affari il 75% del valore di mercato di un'attività dipende dalla sua reputazione, mentre nella vita privata questo dato può avvicinarsi al 100%.

Nel mondo del lavoro, così come nella vita sociale, succede ogni giorno di incontrare persone che non conosciamo e delle quali abbiamo bisogno di informazioni. Se prima ci affidavamo al passaparola dei conoscenti, a lettere di raccomandazione e magari fidarsi della persona stessa, oggi abbiamo un

[3] M. Fertik, *Reputation Economy. Come ottimizzare il capitale delle nostre impronte digitali*, Egea, Milano, 2015, p. 6.

[4] C. Doctorow, *Down and Out in the Magic Kingdom*, St. Martins's Press, New York, 2003.

[5] J. Ryner, *Managing Reputational Risk. Managing Reputational Risk Leveraging opportunities, Curbing Threats*, John Wiley & Sons, Indianapolis, 2003.

mezzo immediato, sicuro e gratuito: Google. In pochi secondi, possiamo collegarci alla Rete, fare una veloce ricerca e farsi immediatamente un'idea della storia di un individuo o di un'azienda sul Web.

Avere indiscriminato accesso alle informazioni di chiunque e di qualunque cosa è un'arma a doppio taglio: da una parte il Web è un preziosissimo strumento a nostra disposizione che ci consente di accedere a un'illimitata fonte di informazioni, dall'altra la sua caratteristica di conservare qualsiasi contenuto venga inserito e di restituirlo ogni qual volta viene richiesto, può rivangare antichi errori, indiscrezioni e comportamenti imbarazzanti. Le conseguenze negative di questi errori possono perseguitarci per anni, almeno che non si adotti un provvedimento proattivo nei confronti delle nostre informazioni che circolano sulla Rete.

Da quando Google è diventato il motore di ricerca più usato al mondo, il problema si è esacerbato. Mentre il gigante della ricerca migliora e affina i suoi algoritmi, le informazioni vengono alla luce a una velocità mai pensata prima.

Le aziende ora usano queste informazioni per valutare qualsiasi cosa, indiscriminatamente prodotti e persone, per prendere le loro decisioni. Se una persona, un prodotto o una società ha un passato torbido, è probabile che lo scoprano nel giro di pochi attimi. Non importa quale decisione sta per essere presa, i dati forniti da una esaustiva ricerca su Google possono influenzare drasticamente i risultati di questa decisione.

Una recente indagine di Doxa 2.0 Research[6] sostiene che tramite Google gli internauti raccolgono la maggior parte delle informazioni che riguardano individui, per motivi sia personali sia professionali. Il 62% fa ricerche su una persona con cui deve collaborare in ambito lavorativo e oltre il 70% esegue indagini su Google prima di acquistare un prodotto o servizio.

Nel mondo dell'economia il valore della reputazione è già molto chiaro, specialmente in ambito aziendale. Si è sviluppato infatti nella teoria economica il concetto di *capitale reputazionale*, ossia l'insieme dei valori e comportamenti sociali che influenzano il potere contrattuale dell'individuo o dell'azienda.

Nell'era del Web, ognuno di noi è un *brand* e dunque ognuno di noi possiede un capitale reputazionale da tutelare e su cui investire per potersi

[6]Studio riportato in A. Agostini, *La tua reputazione su Google e Social Media*, Hoepli, Milano, 2013, pp. 8 e ss.

garantire le giuste possibilità di intrecciare relazioni professionali e personali sia sulla Rete ma anche, e sopratutto, al di fuori di essa.

L'Economia della Reputazione è massima espressione della Società dell'Informazione in cui viviamo, una fase della Storia nella quale l'informazione ha più valore del bene materiale: il corredo informativo sul nostro conto (che oggi si trova soprattutto sul Web) può assumere un valore maggiore delle nostre effettive qualità, al punto di poterle invalidare.

Identity cleaning e diritto all'oblio, tra mito e realtà

Quando ho iniziato la mia ricerca sulla tutela della reputazione *online*, ero molto incuriosito da un termine: *identity cleaner*. L'*identity cleaner* è una figura professionale che si occupa proprio di ciò che il suo nome lascia intendere, cioè di ripulire le identità *online*.

Ripulire un'identità fa pensare che si possa in qualche modo cancellare, o per lo meno rendere non più rintracciabili, le informazioni scomode relative ad un soggetto. In ambito Web, questo si tradurrebbe nell'effettiva facoltà di rimuovere pagine *web* o parte di esse, o trovare il modo di deindicizzare risorse dagli archivi di Google in modo tale che non compaiano più nei risultati di ricerca.

L'idea è affascinante, ancor più se si accosta ad un altro termine: diritto all'oblio. Il diritto di essere dimenticati, di poter ritirare dalla sfera pubblica informazioni non desiderate è un tema che si sta facendo largo nel dibattito giuridico, in particolare dopo una celebre sentenza del maggio del 2014 in cui la Corte Europea lo ha annoverato tra i diritti della personalità e ha imposto a tutti i motori di ricerca di dare la possibilità a qualsiasi utente di fare richiesta di rimozione di pagine lesive della propria reputazione.

La mia indagine è quindi partita dalla ricerca delle effettive possibilità di rimuovere informazioni dal Web, al fine di tutelare il "diritto di essere dimenticati" delle persone.

Il campo di studio è indubbiamente intrigante, ma purtroppo si è rivelato molto presto sterile.

Rimuovere contenuti che non siano di nostra proprietà è semplicemente impossibile, per ovvi motivi. Accedere clandestinamente ad un *server* per cancellare o modificare documenti è un atto illegale: sicuramente fattibile per un bravo *hacker*, ma non certo da prendere in considerazione se il fine

della rimozione è la tutela di un diritto della persona. Il solo modo per poter eliminare un'informazione alla sua fonte è convincere il *webmaster* a farlo e questo può essere fatto in due modi: in via amichevole o in via legale.

La prima strategia non costa nulla in termini di denaro e di tempo, ma non offre grandi garanzie di successo. La seconda può arrivare a costare molto, sia come spesa in avvocati che come tempi della procedura. Quanto alle aspettative di successo, la questione è molto complessa: almeno che il caso non ricada in un chiaro reato di diffamazione o di violazione della *privacy*, tutto rimane nell'area nebbiosa e dai confini incerti del diritto all'oblio, un concetto che da un lato è ancora poco chiaro in ambito giuridico, mentre dall'altro pesta i piedi a molti altri diritti quali diritto di cronaca, manifestazione di libero pensiero, diritto d'informazione.

La teoria sul diritto all'oblio si è sviluppato di pari passo con la diffusione dei mezzi di comunicazione di massa e soprattuto con l'avvento di Internet. I primi casi di riconoscimento del diritto a essere dimenticati si sono manifestati in ambito giornalistico, dove ancor prima dell'era digitale si era creato uno squilibrio tra diritto di cronaca e diritto alla riservatezza.

Nella giurisprudenza italiana, la prima formulazione del diritto all'oblio è contenuta in una sentenza della Cassazione del 1998, come

> (...) giusto interesse di ogni persona a non restare indeterminatamente esposta ai danni ulteriori che arreca al suo onore e alla sua reputazione la reiterata pubblicazione di una notizia in passato legittimamente divulgata.[7]

Con l'arrivo di Internet, il problema non è più quello della ripubblicazione di una notizia, ma la sua persistenza sulle piattaforme *web*, costantemente accessibili e dell'altissimo rischio che questa venga riesumata da un motore di ricerca in situazioni del tutto fuori contesto. Se letteralmente diritto all'oblio significa diritto ad essere dimenticati, oggi si dovrebbe parlare di "diritto alla contestualizzazione": il problema non è cancellare il proprio passato, ma contestualizzarlo[8].

Nel 2014 la Corte di Giustizia Europea si trovò ad esaminare il caso di Mario Costeja Gonzales, il quale aveva citato in giudizio Google Spain rivendicando il diritto di richiedere la rimozione di alcuni *link*, risalenti a sedici

[7] Cass. civ. Sez. III, 09/04/1998, n. 3679.

[8] G. Finocchiaro, *Il diritto all'oblio nel quadro dei diritti della personalità*, "Il diritto dell'informazione e dell'informatica", XXIX, 4 maggio 2014, pp. 591 - 604.

anni prima, relativi alla messa all'asta della sua casa per difficoltà economiche. Secondo Gonzales, quei *link* riconducevano a contenuti che violavano la sua *privacy* e non corrispondevano più alla sua situazione patrimoniale.

Con sentenza del 13 maggio 2014 la Corte Europea stabilì che i cittadini europei hanno il diritto di richiedere al motore di ricerca la rimozione di informazioni associate al proprio nome qualora queste risultino «inadeguate, non pertinenti o non più pertinenti, ovvero eccessive in rapporto alle finalità del trattamento in questione realizzato dal motore di ricerca»[9]

Nelle condizioni sopra descritte, il motore di ricerca è obbligato alla rimozione dei risultati di ricerca che rimandano a contenuti lesivi, anche se questi non sono stati rimossi dai siti originari. La vera novità di questa sentenza è che la responsabilità sul trattamento dei dati personali non è attribuita soltanto ai proprietari dei siti che pubblicano i contenuti, ma si estende anche al motore di ricerca.

Google non accolse bene l'esito la decisione della Corte Europea. D'altro canto, il suo *business* si basa sulla diffusione di informazioni e dare agli utenti il potere di intromettersi nelle scelte del motore di ricerca costituiva un notevole colpo alla sua autorità. Tuttavia, non potendo ignorare la sentenza, nei giorni successivi mise a disposizione un modulo di "Richiesta di rimozione di risultati di ricerca ai sensi della legge europea per la protezione dei dati personali", che consente di fare richiesta di rimozione di *link* che non violano le linee guida di Google (come violazione di *copyright*, contenuti pedopornografici o comunque dichiarati non leciti da un tribunale, per i quali già esistevano *form* appositi).

La sentenza del 13 maggio 2014 e il conseguente modulo di rimozione di Google diede l'illusione che il diritto all'oblio fosse qualcosa di concreto e rivendicabile in via del tutto autonoma da parte di chiunque. Il comune cittadino poteva, comodamente dal divano di casa propria, decidere quali contenuti a lui riferiti sarebbero rimasti negli indici di Google e quali invece dovevano essere rimossi inviando una semplice comunicazione al *team* di Google. La realtà dei fatti non è stata ovviamente all'altezza delle aspettative.

Secondo un'inchiesta del New York Times[10] dell'aprile 2016, la sentenza

[9]Corte di Giustizia dell'Unione europea, 12/5/2017, n. C-131-12.

[10]Mark Scott, *Europe Tried to Rein In Google. It Backfired*, "New York Times", 18 aprile 2016, https://www.nytimes.com/2016/04/19/technology/google-europe-privacy-watchdog.html, 19/4/2017.

del 13 maggio 2014 ha scatenato un effetto boomerang: delle 418.000 richieste di rimozione ricevute negli ultimi due anni (si parla dei dati registrati ad aprile 2016), meno della metà sono state accolte ed è Google, e soltanto lui, ad aver sentenziato quali contenuti fosse giusto rimuovere e quali no. La ragione sta nel fatto che la Corte Europea, dopo aver stabilito che i motori di ricerca devono rispettare il diritto all'oblio, ha delegato ai motori di ricerca stessi l'onere di decidere in quali casi questo viene violato. Google dal canto suo, del tutto riluttante di fronte alle imposizioni dell'Europa, fa il suo gioco in totale libertà.

«Il nostro obiettivo è sempre stato la libertà di espressione e non ritengo che Google possa indossare le vesti del giudice nel determinare quali contenuti debbano essere rimossi» — ha dichiarato Peter Barron, portavoce di Google, concludendo che tuttavia non avrebbero potuto fare altro che adeguarsi alla legge. Il problema è che, a partire da questa "rivoluzione europea" in termini di tutela del diritto all'oblio, è successo proprio questo: Google è diventato giudice di se stesso.

L'oblio, in poche parole, non funziona. Prima di tutto perché non sono ancora ben chiari i criteri per definire il grado di obsolescenza di una notizia tale che la sua rimozione non vada contro il diritto di cronaca; in secondo luogo, perché almeno che non sia un tribunale a stabilirlo, la rimozione di un contenuto è a discrezione del *content* o *service provider* (come per esempio Google); terzo, perché la Rete non dimentica, non può farlo per sua stessa natura.

Tutela della reputazione sui motori di ricerca: strategia sottrattiva e strategia additiva

Tentare di rimuovere o deindicizzare contenuti dal Web significa curare l'identità digitale agendo per sottrazione. Vanno tuttavia considerati i rischi che possono nascere da un approccio di questo tipo.

Certamente, riuscire a rimuovere un risultato negativo dalla pagina dei risultati di Google risolve buona parte del problema in caso di reputazione danneggiata. È già chiaro però come le aspettative di successo di questa operazione siano davvero minime.

Anche nell'eventualità di successo, potrebbero verificarsi effetti collaterali molto pericolosi.

La SERP di Google (cioè la pagina dei risultati di ricerca) va considerata come una pila di oggetti che vengono spinti verso l'alto. Se se ne sottrae uno, quello sotto prenderà il suo posto. Se l'oggetto che sostituisce quello rimosso è un contenuto positivo nessun problema, se è negativo o in qualche modo indesiderato il processo di pulizia torna al punto di partenza.

Sottrarre uno o più risultati dalla SERP significa lasciare dei buchi che Google troverà il modo di riempire. Se non si è effettuato un programma di monitoraggio preliminare su tutto ciò che è indicizzato per la *query* alla quale la SERP corrisponde e non si sono valutati bene i rischi, l'identità digitale rimane esposta ad attacchi inaspettati.

Inoltre, va considerato che quando un risultato viene rimosso, per ingiunzione del tribunale, per accettazione del modulo di richiesta o per qualsiasi altra ragione, il proprietario del contenuto riceve una notifica. Ciò comporta il rischio di ritorsioni da parte di chi ha emesso la notizia con la pubblicazione di altri contenuti, col risultato di portare i riflettori proprio su ciò che si voleva celare.

Quando si parla di far valere il proprio diritto all'oblio, va sempre tenuto in mente qual'è stata la sorte del cosiddetto "padre dell'oblio": Costeja Gonzales, colui che vinse la causa contro Google (sentenza del 13 maggio 2014) e ottene, per se stesso e per tutti i cittadini d'Europa, il diritto di essere dimenticato, è oggi ricordato da tutti proprio per il successo della sua battaglia ma, di conseguenza, per tutte le ragioni per cui l'aveva intrapresa.

In seguito alla celebre sentenza, annoverata come uno dei punti di svolta nella storia del diritto dell'informazione, nacquero moltissimi articoli sulla vicenda che ne ripercorrevano tutte le fasi e che, naturalmente, riportavano nel dettaglio i fatti compromettenti a cui Costeja voleva non essere più associato. Pochi *link* erano stati rimossi con successo in seguito alla sentenza e in seguito alla sentenza magliaia di nuovi *link* si erano creati.

Costeja chiese nuovamente a Google di deindicizzare i nuovi articoli, Google rifiutò. Allora si rivolse all'Autorità spagnola per la tutela della *privacy* perché ordinasse al motore di rimuovere i *link* dai suoi archivi, ma questa volta l'autorità spagnola appoggiò le ragioni di Google. Le notizie relative a Mario Costeja Gonzales erano tutt'altro che obsolete, bensì al centro di un dibattito non solo attuale ma quasi rivoluzionario: il diritto di riservatezza doveva cedere il passo al diritto di cronaca.

Oggi, se si cerca "Mario Costeja Gonzales" su Google, ci si stanca molto

presto di contare quanti contenuti fanno riferimento al caso "Google Spain" e ai motivi perché Costeja aveva citato Google in giudizio. Mezzo mondo ora sa che la casa di Mario Costeja Gonzales era stata messa all'asta per i suoi ingenti problemi economici. Verrebbe da chiedersi quanti milioni di persone in meno lo saprebbero se solo quella sentenza non ci fosse mai stata.

Portare all'estremo l'applicazione del metodo sottrattivo comporta la totale cancellazione di ogni traccia di sé dal Web. Ancora una volta, l'idea è affascinante. Un'identità digitale nulla significa nullificare il potere che questa può esercitare sulla nostra vita reale.

Tuttavia, "uccidere" il proprio Io *online* potrebbe essere molto rischioso. A un'identità digitale di grado zero corrisponde una vulnerabilità ad eventuali crisi reputazionali di grado massimo.

Ancora una volta, la metafora della pila di oggetti ci fa capire il problema: se la pila è vuota, resta il fatto che la spinta verso l'alto rimane. Come viene inserito un oggetto nella pila, questo verrà spinto immediatamente in prima posizione. Se quell'oggetto è in contenuto è negativo, quello sarà l'unico biglietto da visita per chi vorrà incontrarci sul Web.

Tutte queste considerazioni non vogliono portare a concludere che l'approccio sottrattivo alla tutela dell'identità digitale sia sbagliato. Semplicemente è un metodo che, oltre ad offrire basse probabilità di successo, è molto poco controllabile almeno che non venga controbilanciato da un'efficace strategia di addizione.

Anziché togliere oggetti dalla pila, se se ne aggiungono il più possibile nelle prime posizioni, gli oggetti indesiderati scaleranno verso il basso. Questo si traduce, all'atto pratico, che se si è in grado di produrre contenuti efficacemente ottimizzati secondo i criteri di *ranking* di Google, questi andranno (con tempo e pazienza) ad occupare le prime posizioni della SERP, costringendo i risultati più vecchi a scendere verso il basso diventando sempre meno visibili.

Allo stesso modo, se nel frattempo si riesce a rimuovere alcuni contenuti attraverso l'approccio sottrattivo, questi potranno essere sostituiti dai contenuti positivi e ottimizzati anziché da altri su cui non si ha il controllo.

Sono due gli aspetti più importanti da considerare sull'influenza che un risultato di Google e del contenuto a cui si riferisce hanno sulla reputazione di una persona: la visibilità e l'accessibilità.

Mentre la strategia sottrattiva mira a colpire l'accessibilità di un con-

tenuto, cercando di rimuovere il *link* dalla SERP di Google o addirittura rimuovere la pagina stessa, quella additiva mira a colpirne la visibilità: rendere risultati negativi meno visibili promuovendo la visibilità di contenuti positivi.

Tutto si riassume in un detto comune di chi si occupa di *web marketing*: «Il posto migliore dove nascondere un cadavere è la seconda pagina di Google»

La maggior parte delle persone infatti non considera, o considera come meno rilevante, tutti i risultati dopo il terzo della prima pagina. Il 91% degli utenti non approda alla seconda pagina di Google e più del 97% non va oltre la terza. Il che significa che un *link* negativo nella seconda pagina di Google ha un potenziale lesivo ridotto del 91% rispetto a uno in prima pagina.

Se è così difficile rimuovere contenuti dal Web, soprattutto se non creati da noi, è però molto facile produrli e, grazie ai criteri che ci provengono dal *web marketing* (SEO), è possibile farli "piacere" a Google in modo che questi compaiano nelle prime posizioni della SERP, facendo affondare altri *link*, magari proprio quelli indesiderati.

Il concetto di strategia reputazionale additiva si basa sulla produzione e ottimizzazione di contenuti sui quali si esercita diretto controllo, in modo tale da assumere, di conseguenza, il controllo sulla SERP. Viene da sé che qui non si tratta solo di limitarsi a rimuovere o nascondere un contenuto, ma costruirsi una solida identità digitale, in linea con le nostre aspettative. Il metodo additivo è quindi applicabile, ed è più che consigliabile farlo, prima di una qualsiasi crisi reputazionale.

Questa strategia è stata testata su un caso reale. Si trattava di un soggetto che voleva rimuovere un risultato che compariva in quarta posizione sulla SERP quando si cercava il suo nome e cognome su Google.

Fallito il tentativo di agire per sottrazione inviando il modulo richiesta a Google, la SERP è stata tenuta sotto osservazione per circa 10 settimane. In questo periodo di tempo è stata incrementata l'attività *social* (escludendo Facebook, considerata una piattaforma per uso troppo personale), sono stati creati nuovi profili professionali, si è eseguita l'ottimizzazione di pagine di alcuni *blog* per i quali il soggetto scriveva e di alcune fotografie secondo i criteri usati nella *Search Engine Optimization* (SEO).

L'esito è stato che nel giro di poco più di un mese il risultato indesiderato è passato dalla quarta posizione alla dodicesima, il che significa nella seconda

pagina della SERP. Il cadavere era stato seppellito.

Trattandosi di un caso di studio piuttosto contenuto, che non gode del confronto con altri casi simili e che soffre dell'incertezza di metodo che sempre accompagna chi si occupa di ottimizzazione sui motori di ricerca, non può avere un solido valore scientifico. Resta il fatto che al primo tentativo di applicare la strategia additiva a un caso reale di reputazione compromessa il successo è stato totale.

La progettazione di un *crawler* per il monitoraggio della SERP di Google

Durante la fase di osservazione della SERP relativa al soggetto preso in esame, era necessario uno strumento in grado di rilevare e registrare giornalmente la posizione e il contenuto dei primi trenta risultati che venivano restituiti, in modo tale da poter valutare le fluttuazioni alle quali erano soggetti e verificare se le operazioni di correzione stavano avendo il loro effetto.

In particolare, era necessario che almeno il titolo, la URL e la posizione di ogni risultato fosse intabellato in un *file* leggibile da un foglio di calcolo come Microsoft Excel in modo tale che i dati potessero essere comodamente analizzati e confrontati una volta conclusa la fase di sperimentazione. Pensare di effettuare manualmente il monitoraggio ogni giorno e registrare uno per uno i dati necessari per ogni risultato in una tabella Excel era impensabile.

Il tipo di programma adatto per questo tipo di attività era un *web scraper*, un *crawler* in grado di visitare pagine *web* e raccogliere automaticamente i dati necessari.

Insoddisfatto dei programmi già resi disponibili in forma gratuita o a pagamento, ho deciso di svilupparne uno da zero, sfruttando alcune librerie in linguaggio Python che consentivano di ottenere risorse *web* tramite richieste HTTP e di parsificare e navigare il loro contenuto.

Lo *scraper* avrebbe dovuto ottenere il codice HTML della SERP di una determinata *search query*, visitare il documento una volta parsificato e copiare le informazioni richieste in un *file* csv pronto per essere usato in uno *spreadsheet*.

Per individuare l'esatta posizione dei dati da raccogliere sulla SERP, è stato necessario una lunga e accurata ispezione del suo codice, che mi ha portato a capire e schematizzare con precisione la sua struttura HTML.

Presa confidenza con le librerie in Python e con il *markup* della SERP, ho pensato di non limitarmi ad ottenere soltanto i dati necessari per monitorare gli spostamenti dei risultati all'interno della SERP, ma di raccogliere tutto quello che sarebbe potuto servire per l'analisi e il monitoraggio della reputazione sul motore di ricerca.

La parola d'ordine del metodo additivo infatti è controllo. Ogni operazione deve mirare ad assumere il controllo della SERP utilizzando canali sotto la propria gestione, in modo tale che sia possibile influenzare ciò che sulla SERP deve comparire.

Il monitoraggio è la prima, necessaria, fase per assumere progressivamente il controllo della SERP. È inoltre fondamentale durante l'attività di costruzione o riparazione della *personal reputation* e anche dopo, per verificare che le modifiche che si sono riuscite ad applicare manifestino una sufficiente stabilità.

L'idea non è stata quella di creare uno strumento di analisi: stabilire dei criteri universali di valutazione della reputazione va incontro a diverse difficoltà e rischi. La reputazione infatti si gioca spesso su conclusioni intuitive da parte del giudicante nel momento in cui associa un soggetto ad un determinato contesto. Cogliere questi segnali automaticamente (cioè attraverso istruzioni eseguibili da un programma) è molto difficile, così come algoritmi di analisi testuale come la *sentiment analysis* possono svolgere un ruolo soltanto parziale all'interno della valutazione della reputazione e soltanto nel caso in cui il soggetto sia al centro di un effettivo ed ampio dibattito sulla rete.

Lo strumento che ho voluto creare è un "semplice" collettore di dati, estrapolati dal codice HTML, che ho pensato essere rilevanti per una serie di approcci all'analisi della reputazione, al fine di ottenere una serie di informazioni neutre, ossia non interpretate, da utilizzare come più si ritiene utile per le proprie necessità.

In primo luogo, è stato necessario individuare una serie di aspetti della reputazione online in base ai quali stabilire quali dati estraibili dalla SERP fossero rilevanti per l'analisi della reputazione.

I primi due aspetti principali, già citati, sono visibilità e accessibilità.

L'influenza che un'informazione può esercitare sulla reputazione dipende dalla sua visibilità. L'esempio più chiaro è quello della posizione di un risultato nella SERP: le prime posizioni, oltre ad essere viste dalla stragrande

maggior parte degli utenti, sono anche quelle considerate le più autorevoli ed aggiornate. D'altra parte, per quanto oggi le notizie vengano lette per lo più su testate *online*, siamo ancora legati ad un concetto cartaceo di presentazione dell'informazione: ciò che sta in prima pagina è sicuramente la notizia più interessante, autorevole e, sopratutto, aggiornata.

L'accessibilità stabilisce quanto un contenuto può essere effettivamente raggiunto e letto da un utente. Per certi versi, l'accessibilità è subordinata alla visibilità, in quanto un contenuto poco visibile avrà poche probabilità di essere raggiunto. La cosa invece non è vera al contrario: è possibile infatti che sulla SERP alcuni contenuti siano visibili ma non accessibili. È il caso di risorse obsolete non deindicizzate, ossia pagine *web* che sono state cancellate o non disponibili (quelle che danno il celebre errore 404) che non sono state ancora rimosse dagli indici di Google e che quindi vengono ancora restituite sulla SERP.

Un altro aspetto molto importante è il contesto in cui il contenuto è immerso o esso stesso crea. In questo caso, trovare fattori quantificabili è molto difficile e si deve far fronte ad un'eccessiva quantità di variabili dalla definibilità incerta. Per farsi comunque un'idea di quali sono i contesti in cui il nome e cognome del soggetto compare, è stato pensato di valutare la relazione tra *query* di ricerca e altre parole presenti nel corpo di testo delle pagine a cui i risultati sulla SERP si riferiscono. Per dirlo in parole pratiche: se cerco "mario rossi" su Google, quante probabilità ho di imbattermi nella parola "ergastolo" leggendo le pagine indicate dai primi dieci *link* che Google mi ha ristituito?

La possibilità di trovare informazioni utili mentre se ne stavano cercando altre è un altro elemento molto importante per valutare i rischi a cui un'identità digitale è soggetta. Alcuni servizi di assistenza alla ricerca di Google come Autocomplete (le *query* suggerite che compaiono mentre si sta scrivendo nella barra di ricerca) e Ricerche Correlate forniscono suggerimenti di ricerca che aggiungono altre parole chiave alla *query* originale oppure mostrano ricerche simili frequentemente eseguite da altri utenti.

Questo fattore che può influenzare una ricerca si chiama serendipità e può costituire un notevole pericolo per la reputazione personale. Può succedere infatti che mentre un utente, totalmente ignaro che esiste un Mario Rossi terrorista, digiti semplicemente "mario rossi" nella *search query box* di Google e immediatamente gli venga suggerito "mario rossi terrorista". Nessu-

no ovviamente penserà male di Mario Rossi, dal momento che è il nome più comune d'Italia e usato qui solo per semplificare, tuttavia, sono tanti i casi, di cui molti anche in Italia, in cui alcuni individui hanno accusato di diffamazione Google perché associava nelle Ricerche Correlate e in Autocomplete il loro nome a parole come "truffa", "bancarotta", "setta" ecc.

Un altro elemento da considerare nella valutazione di un contenuto in riferimento alla reputazione di un soggetto è la sua età. Sapere in quale data è stato pubblicato un contenuto, o perlomeno quando è stato indicizzato da Google, significa poter valutare l'"età" di un'identità digitale, nonché l'eventuale grado di obsolescenza di una notizia in relazione al diritto di cronaca e quello all'oblio.

Infine, è necessario individuare il responsabile del contenuto: l'autore e il proprietario del sito che lo pubblica

Una volta stabiliti questi aspetti, sono stati individuati tutti i dati effettivamente estraibili dall'HTML sia della SERP, sia delle risorse a cui i suoi risultati si riferiscono, quali *tag* HTML li contenevano in modo tale da istruire il programma a individuarli all'interno del *markup* e copiarne il contenuto in liste (gli *array* del linguaggio Python) che sarebbero state poi stampate in un *file* di testo in formato csv.

Ciò che ne è risultato è uno strumento, battezzato Scrappy per la semplice ragione che il *file* Python su cui è scritto è `scrap.py`, in grado di raccogliere in poco tempo una grande quantità di dati che, se utilizzati a dovere, offrono un'ampia possibilità di analisi dell'identità digitale, ma non solo.

Questo programma infatti è utilizzabile in molti altri campi, specialmente in quello SEO, per analizzare il posizionamento di pagine *web* in base a una determinata *query* o anche più di una (il programma infatti può eseguire la ricerca per un numero a piacimento di *query*, così come può verificare la presenza di quante parole si vuole all'interno di ogni documento).

Per quanto si tratti di un programma di poche righe di codice e dalla struttura rudimentale (non è ancora provvisto di un'interfaccia grafica) ha consentito di poter monitorare e successivamente analizzare le operazioni eseguite durante l'esperimento sull'applicazione della strategia additiva con facilità e precisione.

I primi due capitoli di questa relazione affronteranno il problema dell'identità digitale da un punto di vista concettuale e giuridico, i capitoli 3 e 4 entreranno nel merito dell'analisi della reputazione *online* che si concretizza,

in parte, nella progettazione del programma Scrappy. Negli ultimi due capitoli verranno riportati i risultati di applicazione delle strategie sottrattive e additive su casi reali.

Capitolo 1

L'Io nell'epoca della sua riproducibilità tecnica

1.1 La Società dell'Informazione

Sul finire degli anni '50 l'economista Fritz Machlup stimò che, tra il 1947 e il 1958, negli Stati Uniti le attività economiche legate alla conoscenza come educazione, ricerca, informatica e telecomunicazioni erano cresciute ad un tasso doppio rispetto al prodotto interno lordo dello stesso periodo. Era il primo segnale che qualcosa nella società era cambiato: un bene immateriale, l'informazione, stava assumendo un ruolo sempre più centrale nell'economia americana.

Le intuizioni di Machlup furono confermate solo un decennio più tardi, quando l'industria della conoscenza ricopriva il 40% del PIL degli Stati Uniti.

Nel 1973 Daniel Bell pubblicò *The Coming Of Post-industrial Society*[1], nel quale teorizzava una nuova fase dell'economia e della società, successiva a quella industriale, basata sulla produzione non più di beni materiali ma di servizi e beni informazionali. I dati economici confermarono, negli anni a seguire, la sua teoria: il "fattore informazione" divenne sempre più rilevante nell'economia americana (50% del PIL e dei salari derivavano dalla produzione, trattamento, e distribuzione di beni e servizi informazionali), mentre il progresso tecnologico dell'informatica e la telematica iniziarono ad accelerare.

[1] D. Bell, *The Coming Of Post-industrial Society: A Venture in Social Forecasting*, Basic Books, New York, 1973.

In *The Social Framework of Information Society*[2] parlò espressamente di una nuova Società dell'Informazione, in cui l'informazione, bene immateriale, prevale economicamente sui beni materiali come il lavoro e il capitale.

La conoscenza così si sostituisce al lavoro come fonte primaria del valore, con tutte le implicazioni di ordine sociale e politico che ne derivano. La proprietà privata dunque è destinata a perdere di importanza, soppiantata dal possesso della conoscenza e dal controllo delle informazioni.

1.2 La digitalizzazione dei contenuti

Il fatto che nell'*information society* l'informazione si sia stabilita alla base nell'economia e che sia diventata un nuovo indice di ricchezza, non implica che questo abbia influito anche sul suo costo in termini di risorse. Al grande valore che l'informazione ha assunto come bene immateriale, si contrappone infatti il crollo del suo prezzo come prodotto materiale. Grazie allo sviluppo delle tecnologie digitali produrre, diffondere e archiviare informazioni è diventato facile, economico e veloce. Nel corso degli ultimi decenni del secolo scorso, la tecnologia ha sancito il passaggio dall'era analogica a quella digitale, che ha modificato in maniera sostanziale il tipo di informazione che può essere memorizzato, il modo in cui viene memorizzato e il suo costo. I prezzi si sono abbassati sempre di più e le persone hanno cominciato a produrre e salvare contenuti in digitale.

Il passaggio dall'informazione analogica a quella digitale ha coinvolto gran parte, se non tutti, i settori economici e gli aspetti della nostra vita. La maggior parte dei contenuti che creiamo o di cui usufruiamo è in formato digitale: musica, video, foto, ma anche testi e opere d'arte figurativa oggi vengono prodotti, diffusi e archiviati attraverso apparecchiature digitali.

Nell'era analogica, l'elaborazione, la memorizzazione e il recupero erano diversi in base al tipo di informazione e necessitavano quindi di strumenti dedicati. I fotografi scattavano foto con macchine apposite, su pellicole adeguate e le portavano a sviluppare in laboratori specializzati. I documenti si scrivevano a mano o con una macchina da scrivere e tutti i materiali necessari (carta, inchiostro, penne ecc.) erano venduti da negozi per ufficio. Così funzionava anche per i video e per l'audio: per ogni tipo di contenuto che

[2]D. Bell, *The social framework of information society*, in T. Forest (a cura di), *Microelectronics revolution*, Oxford, 1980, pp. 501-549.

si voleva creare, c'era un dispositivo dedicato e tutto un mondo tecnologico, commerciale e professionale che girava intorno ad esso.

Nel mondo digitale, invece, tutto viene registrato allo stesso modo, cioè come stringhe di zero e di uno, e quindi tutti gli strumenti digitali, almeno in via teorica, sono in grado di farlo. Un'informazione digitalizzata può essere memorizzata su un supporto digitale indipendentemente dal fatto che si tratti di suoni, immagini, testi, immagini o qualsiasi altra cosa. La standardizzazione dell'informazione digitale offre ai produttori grandi opportunità, dal momento che gli strumenti digitali possono essere utilizzati per produrre e memorizzare qualsiasi tipo di contenuto. Essendo questi strumenti prodotti su larga scala, i loro prezzi unitari sono scesi sempre più drasticamente. La standardizzazione alimenta anche la domanda di strumenti tutto fare, in grado di elaborare testi, immagini, suoni e filmati. Esempio di questo tipo di dispositivi è lo *smartphone*, *device* formato tascabile da poche centinaia di euro che consente di fare e processare video e fotografie, scrivere, disegnare, collegarsi a Internet, ma anche cambiare canale della televisione e regolare il termostato dell'impianto di riscaldamento (oltre che telefonare)[3].

La standardizzazione presenta dei vantaggi anche a livello di condivisione e distribuzione delle informazioni. Mentre l'informazione analogica viene condivisa e distribuita su canali diversi, l'informazione digitale può viaggiare attraverso lo stesso *network*. Per questo mentre con lo *smartphone* si può leggere le ultime notizie del quotidiano preferito, vedere un film e ascoltare musica, nel dominio analogico per fare tutte e tre le cose occorre andare in edicola, accendere il televisore o andare al cinema, sintonizzare la radio o azionare il giradischi.

Produrre informazioni oggi è diventato facile, veloce ed economico. Basta un solo dispositivo per creare qualsiasi tipo di contenuto con la forza e agilità del proprio pollice. Non occorre più avere basi di tecnica fotografica per scattare foto di discreta qualità, la stessa cosa vale per i video. Ognuno di noi può scrivere sul Web e avere la stessa visibilità che una volta era esclusiva di giornalisti e scrittori, sui propri profili *social* o, per chi è più professionale, sul proprio *blog*. La produzione di tutti questi contenuti è a costo zero, o meglio, è tutto compreso nella spesa *una tantum* per il dispositivo che si usa

[3]Sugli effetti del multimediale e del *multitasking* sulla società, *cfr.* H. Jenkins, *Convergence culture: where old and new media collide*, New York University Press, New York, 2006.

per creare ognuno di essi. Non si parla solo di economicità, se non gratuità, in termini di denaro, ma tutti i tipi di risorsa coinvolti nella produzione di un contenuto sono minimizzati:

- spazio, perché i dispositivi sono così piccoli, leggeri e *wireless* che possono essere portati e usati ovunque,
- tempo, perché occorre solo quello impiegato per estrarre dalla tasca o dallo zaino il proprio dispositivo e il resto è questione di un *click* o due,
- forza, perché serve solo quella esercitata dal proprio pollice,
- esperienza e talento, perché le nuove tecnologie invogliano a improvvisarsi *video-maker*, fotografi, opinionisti e i risultati a volte superano di gran lunga le aspettative.

Si potrebbe andare avanti ancora per molto.

Oltre a produrle, oggi è economico anche condividere, pubblicare e diffondere informazioni. A partire dal Web 2.0 gli utenti possono interagire con le pagine *web*, creare e caricare contenuti e, quindi, condividerli. Tutto è iniziato con i *blog*, i *forum* e le *wiki* fino ad approdare ai *social network*, piattaforme su cui vengono ogni giorno caricati e condivisi milioni di contenuti di ogni genere.

La grande rivoluzione dell'era digitale è l'incredibile capacità di immagazzinare dati di ogni genere. Inizialmente i costi erano decisamente più importanti: negli anni Cinquanta un *megabyte* costava 70.000 dollari, oggi un *hard disk* da 500 *gigabyte* (500.000 *megabyte*) costa intorno ai 50 euro.

Nel momento che i contenuti sono pubblicati sul Web, questi sono automaticamente archiviati da qualche parte, in qualche *server*. Oggi anche l'archiviazione di informazione ha costi bassissimi, tanto che costa meno comprare altro spazio di archiviazione che svuotare quello già esistente cancellando contenuti obsoleti[4].

Il crollo dei prezzi dell'archiviazione di dati ha reso più conveniente conservare che cancellare. Se una grande azienda di servizi Internet dovesse decidere di svuotare i propri *server*, dovrebbe scegliere con cura i criteri con cui stabilire cosa cancellare o cosa no. Dopodiché occorrerebbe un programmatore o un ingegnere informatico per farlo. I *database* dovrebbero

[4]Fertik, *Reputation Economy*, cit., pp. 17 e ss.

essere idealmente progettati per consentire una facile rimozione di dati superflui, ma nella realtà dei fatti le banche dati sono state create nel corso del tempo con scarsa considerazione per le modalità di eventuale pulizia. Un programmatore esperto può costare cento dollari l'ora e pulire il *database* di un piccolo sito *web* può costare, nel migliori dei casi, almeno mille dollari[5]. Oggi con mille dollari un'azienda può comprare 20 *terabyte* di memoria su disco rigido. Basta pensare a come ogni anno, chiavette USB e *hard disk* esterni raddoppiano di capienza e dimezzano di prezzo per capire quanto l'archiviazione di dati stia diventando sempre più a buon mercato.

In sintesi, l'informazione analogica implica diversi costi: lo strumento con cui viene prodotta e quello con cui viene riprodotta, il supporto (carta, nastro), lo spazio in cui viene conservata ecc. Questi costi si riflettono sull'industria dell'informazione, che era costretta a mettere dei filtri che determinassero quale informazione pubblicare e conservare e quale no: se la carta ha un costo, l'editore dovrà decidere quale romanzo pubblicare e quale no, se l'archivio fisico di una biblioteca è limitato, lo storico dovrà stabilire quali documenti vale davvero la pena conservare e quali no, se un rullino consente di scattare un certo numero di fotografie, il fotografo ci penserà due volte prima di fotografare tutto ciò che gli capita sotto gli occhi.

Nell'era digitale, buona parte di tutti questi costi è stata abbattuta e con questa anche i filtri della conoscenza che prima stabilivano quale fosse informazione di valore e quale no. Con un dispositivo grande pochi centimetri quadrati e dal peso di pochi grammi chiunque può creare contenuti di qualsiasi genere, riprodurli, elaborarli, diffonderli e archiviarli sia suoi propri mezzi che sulle infinitamente capienti memorie dei fornitori di servizi Internet. In questo modo, le aziende accumulano informazioni a costi bassissimi, tanto che preferiscono conservarle tutte piuttosto che scremarle in base alla loro utilità.

1.3 La rivoluzione Internet

Per quanto sociologi ed economisti abbiano cominciato a parlare di Società dell'Informazione a partire dagli anni '60, la grande vera rivoluzione della nuova era è stata la nascita di Internet e successivamente del Web.

[5] Ibidem.

La grande rivoluzione del WWW sta nell'invenzione dell'ipertesto. Il linguaggio HTML e il relativo protocollo HTTP sfruttano la rete non solo per collegare calcolatori e inviarsi informazioni, ma per collegare le informazioni stesse. Il *link* cambia il modo con cui le persone si informano e i processi di conoscenza perdono la loro originaria sequenzialità. Dal Medioevo in avanti, conoscenza e cultura sono sempre state organizzate principalmente sotto forma di libro. Intorno al libro si è creata una *forma mentis* sempre più sofisticata dell'organizzazione del sapere: indici, tassonomie, classificazioni, biblioteche, bibliografie. Il criterio di organizzazione dei contenuti (nella produzione, fruizione, conservazione ecc.) era quello della linearità. I documenti ipertestuali sono invece organizzati a rete, una struttura multilineare e non sequenziale che porta ad una ridefinizione di quel sistema di organizzazione della conoscenza e della cultura. Ad un sistema basato sulle biblioteche se ne sostituisce uno fondato sui motori di ricerca. Al libro si sostituisce il *network* di associazioni che tramite i *link* sono consultabili in modo non lineare e organizzate secondo classificazioni create dall'utente.

Il primo Web, quello degli anni Novanta chiamato Web 1.0, era creato da pagine *web* modificabili esclusivamente dall'autore/proprietario del documento stesso, che doveva avere necessariamente competenze di utilizzo dei linguaggi informatici necessari (HTML, CSS ecc.). In questo modo gli utenti potevano soltanto avere ruolo di fruitori dei contenuti. Gli orizzonti di sviluppo del Web si sono progressivamente ampliati. Nuovi modelli di programmazione, come AJAX, hanno abbassato le barriere tecniche e sono nate nuove piattaforme come i CMS (*content management system*) grazie ai quali gli utenti sono diventati man mano in grado di produrre e condividere i contenuti oltre che consumarli. Con il Web 2.0, fase in cui l'utente può interagire con le pagine *web* e creare contenuti, si apre l'era dei *blog*, delle *wiki* e dei *social network*.

Questi ultimi sono l'espressione massima del Web 2.0 : sulle piattaforme *social* gli utenti, esasperando l'attività di produzione e condivisione di contenuti, arrivano a mettere se stessi sulla Rete, diventano parte di essa e dando così alla luce la loro identità digitale.

1.4 L'identità digitale

Il carattere interattivo del Web 2.0 ha fatto sì che sia diventato una vera e propria comunità composta da tutti i suoi utenti e in una comunità l'identità di chi la compone è una dimensione fondamentale. Le diverse piattaforme del Web 2.0 offrono molte opportunità di costruire e articolare sia la propria identità sociale, cioè quella che si collega al far parte di più cerchie sociali della vita quotidiana come famiglia, colleghi e amici, sia quella personale, che fa riferimento alle caratteristiche del proprio sé che definiscono i tratti della personalità individuale.

Ciò che noi pensiamo di essere, quello che vorremmo che gli altri pensassero di noi e ciò che gli altri effettivamente pensano di noi sono i tre fattori che definiscono l'identità[6]. L'identità si forma all'interno di un complesso processo sociale in cui noi ci raccontiamo e gli altri ci raccontano, un processo che si ridefinisce giornalmente in base alle nostre azioni quotidiane.

Nel mondo *offline* ognuno di noi frequenta diverse reti sociali come la famiglia, gli amici, i colleghi e in ognuna di esse adottiamo linguaggi e comportamenti pertinenti al contesto. In reti sociali distinte possiamo arrivare anche ad assumere caratteri molto diversi: vivaci e sboccati tra amici, seri ed educatissimi in ufficio, per esempio. Ci sono norme più o meno formalizzate che definiscono come ci si debba comportare in una certa cerchia, stabilendo una serie di rituali peculiari. Erwin Goffman parla di rituali della presentazione del sé[7] distinguendo la ribalta dal retroscena. Come in teatro, ciò che portiamo alla ribalta è la parte che vogliamo mostrare in pubblico, mentre nel retroscena lasciamo gli aspetti più privati e personali.

Alcune caratteristiche della comunicazione *online*, come per esempio celarsi dietro un *nick name* o la natura asincrona dei processi comunicativi stessi, sembrerebbero poter dare ampi margini di decisione su cosa portare alla ribalta e cosa lasciare nei retroscena quando costruiamo la nostra identità *online*. Questo perché a differenza della vita *offline*, in quella virtuale si possono scegliere le proprie cerchie sociali, quanto e come parteciparvi e addirittura con quale nome e che immagine identificarsi.

Le piattaforme virtuali offrono un ambiente di sperimentazione dell'identità personale che la realtà *offline* non potrà mai permettersi di fornire.

[6] L. Sartori, *La società dell'informazione*, il Mulino, Bologna, 2012, p. 108.

[7] E. Goffman, *The Presentation of Self in Everyday Life*, Anchor Books, New York, 1959.

Questa peculiarità della *web identity* ha fatto sì che le prime *chat*, *forum* e anche i primissimi *social network* fossero popolati dai *nick name* più fantasiosi. Poter segmentare la propria identità e celarne ogni parte dietro un nome di fantasia aveva un grande fascino e permetteva di esprimere ogni aspetto di sé con una libertà che nel mondo *offline* non sarebbe stata concessa. Le persone potevano costruirsi la propria identità digitale idealizzando e diversificando la propria vera identità, forti anche di una presunta de-responsabilizzazione delle azioni compiute sul Web[8].

Tuttavia, le cose sono andate progressivamente nella direzione opposta. Parallelamente all'assimilazione di Internet nella vita quotidiana, l'interconnessione tra identità *online* e *offline* si è intensificata, fino a diventare inscindibile.

La politica *real name* dei "nuovi" *social network*, come per esempio Facebook, ha imposto agli utenti di utilizzare il vero nome e cognome sul proprio profilo stabilendo così una "forzata" continuità tra identità *offline* e *online*. I *social network* hanno enfatizzato la costruzione sociale dell'identità, costringendo però la presentazione del Sé in uno schema predefinito di categorie come sesso, età, residenza, interessi e istruzione. Come dice la sociologa Laura Sartori, l'architettura stessa dei *social network*

> richiede di riflettere *online* i gusti e le preferenze *offline*, scoraggiando la separazione tra identità virtuale e reale. [...] Un processo squisitamente personale – in cui si ricerca un alto grado di autonomia (la presentazione del Sé) e in cui ognuno può decidere cosa (e come) portare in pubblico – appare sempre più instradato e guidato da altri. In particolare i Social Media sembrano operare una sorta di rivelazione del Sé predefinita secondo un template dove l'interazione sociale passa anche attraverso la piattaforma sociali. [9]

In poche parole, se da un lato sul Web possiamo avere un maggior controllo dei modi in cui decidiamo di presentarci, dall'altro le piattaforme su cui lo facciamo ci hanno progressivamente costretti ad assottigliare il confine tra identità reale e virtuale vanificando di fatto quei benefici che la Rete poteva offrire.

Quello che noi decidiamo di portare alla ribalta non dipende soltanto da noi, ma anche da altri attori sociali in continuo e reciproco processo di

[8] Sartori, *La società dell'informazione*, cit., p. 110.
[9] *Ivi*, p. 112.

interazione. Le piattaforme *social* sono oggi uno strumento molto usato per la presentazione del Sé e per trovare collocazione anche rispetto ad altri soggetti. La costruzione del nostro profilo *online* non è lasciata al caso ma è un atto consapevole per inserisi in uno specifico ambiente digitale. Attraverso l'identità digitale costruiamo la nostra reputazione *online* che oggi, dove il reale e il virtuale non hanno ormai confini così definiti, può avere forti ripercussioni su quella *offline*.

Pubblicare una foto su Facebook, un pensiero o un *link*, significa rivelare consapevolmente qualcosa del proprio modo di essere che incide sulla percezione che gli altri hanno di noi. Proprio come nella vita *offline*, il modo di parlare, le amicizie con relative reti sociali, l'aspetto fisico e il lavoro parlano di noi anche *online*. Le tecnologie Web 2.0 permettono anche a chi legge di poter partecipare alla pubblicazione di una foto, un pensiero o un *link*, attraverso un commento, un *tag*, un *like* o una condivisione. Se prima, nel Web 1.0, chi scriveva aveva il controllo totale su ciò che diceva di sé, oggi l'informazione che circola su di noi dipende anche da ciò che gli altri dicono sul nostro conto.

1.5 La riproducibilità tecnica dell'Io

In sintesi, viviamo in un'epoca in cui l'informazione ha un grande valore economico e sociale ma che allo stesso tempo costa pochissimo, non solo in termini di denaro ma di qualsiasi altra risorsa, come tempo, spazio, forza, abilità e talento, grazie al suo formato digitale. Ambiente principe in cui l'informazione digitale circola, si crea, si diffonde e si conserva è il Web, che ha modificato la struttura della nostra conoscenza, da lineare a reticolare, ha abbattuto certi limiti spazio-temporali e ha sconfitto i filtri dei media tradizionali. Per necessità o tendenza, anche le nostre identità sono entrate a far parte del Web, sotto forma di informazioni prodotte da noi stessi o da terzi. Le nostre informazioni presentano le stesse caratteristiche di qualsiasi altra informazione digitale: sono riproducibili, accessibili e condivisibili da chiunque, a costo praticamente nullo. La facilità con cui l'informazione è riproducibile e divulgabile e archiviabile, l'assenza dei tradizionali filtri del sapere e la non linerarità della struttura della rete fa sì che spesso l'informazione venga decontestualizzata.

Wikipedia alla voce "Era dell'informazione"[10] fornisce una definizione molto interessante e diversa rispetto alle definizioni più tradizionali: «periodo in cui il movimento dell'informazione divenne più veloce del movimento fisico». Per quanto molto astratta questa definizione rende bene l'idea di cosa stia alla base di tutti i fattori che hanno dato luogo alla Società dell'Informazione: la tecnica che ha accelerato a ritmi semplicemente non umani qualsiasi azione intorno all'informazione.

Prima dell'era digitale qualcuno aveva già intuito che se qualcosa può essere riprodotto ad una velocità maggiore di quella della mano umana, rischia di perdere il suo autentico valore e, in ultima analisi, essere decontestualizzata, cioè alienata da ciò che le dà ragione di essere.

Questo è Benjamin, che in *L'opera d'arte nell'epoca della sua riproducibilità tecnica*[11] sostiene che nel momento in cui l'arte diventa riproducibile tecnicamente (cioè ad un ritmo diverso da quello del semplice lavoro manuale tramite cui fino ad allora era sempre stata riprodotta) questa perde la sua aura, perché si ritrova decontestualizzata dal luogo e momento in cui viene creata, ciò che lui chiama *hic et nunc* dell'opera d'arte.

Mentre in realtà l'arte è sopravvissuta alla sua riproducibilità tecnica e anzi ha trovato nella tecnologia nuovi mezzi tramite cui esprimersi, questo non si può dire dell'identità dell'individuo, perché oggi anche l'Io è riproducibile tecnicamente, con esiti a volte devastanti.

Il fatto che la nostra identità digitale sia costituita da informazioni condivise e conservate sulla Rete e che queste informazioni, per loro natura digitale, sono riproducibili, accessibili e condivisibili, fa sì che la nostra identità, il nostro Io, sia riproducibile tecnicamente. Una foto che ci ritrae, un dato anagrafico o un commento sul Web sono tutte riproduzioni di una piccola o grande parte della nostra identità.

Esattamente come Benjamin dice per l'arte, anche l'Io è stato da sempre riproducibile, soprattutto tramite l'arte figurativa e la letteratura. Dipingere il ritratto di una persona o scriverci sopra una storia, un trattato storiografico o semplicemente un articolo di giornale risente però di tutti quei costi che la produzione e la conservazione (e quindi memorizzazione) di questa riproduzione di un'identità implica. Riprodurre una persona doveva perciò avere

[10] Wikipedia, *voce Era dell'informazione*, https://it.wikipedia.org/wiki/Era_dell%27informazione, 19/4/2017.

[11] W. Benjamin, *L'opera d'arte nell'epoca della sua riproducibilità tecnica*, trad. di E. Filippini, Einaudi, Torino, 2000.

uno scopo, la sua identità doveva avere un ruolo nella storia, nella società o comunque nell'interesse della collettività. La facilità con cui oggi possiamo fotografare, riprendere, registrare, in poche parole riprodurre tecnicamente una persona e immediatamente rendere la sua immagine riprodotta pubblica a tutto il mondo rischia di far sì che l'Io digitale si alieni da quello reale.

Se infatti l'arte, come dice Benjamin, ha un suo *hic et nunc*, il qui e ora originario dell'opera che ne stabilisce la sua autenticità, la sua aura, anche l'Io ha il proprio *hic et nunc*, anzi, ne ha più di uno.

La nostra personalità infatti, che attraverso l'azione stabilisce l'immagine che gli altri hanno di noi, si manifesta attraverso una serie consecutiva di momenti unici e irripetibili, di *hic et nunc*, che possono essere molto diversi fra loro, ma tutti necessari a definire l'identità di una persona. Più semplicemente: ciò che l'individuo fa e dice il sabato sera in birreria con gli amici e ciò che dice e fa il lunedì successivo in ufficio sono cose ben diverse ed è bene che queste due fasi, che si consumano in momenti e luoghi diversi (due diversi *hic et nunc* appunto), non entrino in contatto fra loro. Il fatto che ci si comporti in modo diverso in due momenti distinti della propria vita non significa essere incoerenti e non implica che una di queste due manifestazioni del proprio Io sia sbagliata, anzi, probabilmente l'una è necessaria all'altra: se una persona è capace di scherzare la sera con gli amici probabilmente sarà di mente più aperta anche in ambito lavorativo, così come il fatto di essere responsabile in ufficio consente di saper riconoscere i limiti del divertimento.

Ciò che non deve assolutamente succedere è che quello che avviene in birreria avvenga in ufficio. Le conseguenze sarebbero devastanti, come è facile immaginare. Il problema è che oggi il rischio che questo è accada è altissimo, proprio perché l'Io è riproducibile tecnicamente. Basta che, mentre questa persona si trova in birreria, un amico, o lui stesso, impugni il proprio *smartphone* ed ecco che il lunedì successivo tutti i colleghi in ufficio potranno vedere cos'ha detto e fatto il sabato scorso.

Un *hic et nunc* è stato riprodotto tecnicamente ed è stato sovrapposto ad un altro, è stato cioè decontestualizzato. La successione lineare degli *hic et nunc* che definiscono l'Io ha subito un'interruzione. L'Io reale, fatto da una successione ordinata di *hic et nunc*, si scontra con l'Io digitale, che è invece sovrapposizione di *hic et nunc* ripescati dalla storia dell'individuo e cristallizzati.

1.6 Memoria analogica e memoria digitale

Secondo lo psicologo statunitense Daniel Schacter[12], il ricordo umano non è semplice recupero meccanico di fatti dal passato, ma un processo che si basa sulla costante ricostruzione del passato basata sul presente. L'idea è intuitiva: ogni evento del passato è rileggibile in base alle conseguenze che questo ha avuto sul nostro presente. Nell'arco di un'amicizia, è molto probabile che tra due persone venga detto o fatto qualcosa di sbagliato prima o poi, ma quegli eventi spiacevoli verranno in seguito ricordati alla luce dell'amicizia che, malgrado inevitabili inconvenienti, è rimasta immutata. Questo perché ogni frammento del passato viene ricostruito lungo un filo, lineare e continuo, del presente. Il presente è il contesto in cui i nostri ricordi vivono e si intrecciano fra di loro per creare la nostra memoria. La memoria "analogica" quindi è continua, costantemente rimodulata in base al presente ma grazie a questo sempre contestualizzata. Proprio come qualsiasi altra informazione analogica, il ricordo analogico è soggetto a rumore: modifiche, offuscamenti, perdita di dettagli che si aggiungono col tempo. Ma proprio grazie a questo "rumore" ogni evento del passato si armonizza con gli altri e acquista così la sua coerenza con il presente.

La memoria digitale invece, quella costituita da informazioni digitali archiviate su supporti digitali, è immutabile nel tempo e non soggetta a rumore. Il passato digitale non si ricostruisce sul filo del presente, ma è un mucchio di frammenti del passato, cristallizzati così come al momento del loro "campionamento", che si sovrappongono al presente. La memoria digitale non produce l'intero ricordo ma solo gli elementi che ha registrato. Anni di vita vengono amalgamati in un *collage* di informazioni personali, ognuna delle quali era vera in un preciso momento del passato di un individuo. Ecco che dunque l'informazione digitale, per quanto perfetta nella sua risoluzione, è del tutto decontestualizzata.

In poche parole, ciò che differenzia la memoria analogica da quella digitale è il rapporto tra passato e presente: nella memoria analogica il passato è subordinato al presente, che costituisce il contesto in cui trova la sua ragione d'essere, nella memoria digitale passato e presente si trovano sullo stesso piano, si sovrappongono, si mescolano dando luogo ad una sorta di eterno

[12]D. Schacter, *The seven sins of memory: How the mind forgets and remembers*, Houghton Mifflin, Boston, 2001.

presente.

Proprio di memoria digitale parla Viktor Meyer-Schoenberger, noto professore di Internet Governance and Regulation a Oxford, quando definisce, nel suo saggio *Delete* [13] la vera grande rivoluzione dell'età contemporanea, il cambiamento di paradigma della Società dell'Informazione rispetto a tutto il resto della storia dell'umanità: le tecnologie digitali hanno provocato la rottura di un equilibrio antico tanto quanto l'uomo per cui dimenticare è sempre stato la norma e ricordare l'eccezione.

Dimenticare è facile, è una legge biologica, un meccanismo necessario a filtrare le informazioni che i nostri sensi raccolgono ogni istante in enormi quantità su tutto ciò che ci circonda e tutto ciò che avviene dentro di noi. Se il nostro cervello non selezionasse drasticamente tutte queste informazioni non saremmo in grado di comprendere la realtà e prendere decisioni.

Ricordare invece è difficile, faticoso e costoso. Ricordare costa in termini energetici, in quanto è processo che va contro la seconda legge della termodinamica per la quale i sistemi si tendono a evolversi verso l'entropia e il disordine. Costa anche in termini di tempo, fatica e denaro per gli individui e per la società, perché devono compiere grandi sforzi per andare contro la tendenza naturale all'oblio.

Le tecnologie digitali hanno abbattuto nel giro di pochi anni tutti questi costi e così l'equilibrio tra ricordo e oblio si è sovvertito: oggi ricordare è diventato la norma e dimenticare l'eccezione:

> Dall'inizio del tempo, per noi umani, dimenticare è stato la norma e ricordare l'eccezione. A causa della tecnologia digitale e delle reti globali questo bilanciamento è cambiato. Oggi con l'aiuto dell'altissima diffusione tecnologica dimenticare è diventata l'eccezione e ricordare la norma. [14]

Oggi memoriziamo sui supporti digitali qualsiasi cosa, in poche parole ricordiamo tutto. Lo sforzo che prima impiegavamo per ricordare costituiva il nostro filtro per conservare ciò che era davvero utile, ma oggi questo sforzo non è più necessario e con lui anche il filtraggio viene neutralizzato.

Il tramonto dell'oblio nella contemporanea Società dell'Informazione comporta una grave conseguenza: la perdita del controllo che gli individui hanno sulle proprie informazioni.

[13] V. Mayer-Schönberger, *Delete. Il diritto all'oblio nell'era digitale*, Egea, Milano, 2013.
[14] Mayer-Schönberger, *Delete. Il diritto all'oblio nell'era digitale*, cit., p. 6.

La gratutità dell'informazione digitale ha come conseguenza la leggerezza con cui questa viene trattata dalle persone, compresa l'informazione che li riguarda. La struttura del Web 2.0, utilizzata ormai più come strumento di condivisione che di accesso alle informazioni, la diffusa inconsapevolezza degli utenti che le informazioni, una volta condivise, sfuggono al loro controllo e possono essere usate da terzi (aziende, governi ecc.) e la costante ricerca di visibilità dilagata con l'avvento dei *social media* hanno portato alla riduzione della libertà di scegliere se e quando portarsi alla ribalta e se e quando restare al riparo dagli sguardi altrui. Controllare l'immagine di sé sulla Rete è diventato un problema fondamentale.

Soprattutto, la memoria digitale sta cancellando di fatto la possibilità, che ogni essere umano aveva fino ad ora avuto, di ridisegnare periodicamente la propria identità. Nel tempo ognuno di noi si evolve, l'età e l'esperienza ci rinnovano sempre. Il perdigiorno può diventare uomo in carriera, il festaiolo può farsi prete, il *playboy* può sposarsi e dimostrarsi un marito fedele. Ma se le nostre identità passate diventano informazione digitale, indelebile e immutabile, la nostra società diventa incapace di perdonare perché incapace di dimenticare.

1.7 L'Io *online* e l'Io *offline*

Oggi sono sempre di più i casi di licenziamenti, colloqui di lavoro negati, carriere rovinate perché il *collage* digitale che dava forma all'identità in rete dei soggetti interessati aveva generato mostri incontrollabili vaganti per il Web.

Andrea Barchiesi, CEO di Reputation Manager definisce con "stretta di mano digitale" la

> conoscenza indiretta di un individuo ottenuta mediante l'acquisizione di un insieme di informazioni pubbliche raccolte attraverso Internet. Questa anticipa, condiziona e a volte impedisce la stretta di mano fisica. [15]

La forma più comune e banale della stretta di mano digitale è la ricerca su Google del nome e cognome della persona. La stretta di mano digitale è un gesto non percepibile dal soggetto osservato, è un incontro fra due persone

[15] A. Barchiesi, *La tentazione dell'oblio*, Franco Angeli, Milano, 2016, p. 34.

di cui una ne è inconsapevole perché è praticamente impossibile accorgersi che qualcuno sta raccogliendo informazioni su di noi. L'inconsapevolezza comporta l'inevitabilità di quest'incontro, proprio perché uno dei soggetti non è presente e quindi le decisioni le prende solo l'altro. La stretta di mano digitale può venire in qualsiasi momento e in qualsiasi luogo ed è estremamente condizionante: le informazioni su un soggetto possono influenzare a tal punto l'opionine di chi osserva da poter comportare anche l'annullamento del contatto reale.

Nella stretta di mano digitale, i motori di ricerca giocano un ruolo determinante: i risultati che associano a una determinata ricerca influiscono sulle scelte quotidiane e fondamentali come assumere una persona, comprare un prodotto piuttosto che un altro o investire soldi in un'attività.

I motori di ricerca fanno sì che la nostra identità digitale sia verificabile da chiunque in qualsiasi momento, con estrema facilità e velocità. Dunque l'identità digitale ci precede sempre e stringe la mano alle persone a nostra insaputa. Non solo: i motori di ricerca non si limitano a rendere accessibile la nostra identità digitale, ma la costruiscono pure. In base a ciò che ritengono più rilevante per la ricerca effettuata, scelgono, dispongono e gerarchizzano le informazioni di una persona decidendo autonomamente come presentarle. In poche parole i motori di ricerca parlano al posto nostro senza che noi ci si sia espressi se siamo d'accordo oppure no.

L'Io nell'epoca della sua riproducibilità tecnica è un Io diviso tra un Io *offline*, analogico, contestualizzato nel suo presente e costantemente rinnovabile rispetto al suo passato, e un Io *online*, digitale, decontestualizzato, che si accumula al suo passato. L'Io *online* è una riproduzione dell'Io *offline*, ma il paradosso è che l'Io *online* ha più influenza sul destino dell'Io *offline* che non il contrario.

Capitolo 2

Diritto all'oblio e motori di ricerca

2.1 Cos'è il diritto all'oblio

Quando si parla dei rischi legati alla persistenza di informazioni personali sulla Rete, è necessario considerare preliminarmente quali sono i tipi di tutela che la Legge offre in termine di protezione dei dati personali sul *web*. Il diritto che tutela lo "storico" dell'identità digitale chiamato diritto all'oblio.

Definito comunemente come "diritto ad essere dimenticati", come se davvero si potesse rivendicare la cancellazione dalla memoria comune parti del proprio passato, il diritto all'oblio può assumere diverse accezioni: diritto a non essere facilmente trovati, o non essere facilmente visti, diritto alla contestualizzazione dei dati in Rete o diritto a «vedersi rappresentati in modo da riflettere la propria attuale dimensione personale e sociale e, di conseguenza, a non essere rappresentati in maniera non più corrispondente a quella»[1].

La definizione di diritto all'oblio non è ancora univoca, ma è abbastanza chiaro in quali casi questo diritto debba "intervenire": quando il contenuto di una notizia pubblicata crea uno squilibrio tra diritto alla riservatezza e diritto all'informazione.

[1] L. De Grazia, *La libertà di stampa e il diritto all'oblio nei casi di diffusione di articoli attraverso Internet: argomenti comparativi*, "AIC", 4/2013, 29 ottobre 2013.

2.1.1 Prime definizioni di diritto all'oblio

Il diritto all'oblio è un concetto giurisprudenziale abbastanza recente. Lo sviluppo di questo diritto va di pari passo con la diffusione dei mezzi di comunicazione di massa e l'interesse su di esso si è intensificato drasticamente con l'avvento di Internet, il quale, a causa della sua smisurata capacità di archiviazione ed estrema facilità di consultazione, ha rivelato i pericoli di una memoria illimitata e collettiva.

I primi casi di formulazione del diritto all'oblio risalgono tuttavia alla carta stampata, in particolare in ambito giornalistico dove, ancor prima dell'era digitale, era nato uno squilibrio tra il diritto di cronaca e diritto alla riservatezza. In questa fase "pre-Internet" si faceva riferimento al diritto di un soggetto a non vedere pubblicate notizie relative a vicende, già legittimamente pubblicate, rese sostanzialmente obsolete da un consistente lasso di tempo tra l'accadimento e la pubblicazione.

In Italia, il diritto all'oblio è stato menzionato per la prima volta nel 1983 da Amedeo Tommaso Auletta nel suo saggio *Diritto alla riservatezza e "droit à l'oubli"* in cui si poneva il problema

> di stabilire se la persona o le vicende lecitamente pubblicizzate possano sempre costituire oggetto di nuova pubblicizzazione o se, invece, il trascorrere del tempo e il mutamento delle situazioni non la rendano illecita.[2]

Protagonista del dibattito sul diritto all'oblio è il tempo: esso può stabilire la perdita di rilevanza di una notizia nei confronti di un individuo al punto di farne decadere i diritti di cronaca e all'informazione.

Il primo caso di applicazione, seppur ancora implicita, del diritto all'oblio in Italia, risale al 15 maggio 1995, da parte del Tribunale di Roma.

Il caso riguardava un quotidiano nazionale che aveva ripubblicato, per fini puramente promozionali, alcune sue prime pagine, tra cui una del 6 dicembre 1961, la quale riportava la notizia della confessione di un omicidio con tanto di foto e nome del colpevole. Costui, dopo più di trent'anni dal fatto, aveva scontato la sua pena e ricevuto provvedimento di grazia dal Presidente della Repubblica per buona condotta, riuscendo in seguito a ricostruirsi, faticosamente, una vita professionale ed affettiva.

[2] Auletta, *Diritto alla riservatezza e "droit à l'oubli"*, in G. Alpa, M. Bessone, L. Bonechi, G. Caiazza (a cura di), *L'informazione e i diritti della persona*, Napoli, 1983, pp. 127 e ss.

La ripubblicazione su scala nazionale della sua immagine associata alla confessione dell'omicidio comportò gravi ripercussioni suoi rapporti personali e gli fece perdere il lavoro. Citato in giudizio l'editore del quotidiano, il Tribunale condannò questi al risarcimento dei danni morali nei confronti dell'attore, stabilendo che si trattava di reato di diffamazione. La natura diffamatoria della notizia, secondo il tribunale, stava nel fatto che questa non fosse più di interesse pubblico, trattandosi di un fatto di sangue avvenuto molto tempo prima e di nessun valore "storico". Si tratta, pur dietro la formula di un reato di diffamazione, del primo caso di riconoscimento del diritto all'oblio nella giurisprudenza italiana.

Ironia della sorte, dopo quasi vent'anni dalla sentenza, quel quotidiano ha digitalizzato tutti i suoi numeri e li ha resi disponibili al pubblico tramite l'archivio storico del suo sito *web*. Oggi, chiunque può trovare la prima pagina del 6 dicembre 1961 e leggere il titolo «S.S. ha confessato di avere ucciso B.C.».

Nel 1998 la terza sezione civile della Corte di Cassazione si trovò ad esaminare il caso di Mario Rendi[3], imprenditore che aveva citato per danni patrimoniali e non patrimoniali un giornalista e un editore del settimanale "Avvenimenti" su cui era comparso un articolo relativo al suo presunto coinvolgimento in fatti di collusione tra mafia e imprenditoria, già pubblicati in precedenza dalla stampa.

La Cassazione ripercorre il processo decisionale condotto dalla Corte d'Appello, la quale aveva fatto luce sull'attualità e grado di aggiornamento della notizia, facendo riferimento alle tre condizioni, secondo il decalogo dei giornalisti, che rendono lecita la divulgazione di notizie che arrecano pregiudizio all'onore e alla reputazione: la verità oggettiva della notizia pubblicata, l'interesse pubblico alla conoscenza del fatto, la correttezza formale dell'esposizione. La sentenza d'appello aveva tuttavia individuato un ulteriore limite del diritto di cronaca, cioè l'attualità della notizia:

> non è di per sé lecito divulgare nuovamente, dopo un consistente lasso di tempo, una notizia che in passato era stata legittimamente pubblicata.

L'interesse del pubblico, uno dei tre punti che giustificano la diffusione di notizie lesive della reputazione di un individuo, è necessariamente legato

[3]Sentenza n. 3679, cit.

al fattore tempo: una notizia vecchia non rientra più nell'interesse pubblico e, quindi, non ha più senso di essere riesumata. Un soggetto ha diritto di non rimanere esposto per sempre ai danni causati da una notizia lecitamente pubblicata in passato e che successivamente continua a danneggiare la sua reputazione:

> Viene invece in considerazione un nuovo profilo del diritto di riservatezza recentemente definito anche come diritto all'oblio inteso come giusto interesse di ogni persona a non restare indeterminatamente esposta ai danni ulteriori che arreca al suo onore e alla sua reputazione la reiterata pubblicazione di una notizia in passato legittimamente divulgata.

Anche se ancora sotto la copertura del diritto alla riservatezza, si tratta della prima definizione nella giurisprudenza italiana del diritto all'oblio. Poco dopo viene aggiunto:

> Ma, quando il fatto precedente per altri eventi sopravvenuti ritorna di attualità, rinasce un nuovo interesse pubblico alla informazione non strettamente legato alla stretta contemporaneità fra divulgazione e fatto pubblico che si deve contemperare con quel principio, adeguatamente valutando la ricorrente correttezza delle fonti di informazione.

Il fattore tempo dunque non basta a garantire la tutela del diritto all'oblio: se, rispetto alla prima divulgazione della notizia, emergono nuovi elementi di interesse pubblico, le informazioni devono essere aggiornate e quindi il diritto all'oblio cede il passo al diritto di cronaca. In giurisprudenza un diritto non può mai prevaricare su un altro, non si può cioè sospendere un diritto per garantirne un altro. In caso di conflitto fra due diritti, l'unica soluzione è ricorrere ad un bilanciamento. Un contenuto che danneggi l'identità di un individuo può essere dimenticato solo se, nel corso del tempo, non emergano nuovi elementi di interesse pubblico che ne giustifichino un ulteriore aggiornamento.

Nel caso della sentenza 3679/1998, la Corte Suprema ha poi evidenziato come la corte d'appello non si fosse limitata a considerare il notevole lasso di tempo trascorso tra la prima e la seconda pubblicazione, in quanto non comporta di per sé l'applicazione del diritto d'oblio, ma aveva verificato anche che nel frattempo i procedimenti giudiziari nei confronti dell'attore fossero

stati archiviati per comprovata estraneità ai fatti di mafia, cosa che invece il giornalista del settimanale non si era preoccupato di fare dimostrando così mancanza di buona fede e chiara volontà di diffamare.

2.1.2 Diritto all'oblio e Internet

Le definizioni e le applicazioni del diritto all'oblio nell'era pre-digitale, o comunque quando l'uso di Internet era ancora agli albori, rivelano che nei cosiddetti mezzi di comunicazione tradizionali il problema dell'oblio si concentrava sull'eventuale ripubblicazione di una notizia. Giornali, televisione e radio possono riesumare una notizia solo ripubblicandola, in quanto la sua prima e lecita pubblicazione è destinata in breve tempo (il solo istante in cui viene trasmessa in televisione o radio, magari qualche giorno per un giornale) ad essere dimenticata dai più e restare reperibile solo accedendo e frugando in qualche archivio fisico.

Tutto è cambiato con l'avvento di Internet: nella Rete la ripubblicazione di un contenuto non è necessario per la stessa organizzazione dell'informazione nella rete stessa. Nel web niente viene cancellato ma permane potenzialmente per sempre, accessibile tramite un URL. Un contenuto su internet ha una vita infinitamente più lunga rispetto agli altri media tradizionali e la stessa struttura del web fa sì che questo rimanga accessibile per moltissimo tempo da luoghi diversi.

Per quanto riguarda l'oblio in rete, il problema non è più quello della ripubblicazione della notizia, ma della sua permanenza e indeterminata accessibilità. L'obsolescenza dell'informazione non si misura più in base al tempo trascorso tra la sua pubblicazione e la ripubblicazione, ma al tempo di permanenza sulla Rete. Alla questione dell'aggiornamento della notizia, si sostituisce il problema del contesto: si sta parlando dello stesso contenuto della prima collocazione, nello stesso luogo in cui stato pubblicato e con la stessa data, ma che potrebbe comparire in un contesto non più appropriato. L'esigenza del diritto all'oblio su Internet non è più la ripubblicazione di una notizia, ma la sua pertinenza col presente. Si dovrebbe infatti parlare, più che di diritto ad essere dimenticati, di diritto ad essere contestualizzati, il diritto che alle informazioni che ci riguardano sia dato il giusto peso al fine di garantire che la nostra identità non sia travisata sulla Rete.

La tutela del diritto all'oblio su Internet si scontra a livello legislativo, ma più in generale a livello concettuale, con l'indefinitezza di molti aspetti di In-

ternet sul piano giuridico. In questo modo, l'informazione su Internet è stata finora considerata in base alle regole che valevano per i media tradizionali.

2.2 La responsabilità dei motori di ricerca sui dati personali

2.2.1 La sentenza della corte Europea del 13 maggio 2014

Come accennato nell'introduzione, nel 1998 l'avvocato spagnolo Mario Costeja Gonzalez subì un rovescio finanziario e la sua casa fu messa all'asta. Il quotidiano spagnolo *La Vanguardia* pubblicò un annuncio che faceva riferimento alla vendita all'asta di un appartamento di Gonzalez a seguito di un pignoramento effettuato per la riscossione coattiva di crediti previdenziali. La notizia fu pubblicata anche nella versione *web* del quotidiano.

L'avvocato Gonzalez riuscì ad appianare il suo debito ancor prima che la sua abitazione fosse venduta e tutto sembrava essersi sistemato al meglio. Tuttavia, cercando il suo nome e cognome su Google, scoprì che la notizia pubblicata da *La Vanguardia* compariva in posizione piuttosto visibile tra i risultati di ricerca.

Nel 2009, Gonzalez si rivolse all'Agenzia Spagnola di Protezione Dati (AEDP), vale a dire il Garante della *privacy* spagnolo, affinché quel risultato fosse rimosso dal motore di ricerca. L'AEDP accolse l'istanza, ritenendo che sussistessero fondate ragioni per chiamare in giudizio Google e poco tempo dopo la questione approdò alla Corte di Giustizia Europea, la con la sentenza n. C-131/12 del 13 maggio 2014 riconobbe il diritto all'oblio di Mario Costeja Gonzalez e stabilì che i cittadini europei hanno il diritto di richiedere al motore di ricerca la rimozione di informazioni associate al proprio nome qualora questa appaiano

> inadeguate, non pertinenti o non più pertinenti, ovvero offensive in rapporto alle finalità del trattamento in questione realizzato dal motore di ricerca [4]

Si tratta di una svolta giurisprudenziale epocale nella storia del diritto all'oblio e nell'ambito della tutela dei diritti in Rete. In primo luogo, perché ha attribuito a Google, e ai motori di ricerca in generale, la responsabilità del

[4] Sentenza n. C-131/12, cit.

trattamento dei dati personali, con conseguente obbligo di deindicizzazione delle pagine *web* che li contengono, sotto richiesta del titolare di quei dati.

Questo significa che l'operazione di indicizzazione, che prevede la localizzazione delle informazioni pubblicate da terzi, memorizzazione dei contenuti e dei metadati al fine di restituire tali informazioni come risultati di ricerche effettuate da utenti, è a tutti gli effetti trattamento dei dati personali ai sensi dell'art. 2 lettere b) e d) della direttiva 95/46/CE del Parlamento europeo e del consiglio europeo del 24 ottobre 1995.

La logica secondo la quale l'attività di un motore di ricerca è effettivamente trattamento di dati personali, sta nel fatto che l'accessibilità per gli utenti alle informazione tramite i loro servizi avvenga sulla base di dati trattati secondo i protocolli decisi dal motore di ricerca. Nella sentenza si legge che il motore di ricerca raccoglie, estrae, registra e organizza i dati all'interno dei suoi sistemi di indicizzazione, li conserva nei propri *server* e li comunica ai suoi utenti sotto forma di elenchi di risultati. In questo modo il motore di ricerca fornisce all'utente che esegue la ricerca, attraverso la pagina dei risultati associati al nome e cognome di una persona, una visione strutturata delle informazioni relative al soggetto su Internet. Quest'immagine strutturata tocca una molteplicità di aspetti legati alla vita privata e professionale del soggetto, che non avrebbe potuto diffondersi, o almeno con molta difficoltà, senza l'"aiuto" del motore di ricerca.

Si tratta, secondo la Corte, di una violazione dei diritti fondamentali della persona giustificabile soltanto nel caso in cui l'interesse pubblico a conoscere quelle informazioni prevalga sul diritto alla riservatezza del soggetto.

La sentenza inoltre stabilisce che

> il gestore di un motore di ricerca è obbligato a sopprimere, dall'elenco di risultati che appare in seguito di una ricerca effettuata a partire dal nome di una persona, dei link verso pagine web pubblicate da terzi e contenenti informazioni relative a questa persona, anche nel caso in cui tale nome o tali informazioni non vengano previamente o simultaneamente cancellati dalle pagine web di cui trattasi, e ciò eventualmente anche quando la loro pubblicazione su tali pagine web sia di per sé lecita.

Questo passo introduce un'altra novità degna di nota: la rimozione di un *link* dagli indici del motore di ricerca, non comporta l'obbligo per il sito-sorgente ai quali il *link* rimanda. La relativa pagina dunque può restare

visibile anche se il suo *link* è rimosso dal motore di ricerca. L'interessato potrà poi eventualmente agire parallelamente affinché la notizia venga eliminata dalla fonte.

Di fatto, di ciò che compare su Google ne risponde Google, indipendentemente dalle responsabilità del sito-sorgente riguardo ai contenuti che sono stati indicizzati. Inversamente, il sito-sorgente che "subisce l'oblio" da parte del motore di ricerca non ha nessuna nessun margine di obiezione sulla rimozione del *link* alle proprie risorse dagli indici del motore. Se Google stabilisce che una richiesta di rimozione vada accolta, esso è l'unico coinvolto nella decisione, senza alcun diritto di replica da parte di terzi.

La grande svolta sta nel fatto che la responsabilità del trattamento dei dati personali, un tempo attribuita esclusivamente ai gestori dei siti *web*, adesso è stata estesa anche ai motori di ricerca. La ragione è che il modo con cui il motore di ricerca indicizza e restituisce contenuti agli utenti non è neutrale: Google effettua una scelta dei dati, una gerarchizzazione delle informazioni secondo criteri precisi e per altro non divulgati e stabilisce il modo con cui presentarli agli utenti.

Google contribuisce attivamente alla costruzione dell'identità digitale di una persona attraverso i suoi algoritmi di *ranking* e dunque diventa responsabile di ciò che restituisce nella pagina dei risultati e di come li dispone all'interno di essa. Questa è la grande novità introdotta dalla sentenza del caso Google Spain, di gran lunga più rilevante che nella sua effettiva influenza sulla tutela del diritto all'oblio, come dimostrano le conseguenze della sentenza stessa.

2.2.2 La reazione di Google alla sentenza

Google non accolse con favore la sentenza della Corte Europea. D'altra parte, la sentenza andava a colpire la sua stessa *mission*, ossia la creazione del più efficiente e universale sistema di diffusione delle informazioni. Dare agli utenti la possibilità di "metter bocca" su cosa dovesse o non dovesse stare su Google rappresentava un colpo basso alla sua autorità di "signore del Web", oltre che un pesante limite alla sua attività e al suo *business*.

Non potendo restare, in ogni caso, indifferente alla sentenza, solo due giorni più tardi Google pubblicò un modulo attraverso il quale i cittadini europei possono fare richiesta di rimozioni dei contenuti associati al proprio nome direttamente al motore di ricerca. Dopo un mese dalla novità intro-

dotta da Google, Bing, il motore di ricerca di Microsoft, seguì l'esempio pubblicando a sua volta un *form* per richiedere l'eliminazione dei contenuti, addirittura più dettagliata e chiara di quella di Big G.

Si apre così una fase dell'informazione tramite Internet dove il cittadino (europeo) può finalmente partecipare attivamente al trattamento dei propri dati personali che circolano sulla Rete. Le cose tuttavia non sono andate come sperato.

Innanzi tutto, vanno fatte alcune precisazioni sulla natura giurisprudenziale della sentenza e sull'effettivo ruolo del modulo di rimozione messo a disposizione da Google.

La natura del diritto all'oblio riconosciuta dalla Corte europea non copre qualsiasi informazione riguardi un soggetto, bensì solo i contenuti «inadeguati, irrilevanti o non più rilevanti» o in ogni caso eccessivi in relazione agli scopi per cui sono stati pubblicati. Nella sentenza è chiaramente precisato che il diritto alla rimozione dei *link* non scatterebbe «qualora risultasse, per ragioni particolari, come il ruolo ricoperto da tale persona nella vita pubblica, che l'ingerenza nei suoi diritti fondamentali è giustificata dall'interesse preponderante del pubblico suddetto ad avere successo, in virtù dell'inclusione summenzionata, all'informazione di cui trattasi».

Secondo la Corte, i criteri di valutazione per il bilanciamento tra il diritto alla riservatezza e quello all'informazione devono basarsi sugli articoli 7 e 8 della Carta dei diritti fondamentale dell'Unione, ossia:

> *Articolo 7*
>
> **Rispetto della vita privata e della vita familiare**
>
> Ogni individuo ha diritto al rispetto della propria vita privata e familiare, del proprio domicilio e delle sue comunicazioni.
>
> *Articolo 8*
>
> **Protezione dei dati di carattere personale**
>
> - Ogni individuo ha diritto alla protezione dei dati di carattere personale che lo riguardano.
> - Tali dati devono essere trattati secondo il principio di lealtà, per finalità determinate e in base al consenso della persona interessata o a un altro fondamento legittimo previsto dalla legge. Ogni individuo ha il diritto di accedere ai dati raccolti che lo riguardano e di ottenerne la rettifica.
> - Il rispetto di tali regole il soggetto al controllo di un'autorità indipendente.

I diritti descritti in questi due articoli prevalgono «non soltanto - si legge nel considerando 97 della sentenza - sull'interesse economico del gestore del motore di ricerca, ma anche sull'interesse del pubblico a trovare l'informazione suddetta in occasione di una ricerca concernente il nome di una persona» a meno che «risultasse, per ragioni particolari, come il ruolo ricoperto dalla persona nella vita pubblica, che l'ingerenza nei suoi diritti fondamentali è giustificata dall'interesse preponderante del pubblico suddetto ad avere accesso, medieante l'inclusione summenzionata, all'informazione di cui trattasi».

Google dal canto suo, sulla sua pagina dedicata alla rimozione dei contenuti, precisa che considera informazioni di interesse pubblico le frodi finanziarie, la negligenza professionale, le condanne penali o la condotta pubblica di funzionari statali.

Chi fa richiesta di rimozione deve dimostrare che la permanenza sulla Rete dell'indirizzo ai contenuti che lo riguardano lo espone a pregiudizio sociale o lavorativo. Google si occupa successivamente di valutare caso per caso sulla base del bilanciamento tra il diritto dell'interessato alla *privacy* e il diritto all'informazione di tutti gli altri utenti.

Fare richiesta al motore di ricerca dell'eliminazione di *link* a notizie della tipologia che Google già annovera tra quelle pubblicamente rilevanti, comporta quasi sempre un rifiuto istantaneo. Il rifiuto di Google è insindacabile e il ricorrente può soltanto rivolgersi al Garante della *privacy* o agire per vie legali.

La totale discrezionalità del motore di ricerca nella valutazione delle richieste di rimozione può essere scavalcata soltanto con l'intervento di un'autorità giudiziaria. Per il resto, Google è giudice di se stesso. Vale a dire che la sentenza del 13 maggio 2014 ha sì imposto ai motori di ricerca di dare alla possibilità agli utenti di richiedere direttamente "l'oblio", ma ha lasciato carta bianca sulla valutazione delle richieste e, quindi, sul bilanciamento tra *privacy* e cronaca.

I risultati della sentenza in merito ai provvedimenti presi da Google hanno lasciato perplesso il New York Times, che in un'inchiesta dell'aprile 2016[5] ha denunciato che delle 418.000 richieste di rimozione ricevute (circa 572 al giorno), ne siano state accolte meno della metà.

[5]Scott, *Europe Tried to Rein In Google. It Backfired*, cit.

2.3 Autocomplete, Ricerche Correlate e reato di diffamazione perpetrato dal motore di ricerca

2.3.1 La diffamazione automatica dei motori ricerca

Con ordinanza del 21-25 gennaio 2011[6], il Tribunale di Milano accoglie il ricorso di un imprenditore operante nel settore finanziario, la cui attività per altro si avvale ampiamente di servizi di *web advertising*. Il ricorrente ha scoperto che, digitando il proprio nome e cognome nella barra di ricerca di Google, nella finestra di Autocomplete che si apre immediatamente sotto di essa nel momento in cui inizia a scrivere, compaiono, associate al suo nome, le parole "truffa" e "truffatore". La Corte ordina a Google di rimuovere dal proprio *software* l'associazione tra il nome del ricorrente e le parole "truffa" e "truffatore" e fissa una somma per ogni giorno di ritardo in caso di ottemperanza all'ordine impartito. Secondo il Tribunale l'associazione generata dal *software*, anche se in via automatica, risulta essere lesiva dell'onore e della reputazione della persona nominata e si presta ad indurre l'utente a non continuare la ricerca, sospettando che questi conduca attività illecite o di dubbia moralità. Il fatto che, una volta accettato il suggerimento, non compaiano documenti dal contenuto offensivo, non può essere ritenuto rilevante per giudicare la la natura diffamatoria delle associazioni generate.

Contro questo provvedimento, Google propone un reclamo facendo forza su tre argomentazioni. In primo luogo, deduce l'errata interpretazione, da parte del giudice di prime cure, del funzionamento di Google Autocomplete: le associazioni infatti non sono generate dal software, ma sono il risultato delle ricerche più popolari effettuate dagli utenti. In secondo luogo sostiene che la motivazione dell'ordinanza in ordine alla responsabilità di Google come *internet service provider* (ISP) è errata e insufficiente: Google ha ampiamente dimostrato, in fase cautelare, la sua natura di ISP e lo stesso ricorrente riconosce che si tratti di un *host provider*, il quale non ha, secondo il d.lgs 70/2003, alcuna responsabilità per contenuti immessi da terzi che ospita. Infine, è errata e insufficiente anche la motivazione dell'ordinanza in merito al pregiudizio subito dal ricorrente: una piattaforma di hosting e i servizi forniti per migliorarne l'usabilità sono di per sé neutri: un potenziale elemento dannoso per la reputazione può essere ricondotto esclusivamente a contenuti

[6]Riferimenti tratti da S. Peron, *Sulla diffamazione commessa tramite motore di ricerca*, "Responsabilità civile e previdenza", 117, n. 6, 2011, pp. 1327 – 1335.

immessi da terzi e perciò fuori dalla responsabilità dell'ISP. Inoltre, l'idea che l'utente potrebbe farsi di primo acchito, nel leggere le associazioni infamanti, sarebbe smentita immediatamente una volta letti i contenuti inoffensivi delle pagine consigliate dallo stesso suggerimento di ricerca. L'utente di Internet è perfettamente in grado di filtrare e di interpretare i contenuti caricati sul web da terzi e di selezionare criticamente le informazioni a sua disposizione.

In risposta al reclamo, il Tribunale anzitutto chiarisce che il procedimento attraverso il quale opera Autocomplete è assolutamente chiaro al collegio, cioè che si tratti di un servizio svolto in via automatica da un software che mediante un algoritmo matematico raccoglie i termini di ricerca immessi da utenti precedenti e li restituisce in ordine di popolarità. Quando viene usata l'espressione «associazione creata dal *software*», si intende proprio associazioni elaborate dal software attraverso un filtro di termini di ricerca maggiormente utilizzati dagli utenti. Tuttavia Google, oltre ad essere un *host provider*, cioè un soggetto che si limita ad offrire ospitalità ad un sito *web* che rimane sotto la responsabilità di coloro che lo gestiscono in autonomia sui propri *server*, fornisce servizi di motore di ricerca.

Un motore di ricerca è fondamentalmente un *database* di testi, immagini, video e altri contenuti indicizzati che gli utenti possono reperire interrogando il motore di ricerca stesso, solitamente inserendo le parole chiave dell'informazione che stanno cercando. In poche parole un motore di ricerca è come un grande indice analitico: ad ogni parola dell'indice corrispondono una quantità di testi che la contengono, di cui il motore di ricerca può fornire la "collocazione" (cioè il *link*). Per fare questo i motori di ricerca si avvalgono di vari strumenti per intermediare informazioni: una piattaforma tecnologica, dei *database* e dei *software* (*spider*). L'attività di un motore di ricerca è quindi diversa da quella di un *host provider*, perchè non si limita solo ad "ospitare" contenuti, ma ha un ruolo attivo sul trattamento di questi nel momento in cui li elabora e, in qualche modo, li interpreta per presentarli agli utenti.

Nella legislazione italiana la responsabilità degli *internet service provider* è regolamentata del decreto legislativo 70/2003, che distingue la responsabilità in base ai servizi offerti: *mere conduit*, cioè semplice trasporto di informazioni (art. 14), *caching*, ovvero attività di memorizzazione temporanea di informazioni (art. 15) e *hosting*, ossia memorizzazione permanente di informazioni (art. 16). In quest'ultimo caso il *provider* non è responsabile delle

informazioni memorizzate a richiesta dell'utente, ma a due condizioni: che esso non sia effettivamente a conoscenza di attività o informazioni illecite e che, non appena lo sia, agisca immediatamente per rimuovere le informazioni o disabilitarne l'accesso.

L'articolo 17 del d.lgs 70/2003 statuisce che «il prestatore non è assoggettato ad un obbligo generale di sorveglianza sulle informazioni che trasmette o memorizza, né a un obbligo generale di ricercare fatti o circostanze che indichino la presenza di attività illecite»[7]. Qualora però il *provider* abbia conoscenza di presunte attività o informazioni illecite riguardanti un destinatario dei suoi servizi, ha l'obbligo di informare tempestivamente l'autorità giudiziaria o quella amministrativa di vigilanza ed è tenuto a fornire le informazioni in suo possesso idonee ad identificare il destinatario dei suoi servizi, al fine di individuare e prevenire attività illecite[8].

Il Tribunale di Milano ha escluso che in questo caso fosse applicabile la normativa del d.lgs 70/2003, dato che essa riguarda l'attività tipica dell'*host provider*, attività assolutamente non messa in discussione ma non pertinente a ciò che il ricorrente lamenta, ovvero l'abbinamento del suo nome con le parole "truffa" e "truffatore". Questa associazione è esclusivamente frutto della specifica modalità operativa del servizio Autocomplete, un *software* messo a punto da Google in quanto motore di ricerca. Poichè l'associazione del nome del ricorrente con le parole offensive è esclusivamente opera del *software* e non riguarda i contenuti resi accessibili da Google, ne consegue che gli eventuali effetti negativi che tale sistema può determinare sono sotto diretta responsabilità di Google.

Per quanto riguarda il problema sollevato da Google, secondo il quale nell'eventualità che si censurasse a posteriori questo tipo di associazioni, si limiterebbe il diritto di libertà di espressione e libera informazione nei confronti degli utenti, la Corte risponde che la rimozione delle associazioni diffamatorie non limiterebbe in alcun modo la libertà degli utenti di accedere alle informazioni: il servizio Autocomplete infatti non compie alcun intervento sui contenuti memorizzati nel Web, ma li elabora per offrire un'agevolazione agli utenti che in ogni caso, anche se i risultati di questa elaborazione fossero modificati o eliminati, avrebbero comunque libero accesso a tali contenuti. È irrilevante poi il fatto che, trattandosi di un *software* completamente au-

[7] art. 17 comma 1 d.lgs 70/2003.
[8] art. 17 comma 2 d.lgs 70/2003.

tomatico, sarebbe impossibile fare una cernita a priori di termini "buoni" o termini "cattivi": ciò che viene richiesto alla società di Google non è il controllo preventivo sui dati nel sistema, ma quello successivo, a posteriori, sui risultati della sua operatività una volta segnalate le associazioni offensive.

A seguito del reclamo di Google, il Tribunale riconferma la valutazione del giudice di prime cure, secondo il quale la semplice associazione tra nome del ricorrente e parole "truffa" e "truffatore" è da considerarsi diffamatoria. Per quanto il delitto di diffamazione può essere punito solo in caso di dolo, cioè se si usano volontariamente espressioni offensive con la consapevolezza che queste ledano l'altrui reputazione, l'illecito diffamatorio può in ogni caso essere censurato anche semplicemente in sede civile (art. 2043), anche se il fatto non costituisce reato: in ambito civile infatti, a giustificare la pretesa del risarcimento non è la commissione di un fatto costituente reato, ma l'illiceità del comportamento di cui quel danno è derivato. L'esistenza o meno del dolo può anche non essere considerata in un giudizio civile, come nel caso in esame in cui, alla società titolare del motore di ricerca, non si può attribuire un vero e proprio dolo, ma soltanto una colpa . A tale proposito il Tribunale non ritiene di poter condividere la tesi della reclamante secondo la quale la suggestione data dall'abbinamento iniziale delle parole "truffa" e "truffatore" al nome del ricorrente sarebbe comunque subito eliminata dalla lettura dei contenuti inoffensivi del materiale raccolto all'interno della ricerca stessa. Questi contenuti, infatti, non sono immediatamente visibili dall'utente, che dovrebbe prima accettare il suggerimento (per altro fuorviante) per accedere al contenuto e leggerlo. Per farlo dovrebbe essere indotto da qualche interesse, in assenza del quale non approfondirebbe la ricerca e gli resterebbe impressa la prima immediata impressione negativa che l'associazione di parole genererebbe.

2.3.2 Affermazione o domanda

Non è dunque per la sua responsabilità come ISP che Google viene sanzionata, ma per il danno che il prodotto di un suo *software* arreca, anche in assenza di dolo, alla persona nominata, facendo scattare l'applicazione dell'art. 2043 del Codice Civile. L'oggetto del ricorso non è l'attività di *host provider*, e quindi di ISP, di Google, ma il risultato della modalità di un servizio messo a punto da questo. La Corte respinge in toto il reclamo di

Google, soprattutto la tesi secondo la quale non avrebbe considerato la sua natura di ISP e quindi delle sue relative responabilità.

Curiosamente, due anni dopo la stessa Corte di Milano si trova ad affrontare una situazione pressoché identica, ma questa volta assolve Google proprio ai sensi degli artt. 15-17 del d.lgs 70/2003. Secondo il Tribunale, i termini che Autocomplete associa al ricorrente (che in questo caso è il presidente di due associazioni *no profit* che ha trovato il proprio nome associato a "truffa", "truffatore", "plagio" e "setta") non costituiscono un archivio, non sono strutturati, organizzati o influenzati da Google, il quale si limita ad analizzarne la popolarità in base ad un *software*: si tratta di un'attività di *caching*, cioè memorizzazione provvisoria e automatica di dati di cui non è responsabile, come non lo è il sistema di ricerca né i suoi risultati[9].

Per di più, i termini isolati non costituiscono una frase di senso compiuto, né rappresentano quello che Google pensa: l'accostamento di termini in una stringa o un profilo di ricerca non costituisce un'affermazione bensì una suggerimento di ricerca sulla base di dati statistici o indicizzati presenti nella memoria di Google e quindi non può essere considerato atto diffamatorio.

Gli esiti così diversi di due casi molto simili rivelano come la legislazione italiana non fornisca una direttiva definita riguardo al trattamento dei contenuti che servizi come Autocomplete applicano, lasciando ai giudici l'interpretazione della natura di questi *software* e dei loro risultati. Il problema di fondo è che è difficile definire cosa effettivamente rappresenti la stringa di suggerimento generata da Autocomplete: mentre nel processo del 2011 ne era stato valutato soltanto il danno, ma non l'identità, in quello del 2013 viene considerato come suggerimento del tutto neutrale e assolutamente non definibile come affermazione né tantomeno come contenuto generato attivamente da Google. Mentre nel primo caso il ricorrente riconosceva a pieno il fatto che Google rappresentasse un *service provider* e che dunque non fosse responsabile dei contenuti che mette a disposizione, nel processo del 2013 la parte offesa sostiene che, dal momento che la funzione di autocompletamento è messa a punto da Google, questo vada considerato un content provider, in quanto il contenuto visualizzabile tramite questi servizi è prodotto e diffuso dallo stesso Google. La Corte di Milano tuttavia respinge la qualità di content provider di Google e ribadisce la neutralità dei risultati di Autocomplete.

[9]Trib. di Milano, 25/3/2013, n. 68306.

È sensato considerare la stringa generata automaticamente da un *software* un contenuto? Sicuramente, anche se non costituisce una frase di senso compiuto, una stringa come "Mario Rossi truffatore" lascia bene intendere un certo significato e può infondere in chi la legge qualche sospetto. La stringa inoltre è effettivamente generata, anche se in via automatica e sull'elaborazione statistica di altri contenuti, dal *software* di Google ed è dunque da considerare un suo esclusivo prodotto. Ma anche se si volesse considerare queste stringhe come veri e propri contenuti e quindi Google come *content provider*, resta il fatto che, indipendentemente dalla compiutezza del suo significato, la stringa di Autocomplete è tecnicamente una *query*, vale a dire una domanda. Di una domanda quindi, e non di un'affermazione, può essere ritenuto responsabile Google e una domanda non può essere, per sua stessa natura, foriera di cattiva reputazione. Se poi i contenuti accessibili tramite queste *query* confermano ciò che nelle *query* è formulato soltanto come ipotesi, la cattiva luce sotto cui ricadrà l'interessato non può essere riconducibile alla responsabilità del motore di ricerca.

Capitolo 3

Monitoraggio e analisi dell'identità sui motori di ricerca

3.1 Il *check-up* reputazionale: considerazioni preliminari

Il primo passo per poter assumere il controllo dell'identità digitale è imparare a conoscerla. Il gesto è tecnicamente molto semplice: accedere a un *browser*, collegarsi a `www.google.com` (o `.it`), digitare il nome e cognome del soggetto di cui si vuole analizzare l'identità. Quella pagina dei risultati è un vero e proprio profilo, come un profilo Facebook o Twitter, una bacheca che offre pubblicamente informazioni, contenuti e notizie riferibili a quel nome e cognome (non necessariamente al soggetto, ma questo può non essere immediatamente chiaro all'utente). La differenza fra qualsiasi profilo e quello prodotto da Google è che quest'ultimo non è sotto la responsabilità di nessuno, se non dell'algoritmo che l'ha generato.

Monitorare la pagina dei risultati di Google (SERP) significa raccogliere i dati che ci permettano di prevedere l'esito di quella che nel primo capitolo è stata definita "stretta di mano digitale". È chiaro che il monitoraggio della SERP soltanto non basta ad ottenere uno spettro completo dell'identità digitale: la SERP si limita a rappresentare un sommario non ragionato e dall'accuratezza poco controllabile della presenza del soggetto su tutti i canali del *web*, come i *social*, i *blog*, *webzine*, contenuti video, foto ecc. Tuttavia,

l'analisi e il monitoraggio limitati al motore di ricerca merita particolare interesse per due ragioni: la prima è che corrisponde al gesto più basilare ed immediato dell'accesso a Internet per cercare informazioni su qualcosa o qualcuno; la seconda sta nel fatto che l'identità digitale restituita dalla SERP è, per certi versi, la rappresentazione di cosa il motore di ricerca "conosce" di una persona, cosa risponde alla virtuale domanda «chi è Mario Rossi?».

Dopo una prima ispezione della SERP è probabile che l'identità digitale del soggetto ricada genericamente in una delle seguenti quattro categorie:

- **Identità positiva**: compaiono solo risultati positivi ed in linea con le aspettative del soggetto.
- **Identità compromessa**: compaiono uno o più risultati non positivi o non desiderati dal soggetto. Purtroppo basta anche solo un risultato negativo, che si trova in una posizione di alta visibilità, per compromettere l'intera immagine del soggetto.
- **Identità non pertinente**: compaiono risultati non riferibili al soggetto per ragioni per lo più riconducibili a casi di omonimia.
- **Identità nulla**: compaiono risultati che non forniscono alcuna informazione sul soggetto, causa la scarsissima presenza *online* dello stesso. Per quanto questo tipo di categoria si traduca poi nell'identità non pertinente, è uno scenario possibile e benché possa sembrare la condizione più idonea alla salvaguardia della reputazione, è esposta a molti rischi, capiremo in seguito perché.

Questa prima classificazione è molto generica e ha molte zone grigie: anche se un'identità è complessivamente positiva, per esempio, non è detto che contenga solo e soltanto contenuti riferibili, così come un'identità non pertinente può dare luogo a identità compromessa. Soprattutto, non è facile fare una netta distinzione tra identità non pertinente e identità nulla in quanto, eccetto casi estremi, il motore di ricerca restituirà sempre qualcosa una volta interrogato e dunque un profilo relativo a quel nome, sia pur nebbioso e fuorviante, si delineerà in ogni caso.

Se da un lato è quasi banale, da un punto di vista intuitivo, capire se un soggetto gode di una buona reputazione *online* o meno, non è altrettanto facile stabilire dei criteri oggettivi per eseguire un'analisi sistematica e quanto più possibile scientifica dell'identità digitale.

Nell'analisi vanno presi in considerazione vari fattori: i primi due sono fondamentali, positività/negatività e riferibilità/non riferibilità dei risultati. A questi si aggiungono la tipologia prevalente di contenuto - come testo, video, immagini - l'effettivo controllo che il soggetto esercita su ogni contenuto, se questo deriva da una piattaforma *social* e se si riferisce a vita privata, professione o formazione.

Questo tipo di attività richiede metodo, capacità di osservazione e soprattutto costanza nel tempo. Monitorare una volta sola o di rado l'identità digitale non crea le condizioni per poter reagire tempestivamente e con i giusti mezzi all'insorgere, potenzialmente quotidiano, dei contenuti. I motori di ricerca sono realtà molto dinamiche, variano e aggiornano i propri indici continuamente, così come i loro criteri di "risposta" alle richieste dell'utente. La periodicità è un aspetto fondamentale del monitoraggio dello stato di salute della reputazione sulla Rete.

Definire i criteri base dell'analisi della reputazione *online* non è semplice perché è lo stesso concetto di reputazione a non essere facilmente scomponibile in una serie finita di variabili universalmente definibili, alle quali assegnare valori quantificabili. La reputazione è una previsione d'esito di una relazione: conoscendo poco o niente di una persona, un individuo effettua una proiezione su come potrebbe svolgersi un incontro, una conversazione o una collaborazione con lei, basandosi sulle informazioni che raccoglie da fonti esterne. Questa previsione è spesso molto istintiva, si basa su associazioni che facciamo in pochi istanti tra la persona in questione e il contesto in cui è immersa.

Cercare di sistematizzare il processo di analisi e fornire uno schema di valutazione reputazionale universale è rischioso, dal momento che quest'attività oscilla tra il semplice buon senso e le necessità peculiari di ogni singolo caso. Tuttavia, è necessario individuare quali aspetti della presenza sui motori di ricerca devono essere considerati in relazione alla reputazione personale, secondo i quali classificare le informazioni e i dati materialmente reperibili ed estraibili dalla SERP che devono poi essere utilizzati nel processo di analisi, indipendentemente dai criteri scelti per eseguire quest'ultima.

L'identità digitale infatti è il risultato dell'interazione di più forze, come per esempio la visibilità di un contenuto, il contesto in cui è calato o che esso stesso genera, la sua potenziale capacità di aprire casualmente nuovi spunti di ricerca e approfondimento, il suo grado di attualità o obsolescenza.

Una volta compresi tutti gli aspetti della reputazione *online*, il passo successivo è conoscere l'ambiente di lavoro, cioè la SERP, la sua struttura anatomica, il suo comportamento e il suo funzionamento. Capire la SERP ci aiuta a capire chi la usa e dunque cosa può trovarci dentro. Ci sono tanti strumenti più o meno sofisticati per misurare e valutare la *web presency*, ma niente è più efficace che imparare ad utilizzare il motore di ricerca così come si presenta al pubblico generico. Inoltre l'analisi di tutti gli elementi che compongono la SERP è necessaria per capire dove andare a cercare i dati da classificare secondo gli aspetti della reputazione *online* definiti precedentemente.

Compreso il motore di ricerca nella sua incarnazione più immediata (e più usata dall'utenza media), si può passare ad un utilizzo più avanzato, che mira non tanto a monitorare ciò che si trova sulla SERP di un generico nome e cognome, ma a setacciare gli indici di Google per trovare qualsiasi contenuto riferibile al soggetto. Google infatti offre degli Operatori di Ricerca Avanzata che permettono di affinare la ricerca secondo criteri di filtraggio, ordinamento e relazioni fra parole chiave che permette di reperire informazioni meno facili da trovare con l'interrogazione standard.

Per rendere l'indagine ancora più sofisticata, è possibile costruirsi manualmente richieste HTTP sfruttando i parametri del protocollo di ricerca di Google Search da inviare al motore di ricerca.

Questi strumenti di ricerca consentono già un alto livello di monitoraggio, ma hanno un limite: richiedono un lavoro manuale consistente. Sia che si faccia una scansione di una SERP, che si usino operatori di ricerca o si costruiscano richieste HTTP *ad hoc*, la raccolta di dati va effettuata a mano e ciò richiede tempo e metodo. Questo aspetto genera due difficoltà: il monitoraggio su lungo periodo e la normalizzazione dei dati.

Come già accennato in precedenza, i motori di ricerca sono realtà in continuo movimento e necessitano di un controllo periodico, soprattutto se si vuole tracciare l'andamento della reputazione nel tempo o si ha necessità di verificare l'efficacia di interventi migliorativi/riparatori sull'identità di un soggetto sulla SERP[1].

In secondo luogo, è probabile avere la necessità di avere dati normalizzati e pronti per essere utilizzati su strumenti di calcolo come per esempio uno *spreadsheet*. Intabellare dati a mano richiede molto tempo e può dar luogo

[1] Come è stato fatto nel caso di studio riportato nel capitolo 6.

ad errori. Per questo, uno strumento di raccolta automatica di dati sarebbe di gran lunga preferibile alla raccolta manuale di informazioni.

3.2 I sei aspetti della reputazione sui motori di ricerca

La principale ragione della difficoltà di stabilire una serie di criteri universali per valutare della reputazione sul motore di ricerca, sta nel fatto che lo stesso contenuto può giocare un ruolo molto positivo o molto negativo a seconda dei casi: una *rock star* e un politico avranno modi diversi di reagire a uno scandalo riverberatosi sui canali digitali.

Tuttavia si può provare ad individuare gli aspetti che insieme costituiscono lo spettro informativo, valoriale e relazionale di un soggetto. In parole più semplici: quali sono i fattori che determinano l'influenza (negativa o positiva che sia, ma questo dipende da ogni caso specifico) che un risultato di ricerca esercita sull'immagine di un soggetto.

Anche in questo caso, i fattori possono essere molti, variare da soggetto a soggetto e non essere sempre facilmente definibili, per questo si è preferito utilizzare il termine "aspetti" anziché "fattori": qualcosa definito "fattore" dovrebbe essere coinvolto in qualche tipo di calcolo scientifico, ma nell'analisi della reputazione solo per alcuni tipi di dato questo è possibile:

- **visibilità**: quanto un contenuto, una notizia, un nome è visibile sulla SERP,
- **accessibilità**: la possibilità di un utente di accedere ad un contenuto e dunque apprendere l'informazione che riporta,
- **contesto**: in quale contesto l'informazione si trova o che essa stessa genera,
- **serendipità**: la possibilità di imbattersi in una determinata informazione mentre se ne sta cercando un'altra,
- **tempo**: la data in cui un contenuto è stato creato e indicizzato dal motore di ricerca,
- **responsabilità**: le figure coinvolte nella produzione, pubblicazione e condivisione del contenuto

.

Vediamo di seguito nel dettaglio ciascuno di questi aspetti.

3.2.1 Visibilità

Chi si occupa di *web marketing* sa bene che non basta "essere su Internet" per essere trovati. L'effettiva facilità e immediatezza con cui una risorsa *web* può essere individuata da un utente ne stabilisce la sua effettiva esistenza. Se un albero cade in una foresta ma nessuno lo sente, c'è da chiedersi se ha fatto davvero rumore, ma è quasi certo che se un risultato non si trova nelle prime tre pagine della SERP, quel risultato non sarà mai trovato. Questo significa, per l'utente medio di Google, che quel contenuto non esiste.

Gli utenti vanno sempre più di fretta, navigano il Web distrattamente, dedicano poco tempo alle ricerche che eseguono abituati ad avere le risposte che vogliono sull'istante (e non a caso, Google ha provveduto a questa necessità con il suo servizio di Istant Search), quasi mai hanno tempo o voglia di approfondire una notizia e neanche verificare che sia vera o falsa.

Probabilmente, siamo ancora legati a un concetto "giornalistico" dell'esposizione dell'informazione, per cui le notizie più importanti sono in prima pagina e riportano titoli che occupano metà del foglio di carta. La rilevanza grafica di un contenuto ci fa pensare automaticamente che non solo sia il più pertinente, ma anche il più autorevole, più affidabile e, soprattutto, il più aggiornato (proprio perché i "titoloni" riportano le ultime notizie).

Per quanto gli algoritmi di Google siano sempre più sofisticati nella valutazione della rilevanza e della qualità di un contenuto, non bisogna dimenticare che non c'è nessun direttore editoriale ad allestire la SERP di Google e che i criteri di rilevanza del motore di ricerca sono diversi da quelli di un giornalista. Se Mario Rossi è stato condannato per truffa in primo grado ma assolto in secondo, ed entrambi le notizie sono sul Web, non è scontato che in un'ipotetica ricerca "processo Mario Rossi" l'articolo dell'assoluzione compaia prima dell'articolo della condanna: se la pagina *web* che riporta la notizia della condanna in primo grado ha molti *backlink* (cioè è stata "linkata" da altri siti) e contiene molte parole chiave è probabile che compaia nei primi risultati. Il secondo, magari meno ottimizzato per i *crawler* del motore di ricerca, potrebbe comparire alla decima o quindicesima posizione o persino non comparire mai. Per chi si limita a guardare i primi risultati (quasi tutti gli utenti), Mario Rossi è tutt'ora condannato per truffa.

I numeri parlano chiaro: secondo una celebre ricerca[2] di Chitika[3], purtroppo risalente al 2013 e non più aggiornata, il 91,5% degli utenti si ferma alla prima pagina della SERP, il 4,8% si spinge alla seconda e solo l'1,1% si sforza di arrivare alla terza pagina. Il restante 2,6% va oltre la terza pagina dei risultati, ma come si può immaginare la percentuale si abbassa sempre più che il numero di pagina cresce. Questo significa che se un contenuto si trova in trentunesima posizione, cioè in cima alla quarta pagina sella SERP, per 97 persone su 100 non esiste.

In Tabella 3.1 i cui dati sono tratti sempre dalla ricerca di Chitika, sono rappresentate le percentuali di traffico su ogni singolo risultato per i primi 20 (che corrispondono alle prime 2 pagine).

Tabella 3.1: Il valore della posizione dei risultati di ricerca sulla SERP di Google

Rank Google	**% traffico**
1	34,35%
2	16,96%
3	11,42%
4	7,73%
5	6,19%
6	5,05%
7	4,02%
8	3,47%
9	2,85%
10	2,71%
11	1,11%
12	0,85%
13	0,70%
14	0,57%
15	0,48%
16	0,39%
17	0,33%
18	0,28%
19	0,27%
20	0,29%

L'aspetto visibilità non consiste solo nella posizione di un risultato nella

[2]Chitika, *The Value of Google Result Positioning*, "Chitika.com", 7 giugno 2013, https://chitika.com/google-positioning-value, 19/4/2017. Dati aggiornati al 6 dicembre 2013.

[3]Importante agenzia americana per la pubblicità *online* e *mobile*.

SERP: titolo e *snippet* sono la prima cosa che l'utente legge del documento a cui il risultato si riferisce. Ciò che compare in titolo e *snippet* gioca un ruolo determinante nella decisione che l'utente di collegarsi e leggere quel contenuto.

Inoltre, foto, video presenti sulla SERP nelle ricerche verticali o nel Google Graph possono attirare l'attenzione prima del testo dei risultati. I contenuti multimediali sulla SERP corrispondono alla foto su un *curriculum vitae* o su un passaporto.

La questione della visibilità è una dei protagonisti della disciplina della *online reputation management*. Se da un lato detiene l'indice massimo di pericolosità nel caso metta in evidenza contenuti compromettenti, dall'altro è il fattore su cui è più facile intervenire in caso di crisi reputazionale, a patto di conoscere le tecniche giuste.

3.2.2 Accessibilità

Un contenuto deve essere, oltre che visibile, accessibile perché possa influenzare la reputazione di un'identità digitale. Una notizia può essere presente sul Web ma non essere facilmente accessibile perché magari non è indicizzata dai motori di ricerca e nessun'altra pagina la punta con un collegamento ipertestuale e dunque per poterla raggiungere l'utente dovrebbe conoscerne l'esatto indirizzo. In un caso del genere (piuttosto raro) indipendentemente dalla positività o negatività del contenuto della pagina, questa avrà un impatto pressoché nullo sull'identità digitale di un soggetto, proprio perché molto difficilmente qualcuno si imbatterà in essa.

In quest'ottica, la stessa visibilità costituisce un importante fattore per l'accessibilità: una pagina poco visibile sarà anche poco accessibile, un *link* che si trova alla trentesima posizione della SERP ha meno probabilità di essere visto e quindi il suo contenuto è meno accessibile rispetto a un *link* in prima o seconda posizione.

Ci sono però dei casi in cui una pagina *web* è visibile sulla SERP ma non accessibile: si tratta di casi in cui una pagina è stata rimossa, per una qualsiasi ragione, dal *server* in cui si trovava, ma il suo indirizzo si trova ancora negli indici di Google.

Questo significa che quella pagina compare sulla SERP, magari anche in posizioni ad alta visibilità, ma cliccandoci sopra restituirà un errore 404 o

una risposta HTTP di tipo 400 o 500 a seconda se il problema di accessibilità del contenuto è lato *server* o lato *client*.

Conoscere l'effettiva accessibilità di un contenuto visibile sulla SERP ha un'importanza piuttosto rilevante: e pagine non accessibili, in particolare quelle cancellate e quindi non più esistenti, sono considerate informazioni obsolete da parte di Google. Il loro status di obsolescenza fa sì che Google sia disposto di buon grado a rimuoverle dai propri indici, tant'è che ha messo a disposizione un apposito modulo per la segnalazione di URL obsolete per poterle poi prontamente rimuovere.

3.2.3 Contesto

Il problema del contesto è un altro aspetto principe della reputazione *online*. Il fatto che il nome e cognome o la fotografia di un soggetto compaia in un determinato contesto influisce sull'opinione che gli altri possono avere su di lui. È il solito discorso della birreria e dell'ufficio di lavoro: avere una birra in mano nel primo contesto è pienamente accettabile e non deprecabile, ma non lo può essere nel secondo.

La decontestualizzazione è una delle cause più frequenti e più gravi della lesività reputazionale e il vero diritto che manca in rete non è tanto il diritto ad essere dimenticati, ma quello di essere contestualizzati.

Monitorare in quali contesti il nome di un soggetto compare, o quali contesti compaiono durante la ricerca del suo nome e cognome su Google è di primaria importanza.

Le informazioni visibili sulla SERP danno luogo ad un panorama più o meno completo e veritiero dello spettro valoriale e relazionale del soggetto sintetizzato dai contesti in cui il suo nome compare.

È evidente come il fattore visibilità e il fattore contesto interagiscano l'uno con l'altro nella costruzione della reputazione del soggetto: un contesto che nella realtà ha un basso grado di relazione col soggetto ma che ha grande visibilità sulla SERP assume una rilevanza prepotente sulla reputazione personale. Se per esempio Mario Rossi compare in un articolo di cronaca che riguarda un processo per truffa in cui è coinvolto solo marginalmente ma questo articolo compare nelle prime posizioni, oppure il suo nome compare in titolo o *snippet* di un risultato per la ricerca di quel fatto di cronaca, poco importa se poi leggendo l'articolo fino in fondo si scopre che il ruolo di

Mario Rossi nella vicenda è marginale o persino positivo. Per l'utente pigro e frettoloso Mario Rossi è coinvolto fino in fondo in un caso di truffa.

3.2.4 Serendipità

"Serendipità" è un termine coniato dallo scrittore Horace Walpole (*serendipity*) ispirandosi alle avventure di tre personaggi della fiaba persiana *Tre principi di Serendippo*, i quali si salvano in diverse situazioni grazie a scoperte fatte per caso, mentre cercavano cose ben diverse.

Wikipedia definisce così il termine serendipità:

> lI termine serendipità è un neologismo che indica la fortuna di fare felici scoperte per puro caso e, anche, il trovare una cosa non cercata e imprevista mentre se ne stava cercando un'altra.[4]

L'esempio più celebre di serendipità nella storia è il caso di Cristoforo Colombo: mentre cercava il Giappone ha trovato l'America. Una "felice scoperta" per il navigatore genovese e per tutto il Vecchio Continente, ma c'è da chiedersi se per Pellerossa e Amerindi quella scoperta sia stata altrettanto felice.

La serendipità trova la sua massima espressione nella struttura reticolare della conoscenza tipica del Web. Mentre navighiamo in Internet siamo tutti potenziali Cristoforo Colombo, ma anche potenziali Pellerossa: le molteplici accuse a Google per diffamazione perpetrata attraverso Autocomplete e Ricerche Correlate lo dimostrano.

Autocomplete, Ricerce Correlate e Istant Search sono i più efficaci stimolatori di serendipità sul motore di ricerca: relazioni tra persone, fatti, luoghi, aggettivi si intromettono nella ricerca di qualcosa nel momento stesso in cui l'utente sta formulando l'oggetto della sua ricerca o immediatamente dopo, aumentando notevolemnte il rischio di "dirottamento" rispetto a ciò che aveva intenzione di cercare.

In Autocomplete le *query* visualizzate rispecchiano l'attività di ricerca degli utenti e i contenuti delle pagine *web* indicizzate da Google. In poche parole offre una sintetica panoramica di cosa gli altri hanno cercato e ciò che esiste riguardo all'argomento della *query*. Si tratta di una «efficace e affilata macchina di correlazione che influenza in modo forte la ricerca nella

[4]Wikipedia, voce *Serendipità*, https://it.wikipedia.org/wiki/Serendipità, 18/4/2017.

sua costruzione guidando gli utenti direttamente ai punti chiave saltando la keyword semplice (nome e cognome)»[5].

Ricerche Correlate è molto simile ad Autocomplete, anche se, rispetto al precedente, tende più ad offrire consigli di ricerca il più possibile pertinenti con la *query* originale e non con quello che cercano gli altri utenti. Anche in questo caso il potere di dirottare la ricerca originaria è molto forte.

Grazie a Istant Search, l'utente può vedere, durante la digitazione della sua *query*, comparire risultati basati su ipotesi probabili di composizione in modo che questi possa farsi un'idea istantaneamente di cosa potrebbe trovare e aggiustare di conseguenza i termini. Questa caratteristica di costruzione progressiva della ricerca rende più probabile la composizione di *query* più complesse di quelle che si aveva in mente e dunque intensificare il grado di serendipità della ricerca.

3.2.5 Tempo

Questo aspetto è molto controverso quando si parla di "notizia" sul Web e si lega anche molto ai fattori visibilità e contesto. La data di pubblicazione di un contenuto è solo uno dei fattori di *ranking* utilizzati da Google per stabilirne la sua rilevanza in base ad una parola chiave e, purtroppo in materia di diritto all'oblio, non uno dei più importanti.

Il fattore tempo è elemento essenziale nella definizione di oblio, con tutte le controversie che si trascina dietro nel dibattito sull'effettiva possibilità di stabilire una data di scadenza all'informazione, dopo la quale la comunità non ha più diritto ad accedervi. In secondo luogo, il tempo è una delle potenziali cause di lesività reputazionale: alcuni contenuti, foto o fatti sono accettabili in un determinato lasso temporale. Ciò che è appropriato oggi potrebbe non esserlo domani. Ogni contenuto ha una data di pubblicazione, ogni pagina indicizzata ha una data di indicizzazione. La media delle età dei risultati di ricerca è l'età della nostra immagine digitale.

3.2.6 Responsabilità

Capire quali persone sono coinvolte nella circolazione di un contenuto *online* è importante nel momento in cui si voglia prendere provvedimenti di rimozione di quel contenuto, soprattutto se si è deciso di agire per vie legali.

[5]Barchiesi, *La tentazione dell'oblio*, cit., pp. 114 - 115.

Il problema della responsabilità del contenuto online è più complesso del contenuto *offline*. Nella comunicazione tradizionale ci sono solo due attori nella diffusione di una notizia: l'autore del contenuto e chi lo pubblica.

Con i media digitali tutto cambia, soprattutto nel Web 2.0 dove spettatore e autore si fondono in una figura sola non definita e incontrollabile. Nel contenuto *online* entrano in gioco quattro figure alle quali corrisponde ciascuna un ruolo, e una responsabilità, ben precisa:

1. **Produzione**: l'effettivo autore materiale del contenuto, di qualsiasi genere si tratti.

2. **Pubblicazione**: il sito *web* che pubblica il contenuto.

3. **Diffusione**: colui che riporta il contenuto su un altro dominio, come per esempio lo condivide su Facebook, lo ripubblica sul proprio blog, lo invia per email o messaggio privato ai conoscenti.

4. **Indicizzazione**: il motore di ricerca che classifica il contenuto in base a delle parole chiave, gli attribuisce un punteggio di qualità e lo rende un'entità che può essere oggetto di ricerca.

3.3 Anatomia della SERP di Google

Con SERP (Search Engine Result Page) si intende la pagina (in realtà sono più di una) nella quale compare la lista di *link* che il motore di ricerca restituisce in seguito all'interrogazione da parte dell'utente mediante la digitazione e invio di una stringa testuale che corrisponde all'oggetto della ricerca (*search query*). La posizione di un *link* nella SERP corrisponde alla rilevanza che il motore di ricerca assegna al documento corrispondente in relazione alla *search query*. Più il *link* è posto in evidenza, più è considerato rilevante.

Nel tempo Google, come anche gli altri motori di ricerca, ha reso sempre più sofisticata la presentazione dei risultati al fine di assistere al meglio l'utente in base al tipo di contenuto che sta cercando. La SERP si è evoluta negli anni da una semplice lista di URL a una struttura dati ricca e articolata, la cui complessità dipende dalla quantità e dalla rilevanza delle risorse trovate in relazione alla ricerca dell'utente.

Inizialmente la SERP conteneva un elenco di *link* a pagine *web*, corredate da alcune righe di descrizione (non sempre chiarissime ed esaustive). In seguito Google si è arricchito ed evoluto, integrando annunci a pagamento (Google AdWords), il motore di ricerca per immagini (Google Images) fino ad arrivare a mappe, *news*, video, prodotti di vendita (Google Shopping), ecc.

Nel 2007 Google annunciò Universal Search, una rivoluzione nel mondo dei motori di ricerca che consisteva nel mostrare in un'unica pagina i risultati aggregati e classificati per formato. Prima di allora, le varie proprietà di Google, come per esempio Google Images, potevano essere consultate, ma apparivano in SERP separate, selezionabili singolarmente. Iniziò così l'era dei contenuti multiformato, la SERP si trasformò da semplice pagina testuale a contenitore multimediale che mira a fornire una panoramica immediata e completa di tutto ciò che il Web offre su un determinato oggetto di ricerca.

Se si cerca il nome di un personaggio famoso, come per esempio "papa francesco", si può avere un'idea abbastanza esaustiva di quasi tutti gli elementi che possono costituire una SERP particolarmente complessa:

La ricchezza e complessità di una pagina dei risultati dipende dalla popolarità dell'oggetto della ricerca. Nella Figura 3.1 si può avere un'idea abbastanza esaustiva di cosa sia l'Universal Search e come si presenti oggi una SERP di un personaggio famoso, come quella per la ricerca della stringa "papa francesco".

I risultati di ricerca sono distribuiti su due colonne: una principale fornisce una lista di *link*, per lo più presentati nel formato "classico", cioè titolo della pagina + URL + descrizione (*snippet*), una colonna a destra invece mira a creare una sorta di *identikit* del personaggio che si è cercato. In più parti della SERP compaiono risultati classificati in base alla tipologia: vediamo infatti una sezione "Notizie" nella colonna principale, nella colonna di destra una sezione "Immagini" e una sezione "Libri". In fondo alla pagina si trovano le ricerche correlate, cioè una serie di suggerimenti di ricerca che contengono la *query* originale o parte di essa. Le ricerche correlate si trovano anche nella colonna di destra, ma consistono non tanto in variazioni della *query* originale, bensì suggerimenti su altri personaggi dello stesso calibro di Papa Francesco.

Vediamo ora nel dettaglio ogni sezione della SERP:

Figura 3.1: Pagina dei risultati di ricerca di Google per “papa francesco”

1. ***Search query box***

   È il *form* all'interno del quale l'utente inserisce la *search query*, una volta premuto invio o cliccato sulla lente il motore di ricerca restituirà tutti i risultati che compongono la SERP (grazie al servizio "Istant Search", oggi quest'ultimo passaggio manuale non è sempre richiesto, i risultati compaiono automaticamente durante la digitazione della *search query* e dunque variano in tempo reale man mano che la stringa di ricerca viene formulata).

   Mentre si digita la *query* è molto probabile che si possa osservare la comparsa di una finestra sotto la *query box* all'interno della quale compaiono suggerimenti di ricerche correlate che si modificano in tempo reale man mano che la *query* viene formulata. Si tratta del servizio "Autocomplete".

   Accanto alla *query box* si trova un l'icona di un microfono per effettuare la ricerca vocale, nella Google Image Search, al posto del microfono si trova l'icona di una fotocamera che consente di caricare un'immagine per ottenere immagini simili.

2. **Menu della ricerca verticale**

   La ricerca verticale è un tipo di ricerca che viene effettuata in base al tipo di contenuto: immagini, video, notizie, mappe, libri ecc. Seguendo le voci del menu si approderà ad una SERP costituita da soli contenuti della tipologia selezionata.

3. **Informazioni sui risultati**

   Questa sezione fornisce alcune informazioni sui risultati restituiti, come una stima approssimativa dei risultati trovati e il tempo impiegato a trovarli.

4. **Risultati sponsorizzati (*paid search*)**

   Risultati sponsorizzati dalle aziende attraverso Google AdWords.

5. **Risultati verticali**

   Sono risultati classificati in base al tipo di contenuto, come immagini, video, libri, notizie ecc. Sono presentati in modo diverso dai risultati "standard" e possono comparire sia nella colonna di sinistra che in quella di destra.

6. **Risultati naturali o organici**

   È la lista di risultati di ricerca "standard", sono ordinati in base alla rilevanza che i complessi algoritmi del motore di ricerca assegnano loro in base alla *query* dell'utente. Il *rank* di ogni risultato non dipende solo dalla rilevanza in base alla *query*, ma anche dallo storico di ricerca dell'utente (se autenticato nell'*account* di Google), dalla posizione geografica, dalla lingua e da moltissimi altri fattori più o meno noti.

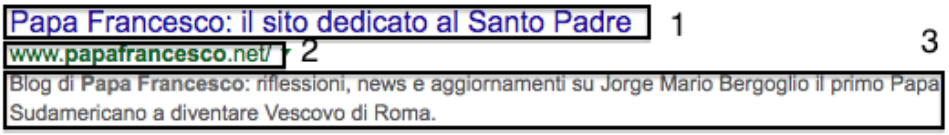

Figura 3.2: Risultato di ricerca organico standard

Un risultato organico nella sua forma standard prevede, come visibile in Figura 3.2, una struttura composta da tre elementi principali:

(a) Titolo della pagina *web*, ricavata dal *tag* `<title>` del documento HTML relativo. È anche l'*anchor text* del *link* alla pagina *web*.

(b) La URL della pagina *web*, non cliccabile.

(c) La descrizione del contenuto della pagina, se presente nella pagina HTML, il testo è estratto dal *tag* `<meta name="description">` contenuti nella sezione `<head>`, se il *tag* non è stato compilato il motore di ricerca lo estrapola solitamente dal testo del `<body>`.

7. **Ricerche correlate**

   Suggerimenti di ricerche inerenti alla *query* originale con l'obiettivo di stimolare l'utente a cercare con *query* più specifiche e possibilmente più rilevanti che possano soddisfare meglio la loro necessità.

8. **Google Knowledge Graph**

   Tradizionalmente i risultati di ricerca sono derivati dal *crawling* del motore di ricerca che analizza pagine *web*. Tuttavia, Google sta creando un *database* di informazioni che vadano oltre la tradizionale ricerca su risorse *web* di terze parti. Il Knowledge Graph, grafo della conoscenza, è un *database* strutturato di informazioni che consentono Google che permette di rispondere alle richieste degli utenti senza dover consultare pagine *web*.

   In un certo senso il Knowledge Graph è un passo successivo alla ricerca verticale, e mira a fornire risposte molto dettagliate e pertinenti direttamente nei risultati di ricerca senza che l'utente abbia bisogno di andare a visitare la pagina *web* di cui è restituito il *link*.

   Le fonti di Google per costruire il suo *database* sono state inizialmente Freebase, Wikipedia e il CIA Fact Book ma col tempo i siti di riferimento sono diventati sempre di più.

   Google Graph si presenta all'utente come un *box* nella colonna di destra o, in alcuni casi nella così detta posizione 0, cioè in cima alla lista dei risultati della colonna centrale.

   Questo *box* aggrega informazioni legate alla ricerca effettuata, solitamente per *query* di una certa rilevanza. Una persona comune è difficile che compaia nel Google Graph, ma nel caso di un personaggio famoso o un artista è probabile che compaia una foto con piccola biografia (anno di nascita, morte, studi effettuati ecc.), un elenco di immagini e altre informazioni che in qualche modo sono collegate a questa persona.

3.4 Ricerca avanzata con Google Search

3.4.1 Operatori di ricerca avanzata

Di base, gli Operatori di Ricerca Avanzata sono una serie di comandi speciali da inserire nella barra di ricerca insieme alla *query*. Questi operatori "forzano" Google a cercare in un determinato modo e secondo criteri non generici. Si tratta di uno strumento molto potente in quanto sono capaci di scremare risultati che Google tende a mostrare in modo massiccio consentendo di focalizzarsi su oggetti di ricerca molto specifici.

Il loro utilizzo esperto può essere fondamentale per trovare tutto ciò che Google ha indicizzato su un nome e cognome, andando oltre quello che la semplice ricerca standard consentirebbe di trovare.

`filetype`

Consente di restringere la ricerca a documenti come PDF, Word, Excel o Powerpoint semplicemente specificando l'estensione del documento nella ricerca in questo modo:
`mario rossi filetype:pdf`
Può essere molto utile per scoprire se esistono documenti indicizzati a un nome e cognome. Per esempio, documenti ufficiali come verbali, atti del tribunale, nonché pubblicazioni accademiche sono molto spesso reperibili in formato PDF.

`site`

Permette di trovare tutte le pagine indicizzate per una determinata *query* all'interno di un solo sito *web* (cioè un solo dominio). È possibile anche aggiungere *sub-direcotory* al dominio per restringere ancor più la ricerca.

`inurl` e `allinurl`

Restringe il campo a risultati che contengono la *query* nella propria URL. Molto utile per trovare pagine che contengono una determinata *keyword* nel proprio indirizzo. `inurl` e `allinurl` svolgono esattamente la stessa funzione, con la sola differenza che il primo esegue la ricerca per parola singola mentre la seconda per parole multiple, vale a dire che cercare `inurl=mario inurl:rossi` equivale a `allinurl:mario rossi`.

`intext` e `allintext`

Risultati di pagine che contengono la *query* (parola singola in caso di `intext`, parola multipla in caso `allintext` nel `body` del documento HTML. Utile per trovare pagine che contengono il testo più rilevante o in ogni caso meglio ottimizzato per la parola chiave.

`intitle` e `allintitle`

Restituisce pagine in cui la *query* di ricerca è contenuta nel *tag* `title` del documento HTML. Questo è il testo che compare nella maggior parte dei casi come titolo del risultato di ricerca. In poche parole, grazie a questo operatore è possibile trovare tutti i risultati che contengono l'intera *query* nel titolo stesso del risultato.

`inanchor` e `allinanchor`

Restituisce le pagine in cui la *query* è contenuta in un elemento ipertestuale.

`daterange`

Operatore che filtra risultati corrispondenti a pagine caricate o aggiornate in un determinato periodo temporale. In termini di ricerca legata all'identità e alla reputazione è un operatore molto utile perché consente di sapere la data di un determinato contenuto e di valutarne la sua obsolescenza in un'ottica di diritto all'oblio.

`related`

Restituisce pagine che Google ritiene simili a quella dichiarata tramite ricerca così definita `related:http//:www.sitoweb.it`. Per stabilire quali pagine sono da considerarsi simili, Google esamina la provenienza di eventuali *backlinks* sulla pagina dichiarata, ossia guarda quali siti ospitano *link* che puntano quella pagina e successivamente controlla quali altri siti queste pagine puntano dei *link*.

`info`

Grazie a questo operatore è possibile sapere se una pagina è stata indicizzata o meno da Google: fornisce il titolo e lo *snippet*, il *link* alla pagina, un *link* alla copia *cache* della pagina.

`cache`

Fornisce la versione della pagina al momento in cui è stata indicizzata da Google. Anche questo può essere a scopi di valutazione di obsolescenza

di informazioni: la versione della pagina conservata nella *cache* di Google potrebbe essere diversa dal contenuto originale (perché aggiornato in seguito).

3.4.2 Parametri HTTP del Protocollo di Ricerca di Google Search

Se si inserisce "mario rossi" nella barra di ricerca sulla pagina di `www.google.it`, nel momento io cui si preme invio si potrà notare che nella barra degli indirizzi del browser che si sta usando compare una URI di questo tipo:
`https://www.google.it/search?q=mario+rossi`

Se invece la stringa viene inserita direttamente nella barra degli indirizzi, in questo caso, un *browser* Firefox, ecco che la URI che ne risulta sarà molto più complessa:
`https://www.google.it/search?q=mario+rossi&ie=utf-8&oe=utf-8&client=firefox-b&gfe_rd=cr&ei=GlZiWIilOdTA8gfv9q_YCQ`

Ciò che si può osservare, è che entrambi le URI generate condividono la prima parte, cioè `https://www.google.it/search?q=mario+rossi`, la richiesta inviata da Firefox aggiunge una stringa ulteriore, nella quale si può individuare un `client=firefox` che ci fa pensare che tutto ciò che è contenuto nella stringa si tratti di una serie di informazioni relative alla versione integrata a Firefox di Google Search.

Se adesso si prende la URI "pulita" `https://www.google.it/search?q=mario+rossi` e la si copia e incolla nella barra degli indirizzi di qualsiasi *browser* (che sia Firefox, Chrome, Safari, Opera ecc.) si noterà che la pagina restituita sarà esattamente la stessa, cioè la SERP della ricerca per "mario rossi".

Questo ci fa capire che quando si invia una richiesta a Google, quello che facciamo non è altro che generare e inviare una richiesta HTTP alla base di dati di Google, il quale restituirà come risposta la SERP dei risultati ottenuti dalla ricerca.

Una richiesta di ricerca di Google è una tipica richiesta HTTP di tipo GET, che restituisce risultati in HTML o in alternativa (se richiesto) in XML. La stringa che viene dopo `https://www.google.it` è il *path* che descrive la *query* di ricerca ed è costituito da `/search?`, sempre presente anche se a volte sostituita dal simbolo `#`, e da una serie di paramentri separati da `&` che

sono chiamati parametri di *input*. Tutti questi parametri fanno parte del protocollo di ricerca di Google.

Il ruolo del paramentro `q` è facilmente intuibile: il suo valore è la stringa della *query* di ricerca inserita dall'utente, che aggiunge un `+` laddove la stringa sia composta da più parole separate da spazio. Il parametro `ie` imposta la codifica dei caratteri che è usata per interpretare la *query* di ricerca, che nel caso dell'esempio sopra è `utf-8`, mentre `oe` imposta la codifica usata per codificare i risultati. Il paramentro `client` indica il *front end* da cui viene inviata la richiesta, in questo caso Firefox. Il paramentri `gfe_rd` e `ei` non sono presenti nella documentazione fornita da Google. Tuttavia da una ricerca sul *web* sembra che `gfe` stia per "Google Front End" e `rd` significhi "Redirect", mentre il valore `cr` sia "Country": il paramentro indica quindi che è avvenuto un reinderizzamento alla versione di Google del proprio paese. Il parametro `ei` invece sta per "Engine Id" ed è un *cookie* inviato al *desktop*.

Tutto ciò significa che, a patto di conoscere bene il protocollo di ricerca di Google, è possibile costruirsi delle richieste HTTP personalizzate che ci consentono di affinare la ricerca ad un livello ancora più sofisticato degli Operatori di Ricerca Avanzata.

I parametri di *input* del protocollo di ricerca sono molti, qui di seguito sono elencati i più importanti e i più interessanti[6].

- `q`: stringa della *query* di ricerca.
- `num`: consente di definire il numero di risultati per pagina. Il valore del parametro non può superare il numero 1000.
- `start`: specifica il numero di indice del primo risultato che deve essere restituito nella lista dei risultati. In altre parole, si può stabilire se Google deve restituire i risultati a partire dal primo oppure dal decimo o dal ventesimo, ossia dalle pagine della SERP successive alla prima.
- `rc`: richiede un accurato conteggio dei risultati fino a 1.000.000 documenti
- `site`: limita la ricerca a contenuti appartenenti ad un determinato dominio.

[6]Google, *Google Protocol Reference*, https://www.google.com/support/enterprise/static/gsa/docs/admin/72/gsa_do 19/4/2017.

- `sort`: specifica un criterio di ordinamento dei risultati, come per esempio per data.
- `client`: determina il *front end* da cui far partire la richiesta HTTP, per esempio `firefox-a` se si vuole impostare Mozilla Firefox
- `output`: seleziona il formato dei risultati di ricerca, come `html` o `xml`.
- `partialfields`: restringe la ricerca a documenti con *meta-tag* il cui valore contiene la parole specificate
- `requiredfields`: restringe la ricerca a documenti che contengono esattamente i *tag* `name` o coppie di nome-valore specificati
- `pws`: abilita/disabilita la personalizzazione
- `filter`: quando settato su 0, include risultati omessi
- `complete`: attiva (=1) o disattiva (=0) le funzioni di Autocomplete e Google Instant
- `nfpr`: attiva (=1) o disattiva (=0) l'autocorrezione
- `ncr`: imposta il *redirect* alla versione di Google del Paese di un determinato Paese. Consente così di settare il Paese da cui vogliamo ottenere la versione di Google indipendentemente dalla posizione geografica da dove viene inviata la richiesta.
- `safe`: abilita/disabilita filtro per contenuti per adulti
- `tbm`: serve a selezionare ricerche speciali, come per immagine, video, notizie ecc.:
  - Applicazioni: `tbm=app`
  - Libri: `tbm=bks`
  - Immagini: `tbm=isch`
  - Notizie: `tbm=nws`
  - Brevetti: `tbm=pts`
  - Video: `tbm=vid`
- `tbs`: consente di limitare la ricerca a un determinato periodo di tempo:

  - qualsiasi data: `tbs=qdr:a`
  - ultimo secondo: `tbs=qdr:s`
  - ultimo minuto: `tbs=qdr:n`
  - ultimi 10 minuti: `tbs=qdr:n10` (and così via per tutti numeri di minuti)
  - ultima ora: `tbs=qdr:h`
  - ultime 12 ore: `tbs=qdr:h10` (e così via per tutti i numeri di ore)
  - ultimo giorno: `tbs=qdr:d`
  - ultima settimana: `tbs=qdr:w`
  - ultimo mese: `tbs=qdr:m`
  - ultimo anno: `tbs=qdr:y`
  - un intervallo di tempo specifico, per esempio dal 2 marzo 1984 al 5 giugno 1987: `tbs=cdr:1,cd_min:3/2/1984,cd_max:6/5/1987`
  - ordinamento per data: `tbs=sbd:1`
  - ordinamento per rilevanza: `tbs=sbd:0`

- `lr`: restringe la ricerca a pagine in una lingua specificata.
- `hl`: impostazioni di lingua forniti dal browser
- `cr`: la regione da cui il risultato dovrebbe provenire
- `gr`: limita i risultati a una certa regione
- `gcs`: limita i risultati a una certa città, oppure latitudine e longitudine
- `ie`: *input encoding*, cioè imposta la codifica dei caratteri usata per interpretare la stringa di ricerca
- `oe`: *output encoding*, cioè imposta la codifica dei caratteri usata per restituire i risultati
- `sitesearch`: svolge esattamente la stessa funzione dell'operatore `site:`
- `access`: specifica se la ricerca deve includere solo contenuti pubblici (=p), protetti (=s) o entrambi (=a).

Capitolo 4

Scrappy: progettazione di un *SERP Scraper* per il monitoraggio dell'identità digitale

4.1 Monitoraggio automatico della SERP

L'analisi manuale della SERP ha un ruolo fondamentale e insostituibile nel monitoraggio dell'identità digitale. Oltre ad essere la soluzione più immediata e intuitiva per fare un *check up* reputazionale, la sua grande importanza risiede nel fatto che quanto va cercato, per condurre un'analisi sull'identità digitale di un soggetto, è tutto ciò che gli altri possono trovare su di lui. Agire, pensare e comportarsi come se si fosse un utente in cerca di informazioni su quella persona e, soprattutto, usare i suoi stessi strumenti è senza dubbio la migliore strategia per prendere coscienza di come gli altri possono conoscerla navigando sul Web. Tuttavia, il processo è lento e non può essere fatto con una periodicità giornaliera.

Per eseguire un monitoraggio continuativo nel tempo, quanto più possibile omogeneo nel tipo di dati che vengono raccolti e archiviati, è necessario uno strumento in grado di sostituirsi al lavoro umano e ispezionare automaticamente il contenuto della SERP per restituire dati strutturati e normalizzati.

Il grande vantaggio di poter ottenere dati normalizzati non sta soltanto

nel rendere più veloce, e quindi più comoda, la raccolta di informazioni, ma consente di poter utilizzare quest'ultime all'interno di programmi per l'analisi di dati come per esempio Microsoft Excel o Google Sheets.

Ciò che un programma del genere deve eseguire è, data una *search query* - la stessa che verrebbe inserita nella *query box* di Google - simulare la ricerca e la visita della SERP restituita per raccogliere tutti i dati necessari all'analisi dell'identità digitale sul motore di ricerca.

4.2 *Web Scraping*

L'idea è stata quella di creare un programma molto usato in ambito di *data mining* e nella SEO: un *web scraper*. Il *web scraping* (lettermalmente "raschiatura del Web") è una pratica di raccolta automatica di dati da pagine *web* attraverso un programma *software* che simula la navigazione umana. Il principale scopo dello *scraping* è l'estrapolazione delle informazioni dal *corpus* di un testo disponibile sulla rete Internet. I dati sono estratti, elaborati e archiviati in una base di dati.

Uno *scraper* può accedere al *World Wide Web* direttamente usando il protocollo HTTP oppure attraverso un *browser*, visita automaticamente documenti *web* e ne copia i contenuti o parte di essi in un *database* o in uno *spreadsheet*. Per fare *scraping* di una pagina *web* è necessario che questa venga ottenuta, parsificata e infine copiata per intero o nelle parti necessarie su un altro *file*.

Lo *scraping* è di base un'attività legale, perché non fa niente di diverso da quello che potrebbe fare manualmente un utente umano che naviga sul Web, ma di fatto si tratta di visitare automaticamente e copiare documenti *web* in una quantità e ad una velocità che un essere umano non potrebbe mai sostenere. Questo può non "piacere" ai siti *web* per una serie di ragioni. Per esempio, questa attività può danneggiare il sito stesso: la visita massiva di un sito da parte di uno *software* consuma le risorse del *server* che lo ospita, rendendo più lenta o impedendo del tutto la visualizzazione dei documenti a tutti gli altri utenti. Inoltre l'attività di *scraping* è spesso usata per tecniche illecite di SEO, come duplicazione di contenuti o *link spam* (un ottimo programma commerciale di *scraping* ScrapeBox si è per esempio fatta una pessima reputazione in quanto è stato impiegato in modo massiccio nella

generazione automatica di commenti *spam* nei *blog* al fine di ottenere grandi quantità di *link* che puntano a siti commerciali).

4.3 Quali dati raccogliere

L'idea alla base della progettazione di uno *scraper* da utilizzare per l'analisi della reputazione *online*, è stata quella di creare uno strumento in grado non tanto di analizzare la reputazione, ma di raccogliere tutti i dati ritenuti utili a una vasta serie di analisi e forme di monitoraggio della stessa.

La prima fase della progettazione del programma è stata quella di stabilire quali dati dovessero essere raccolti e in quale formato e struttura dovessero essere salvati.

Il criterio di identificazione dei dati necessari è stata la loro classificazione in base ai sei aspetti della reputazione *online* discussi nel capitolo precedente.

4.3.0.1 Visibilità

I dati relativi all'aspetto della visibilità sono stati di facile individuazione:

- **rank** : la posizione del risultato all'interno della colonna principale della SERP
- **title**: il titolo del risultato
- **snippet**: la breve descrizione sottostante il titolo

Benché titolo e *snippet* facciano parte anche dell'aspetto del contesto, in quanto danno informazioni rilevanti sul contenuto del risultato, sono le prime informazioni che l'utente legge e stabiliscono se l'utente deciderà di approfondire l'informazione seguendo il *link* offerto dal risultato oppure ignorarlo.

Il *rank*, come già largamente discusso, è il fattore fondamentale della visibilità di un contenuto e sull'influenza che esercita sulla prima impressione che un'identità digitale può fare sull'utente.

4.3.0.2 Accessibilità

Oltre ai fattori di accessibilità di un contenuto stabiliti dal suo grado di visibilità (*rank*) ciò che può essere facilmente ottenuto è lo *status code* delle

risposte HTTP una volta inviata la richiesta ai *link* dei risultati di ricerca. Se la risposta è 200 significa che la richiesta è stata accettata e la pagina è accessibile, se restituisce un numero nell'ordine dei 400 o 500 significa che per qualche ragione, lato *client* o lato *server* che sia, la risorsa non è disponibile e dunque non è accessibile. Se la risposta è un 404, significa che è obsoleta e che quindi può essere facilmente rimossa inviando una segnalazione a Google.

Contesto

Il contesto è l'aspetto più difficile da formalizzare e ridurre in dati quantificabili. L'idea, è stata quella di verificare e contare semplicemente le occorrenze di una o più parole all'interno del corpo di testo delle pagine *web* a cui i risultati si riferiscono.

Nella pratica, permette di verificare quante volte è possibile imbattersi, per esempio, nella parola "ergastolo" (parola effettivamente reperibile nel primo risultato per "mario rossi" in data 18/5/2017[1]), eseguendo una ricerca per la *query* "mario rossi".

Il risultato è ben diverso dal cercare "mario rossi ergastolo", in quanto questa *query* restituisce tutti i documenti in cui compare contemporaneamente sia la stringa "mario rossi" che la stringa "ergastolo". Nel nostro caso invece, è possibile verificare in quanti risultati restituiti per la ricerca "mario rossi" è contenuta la parola "ergastolo" e con quale frequenza (nonché, grazie agli altri dati, con quale visibilità e a quale livello di accessibilità).

Serendipità

Dati relativi alla serendipità sono di facile di individuazione sulla SERP: le ricerche correlate in fondo alla pagina. Ottenerle e, magari, ripetere il processo di estrazione di informazioni della SERP che generano a loro volta può ampliare notevolmente lo spettro di indagine sulle possibili informazioni reperibili su un soggetto e la percezione della probabilità con cui queste possono essere trovate involontariamente da un utente.

[1] Wikipedia, voce *Mario Rossi (terrorista)*, https://it.wikipedia.org/wiki/Mario_Rossi_(terrorista), 18/4/2017.

Tempo

Alcuni risultati (specialmente notizie da quotidiani, *blog* e *forum*) riportano la data di indicizzazione della pagina (o a volte la data stessa di pubblicazione, nel caso il sito *web* fornisca a Google questa informazione, come per esempio quelli su piattaforma WordPress).

Tuttavia non tutti i risultati mostrano la data sulla SERP, per cui non sempre è semplice sapere quando un documento è stato pubblicato o indicizzato.

Per quanto si sia cercato di ricorrere ad alcuni "trucchi" per ottenere la data di indicizzazione di tutti i risultati presenti in SERP, il successo è stato solo parziale e molto poco controllabile. È stato scelto alla fine di escludere la funzione di estrazione della data dal programma.

Responsabilità

Per ogni pagina *web* individuare il responsabile del sito e i suoi contatti. Tramite il protocollo di rete Whois è possibile risalire a quale *internet provider* appartiene un determinato dominio e ottenere tutte le informazioni del suo intestatario.

4.4 Scelta del linguaggio di programmazione e moduli impiegati

4.4.1 Perché Python

I principali linguaggi di programmazione utilizzati per creare *scraper* sono Python e PHP[2]. La scelta è ricaduta immediatamente su Python per due ragioni fondamentali: l'ampia documentazione disponibile sull'impiego di questo linguaggio in applicazioni di *data mining* e analisi di dati (*scraping* e *crawling* incluso) e l'esistenza di alcune librerie fondamentali ai fini del progetto, in particolare Requests e Beautiful Soup.

Inoltre, la semplicità della sintassi di Python ha reso il suo apprendimento veloce e di immediato impiego per i propositi stabiliti.

La versione utilizzata di Python è la 3.5, preferita alla versione 2.7, di *default* sul calcolatore su cui è stato scritto ed eseguito il programma (Mac-

[2]Perlomeno, nella documentazione reperibile in rete e nella manualistica dedicata.

book 2009 Osx 10.9.5), per poter far girare alcune librerie non disponibili per Python 2.7, tra le quali proprio Requests e Beautiful Soup.

Complessivamente, all'interno del programma si è fatto uso di sei moduli Python:

- Requests
- Beautiful Soup
- Whois
- Re
- Sleep
- Csv

Nelle prossime sezioni saranno esposti nel dettaglio.

4.4.2 Requests

La prima necessità per iniziare a fare *scraping* era poter inviare richieste HTTP a Google Search per ottenere la SERP da analizzare.

La scelta era inizialmente ricaduta sulla libreria `urllib2`[3], già facente parte della libreria standard di Python, un modulo che definisce funzioni e classi per aprire URL. Tuttavia, per quanto il modulo risultasse funzionare egregiamente per semplici indirizzi *web* (come ad esempio `https://www.google.it`), risultava inefficiente per richieste HTTP più complesse, come proprio le richieste Google (es. `https://www.google.it/search?q=mario+rossi`).

Il modulo `urllib2` infatti mette a disposizione quasi tutte le principali funzionalità HTTP, ma richiede molto lavoro (addirittura anche l'*overriding* di metodi) per effettuare anche le operazioni più semplici.

Come consigliato sulla stessa documentazione di `urllib2`, il modulo Requests[4] è risultato di utilizzo molto più semplice per gli scopi preposti.

Requests è una libreria HTTP con licenza Apache2, creata da Kenneth Reitz, che permette di inviare richieste HTTP/1.1 utilizzando Python con la possibilità di aggiungere facilmente contenuti come *header*, *data form* e parametri.

[3]Python Software Foundation, https://docs.python.org/2/library/urllib2.html, 19/4/2017.

[4]Python Software Foundation, http://docs.python-requests.org/en/master/, 19/4/2017.

4.4.3 Beautiful Soup

Beautiful Soup[5] è il modulo fondamentale che ha permesso la visita dei documenti HTML e l'estrazione dei dati. Si tratta di un *parser* di documenti HTML e XML molto potente e frequentemente utilizzato nello *scraping*.

Il suo nome trae ispirazione da una poesia recitata da un personaggio di *Alice nel Paese delle Meraviglie* Mock Turtle, il quale si presenta come un incrocio tra una mucca e una tartaruga. La filastrocca, così come il personaggio stesso che la recita, ironizza sulla zuppa Mock Turtle inglese, a base di carne bovina e non di tartaruga.

Come il gioco di parole nel romanzo di Caroll, il modulo Beautiful Soup mira a dare senso a ciò che senso non ha: aiuta a formalizzare e organizzare pagine *web* disordinate e a riparare HTML corrotto. Il *tag soup* infatti è la definizione, nel gergo informatico, del codice HTML sintatticamente o strutturalmente mal formato. Un *parser* HTML che si rispetti, elemento fondamentale di un *web browser* così come di un *crawler*, deve essere in grado di interpretare *markup* HTML anche se contiene errori di sintassi o di struttura. Deve essere dunque un *tag soup parser*.

La popolarità di Beautiful Soup sta nella sua capacità di parsificare pagine *web* e fornire una semplice ma molto potente interfaccia per navigare, interrogare e modificare l'albero ottenuto dal documento HTML (o anche altri formati). È uno strumento utilissimo per dissezionare un documento ed estrarre ciò che si vuole.

Un altro grande vantaggio di Beautiful Soup è che converte automaticamente documenti in ingresso in Unicode e restituisce in *output* documenti in UTF-8. In questo modo il programmatore non deve preoccuparsi del sistema di codifica.

4.4.4 Whois

Libreria creata da Richard Penman, Whois[6] fornisce informazioni di registrazione di un dominio *web*.

Semplicemente fornendo un indirizzo *web*, il modulo è in grado di restituire una vasta gamma di informazioni sul dominio di quell'indirizzo, come il

[5]Crummy, https://www.crummy.com/software/BeautifulSoup/, 19/4/2017.
[6]Bitbicket, https://bitbucket.org/richardpenman/pywhois, 19/4/2017.

nome e l'indirizzo del proprietario del dominio, l'indirizzo *e-mail*, il servizio *server* in cui risiede il sito *web* e molto altro.

4.4.5 Re

Celebre modulo per utilizzare le espressioni regolari in Python, Re[7] è stato impiegato per individuare ed estrarre alcuni dati nel testo HTML.

4.4.6 Sleep

Sottomodulo del modulo Time della libreria standard di Python, è stato impiegato per creare delle pause tra l'invio di una richiesta HTTP e l'altra onde evitare che Google si accorga di essere interrogato da un *software* e non da un essere umano e blocchi l'indirizzo IP impedendo quindi al programma di procedere con lo *scraping*.

4.4.7 Csv

Altro modulo della libreria standard, consente di scrivere e leggere *file* di testo in formato CSV (*Comma Separed Values*), il formato più comune di importazione ed esportazione per *spreadsheet* e basi di dati.

4.5 Ispezione del codice HTML della SERP di Google

Stabiliti quali dati raccogliere e con quali mezzi, la fase successiva è stata quella di capire dove andarli a cercare.

È stata dunque condotta un'ispezione della SERP ottenuta con la ricerca della stringa "mario rossi" per individuare ogni singolo risultato della colonna centrale e in quali *tag* HTML fossero contenuti i dati necessari.

Gli elementi che andavano trovati sulla SERP erano, per ogni risultato di ricerca:

- titolo
- snippet
- URL

[7] Python Software Foundation, https://docs.python.org/2/library/re.html, 19/4/2017.

Una volta ottenuta la URL sarebbe poi stato possibile ottenere gli altri dati (*status code*, proprietario, occorrenze *keyword*) dalla pagina *web* linkata.

Per eseguire l'ispezione è stato impiegato Firefox Developer Edition, la versione per sviluppatori del *browser* Firefox che consente di visualizzare il codice sorgente di pagine *web* e di localizzare visivamente ogni elemento HTML sulla pagina renderizzata attraverso i selettori CSS.

Per verificare che il *markup* visualizzato in Firefox Developer Edition corrispondesse effettivamente a quello restituito attraverso il modulo Requests di Python, è stato creato ed eseguito il seguente programma:

```
import requests
from bs4 import BeautifulSoup
def pagina(url):
    page = requests.get(url)
    soup = BeautifulSoup(page.text, 'html.parser')
    file = open('pagina.html','w')
    file.write(soup.prettify())
    file.close()
url = input('Inserire URL: ')
pagina(url)
```

Il programma ottiene il documento, lo parsifica e scrive il codice (già indentato grazie al metodo `.prettify()` di Beautiful Soup) su un *file* HTML che viene restituito in *output*.

Aprendo il *file* con un editore di testo si avrà l'intero *markup* della SERP, leggibile e modificabile *offline* (se invece si apre con un *browser*, si può notare che pagina renderizzata è uguale circa per il 90% all'originale, anche per lo stile e comportamenti dinamici, dal momento che CSS e Javascript sono embeddati nell'HTML).

Confrontando il codice di Firefox Developer Edition e quello restituito da Requests è saltato subito all'occhio come il secondo sia molto più pulito, libero da troppi `div` annidati e (in apparenza) superflui presenti invece nel primo. Talvolta gli attributi `id` e `class` dello stesso elemento hanno nomi diversi nelle due versioni del *markup*.

Questa non perfetta corrispondenza tra il codice sorgente ottenuto dal *browser* e quello ottenuto con Requests non ha comunque impedito di trovare

comunque i dati necessari, anzi ha semplificato il lavoro grazie al codice più pulito.

Tutti i risultati organici della colonna centrale (la quale è un `div` identificato come `id="center_col"`) sono situati in un `div` identificato come `id="search"`. Da questo `div` sono esclusi eventuali risultati sponsorizzati e le ricerche correlate in fondo alla pagina, così come tutti i risultati della colonna di destra, i quali sono contenuti nel `div id="rhscol"`.

Ogni risultato organico è contenuto in un `div` con attributo `class="g"`, sia che esso sia un risultato standard che un risultato verticale. Sono esclusi dai risultati `class="g"` eventuali risultati in posizione cosiddetta "0", risultati scelti da Google da uno dei primi dieci (non necessariamente il primo) e messo in evidenza in una finestra che sta prima di tutti gli altri.

Ad un risultato standard come quello in Figura 4.1 (che è il primo risultato di ricerca della *query* "mario rossi"),

Figura 4.1: Risultato di ricerca

corrisponde il *markup* visibile in Figura 4.2.

All'interno del `div class="g"` ci sono due elementi principali:

- `<h3 class="r">`: contiene il titolo del risultato di ricerca e al suo interno si trova l'URL della pagina web a cui si riferisce.
- `<div class="s">`: contiene l'URL mostrato sotto il titolo del risultato di ricerca e il testo dello *snippet*

Per quanto riguarda risultati verticali, la cosa si fa più difficile, in quanto ognuno si comporta a suo modo e la loro struttura è molto più complessa. Dalle tipologie di risultati verticali che si possono trovare nella colonna centrale (per lo più risultati di Google Image e Google News) si può osservare che anche questi presentano comunque un tag `<h3>` che racchiude il titolo del risultato (presentando la formula fissa "Immagini relative a...", "Notizie relative a..." oppure "Prima pagina", ecc.)

Nel caso di Google Image, il *tag* `<h3>` è riporta un attributo `class="_DM"` e al suo interno si trova l'URL della ricerca verticale per immagini della

```
<div class="g">
    <h3 class="r">
        <a href="/url?q=https://it.wikipedia.org/wiki/Mario_Rossi_(
        terrorista)&amp;sa=U&amp;
        ved=0ahUKEwiBnKDloMXRAhWJ1xoKHfE-CPsQFggZMAE&amp;
        usg=AFQjCNFLApRsiuQqEyEzzEm8ot0ByzBLvw"><b>Mario Rossi</b> (
        terrorista) - Wikipedia</a>
    </h3>
    <div class="s">
        <div class="kv" style="margin-bottom:2px">
            <cite>https://it.wikipedia.org/wiki/<b>Mario</b>_<b>Rossi</b>_(
            terrorista)</cite>
            <div class="_nBb"><div aria-expanded="false" aria-haspopup="
            true" data-ved="0ahUKEwiBnKDloMXRAhWJ1xoKHfE-CPsQ7B0IGjAB"
            onclick="google.sham(this);" style="display:inline" tabindex="0
            ">
                <span class="_O0"></span>
            </div>
            <div class="am-dropdown-menu" role="menu" style="display:none"
            tabindex="-1">
                <ul>
                    <li class="_Ykb"><a class="_Zkb" href="/url?q=http://
                    webcache.googleusercontent.com/
                    search%3Fq%3Dcache:oYBo0DnARDoJ:https://
                    it.wikipedia.org/wiki/Mario_Rossi_(
                    terrorista)%252Bmario%2Brossi%26hl%3Dit%26ct%3Dclnk&amp
                    ;sa=U&amp;ved=0ahUKEwiBnKDloMXRAhWJ1xoKHfE-CPsQIAgcMAE&
                    amp;usg=AFQjCNFq-176XbhSS26ipCYkcKqO-Xb_bg">Copia cache
                    </a></li>
                    <li class="_Ykb"><a class="_Zkb" href="/
                    search?q=related:https://it.wikipedia.org/wiki/
                    Mario_Rossi_(terrorista)+mario+rossi&amp;tbo=1&amp;sa=X
                    &amp;ved=0ahUKEwiBnKDloMXRAhWJ1xoKHfE-CPsQHwgdMAE">
                    Simili</a></li>
                </ul>
            </div>
        </div>
    </div>
    <span class="st"><b>Mario Rossi</b> (Genova, 19 agosto 1942) è un
    terrorista italiano. Fu a capo del <br>
        gruppo terroristico Gruppo XXII Ottobre. Indice. [nascondi]. 1
        Biografia; 2 Periodo<br>
         ...</br></br></span>
        <br/>
    </div>
</div>
```

Figura 4.2: *Markup* del risultato di ricerca

query. Il risultato Google News invece presente un tag `<h3>` con attributo `class="_MRj"` che non contiene *link* al suo interno, tutti i link ad ogni singola notizia mostrata sono invece raccolti in una lista (`<ul class="_vio">`).

I risultati di ricerca hanno quindi strutture piuttosto eterogenee a seconda della tipologia di contenuto, l'unica regola che sembra si possa assumere è che per ogni risultato corrisponde un *tag* `<h3>`, indipendentemente dagli attributi con cui viene identificata. Osservando tutto il codice HTML della SERP, si è potuto notare che questo tag viene impiegato soltanto per i titoli dei singoli risultati di ricerca della colonna di sinistra (i `div class="g"` invece sono impiegati anche nella colonna di destra).

4.6 Definizione del *task* del programma ed elementi di *input* e *output*

Prima di iniziare la scrittura vera e propria del programma, è stato necessario chiarire quale fosse il compito del programma e quali informazioni fornirgli affinché lo svolgesse.

Il programma deve simulare la ricerca su Google Search da parte di un utente per una o più *query* di ricerca. Da ogni SERP ottenuta deve raccogliere *rank*, titolo, *snippet* e URL di ogni risultato della colonna centrale. Successivamente, deve ottenere la risorsa *web* a cui il risultato si riferisce e ottenere lo *status code* della richiesta HTTP inviata, i dati relativi al proprietario del dominio e contare le occorrenze di una o più parole definite dall'utente all'interno del testo della pagina. Infine, per ogni ricerca correlata presente in fondo alla SERP deve ripetere il *task* dall'inizio.

Gli elementi di *input* del programma devono essere:

1. una lista di stringhe che corrispondono alla *query* di ricerca,
2. una lista di stringhe che corrispondono alle parole di cui verificare le occorrenze,
3. il numero di risultati che devono essere restituiti dal motore di ricerca,
4. il numero di volte per cui il *task* deve ripetersi per le ricerche correlate,

In *output* il programma deve restituire un *file* csv contenente i seguenti dati (con intestazione):

1. **Query**: stringa di ricerca dell'input 1 o di una ricerca correlata
2. **Rank**: posizione nella SERP del risultato
3. **Title**: titolo del risultato
4. **Snippet**: descrizione risultato
5. **URL**: indirizzo della pagina relativa al risultato
6. **Status Code**: lo *status code* della richiesta HTTP
7. **Resp**: dati relativi al proprietario del dominio, quali nominativo, indirizzo postale, indirizzi *e-mail*
8. **Keyword1**: numero occorrenze della prima parola dell'*input* 2
9. **Keyword2**: numero occorrenze della seconda parola dell'*input* 2
10. **Keyword*n***: numero occorrenze della n-esima parola dell'*input* 2

Il *task* prevedeva inizialmente anche l'estrazione della data di indicizzazione dei risultati, tuttavia, vista la scarsa "eleganza" della soluzione trovata e le sue scarse probabilità di successo, è stato deciso di non includere questa funzionalità nel programma finale.

4.7 Spiegazione della funzione principale del programma

Tutta l'attività di *scraping* del programma è concentrata in un'unica funzione definita `scrap` che restituisce una lista multipla (matrice) le cui righe corrispondono ai dati raccolti per ogni risultato di ricerca. La funzione `scrap` esegue il *task* per una sola *query*, quindi in caso ne siano state inserite più di una la funzione viene richiamata per ognuna di esse. Il resto delle funzioni all'interno del programma servono ad accorpare tutte le matrici corrispondenti ad ogni *query*, creare il *file* csv e a gestire l'interfaccia testuale. La funzione `main()` si occupa della gestione generale del programma.

La funzione ha i seguenti parametri:

- `query`: la stringa che rappresenta la *query* di ricerca e che viene utilizzata per costruire la richiesta HTTP da inviare a Google.

- `keywords`: è una lista di stringhe che rappresentano le parole di cui conteggiare le occorrenze.
- `results`: numero dei risultati che la ricerca di Google deve restituire.
- `corr`: il numero di volte che il processo deve essere effettuato per le ricerche correlate, se è 0 le ricerche correlate non vengono prese in considerazione, se è 1 la funzione viene richiamata (tramite ricorsione) assegnando al parametro `query` ciascuna ricerca correlata trovata in fondo alla SERP.
- `matrix = []`: si tratta di una lista inizialmente vuota, che viene riempita via via di altre liste che corrispondono ognuna ai dati raccolti per ogni risultato di ricerca. Per ogni volta che la funzione viene richiamata all'interno della funzione stessa per le ricerche correlate, questo parametro assume il valore della matrice contenente i dati fino a quel momento raccolti.

4.7.1 Costruzione della *query string* della richiesta HTTP

La prima cosa che esegue la funzione, è la costruzione della *query string* che contiene le informazioni di ricerca della richiesta che verrà inviata a Google.

Di base in una richiesta come

```
https://www.google.com/search?q=mario+rossi
```

la *query string* è tutto ciò che viene dopo il simbolo `?`.

Come è già stato visto, il protocollo di ricerca di Google è costituito da un ricco assortimento di parametri, molti dei quali sostituiscono gli operatori di ricerca avanzata e altre operazioni di filtraggio dei risultati.

Ciò che serve per avere una SERP il più possibile non influenzata dai vari fattori con cui Google modifica una SERP "standard" è il semplice parametro `q`.

La `http://www.google.it/search?q=mario+rossi` è già sufficiente per restituire una lista di risultati abbastanza neutra (non influenzata cioè da eventuali personalizzazioni del *client* come localizzazione geografica, cronologia, *login* a Google+).

Altra cosa necessaria è stabilire la quantità di risultati da restituire. Il parametro che fa al caso nostro è `num`, che fissa un massimo di *link* da restituire, così che, per esempio, `num=10` includa 10 risultati.

La URL necessaria per avere i primi 10 risultati per "mario rossi" ha dunque questa composizione: `http://www.google.it/search?q=mario+rossi&num=10`.

Una volta compreso come costruire la URL, entra in gioco il modulo Requests di Python, che consente di costruire la URL definendo tutti i parametri necessari all'interno di un dizionario da assegnare all'argomento `params` del metodo `.get`, insieme alla URL di base, in questo modo:

```
search_engine = 'http://www.google.it/search'
payload = { 'q' : query, 'num' : results }
my_headers = { 'User-agent' : 'Mozilla/53.0' }
serp = requests.get( search_engine, params = payload, headers = my_headers )
```

La variabile `search_engine` è la stringa di base della URL per inviare richieste a Google Search, a cui deve essere aggiunta la *query string*. Il dizionario `payload` contiene i parametri di ricerca della *query string* a cui viene assegnato il valore dei parametri passati in *input* alla funzione, già definiti in precedenza.

`my_headers` imposta l'`user-agent`, in questo caso l'ultima versione di Mozilla Firefox. Se questo parametro non viene impostato Requests indica un `user-agent` proprietario di *default* che potrebbe essere rifiutato dal *server* percependolo come una richiesta effettivamente invitata da un *bot*.

La variabile `serp` è la risposta della richiesta HTTP di tipo GET inviata da Requests con la URL costruita con i vari parametri dichiarati. È, a tutti gli effetti, la SERP per la ricerca "mario rossi"!

4.7.2 Parsificazione del documento HTML e creazione dell'oggetto `BeautifulSoup`

Una volta ottenuta la risorsa HTML, è il momento di renderla navigabile. Tutto viene eseguito dal metodo `BeautifulSoup` che provvede alla parsificazione del *markup* HTML e alla creazione di un oggetto `BeautifulSoup` che rappresenta l'albero ottenuto pronto per essere navigato e interrogato attraverso i metodi offerti dal modulo.

```
serp_soup = BeautifulSoup( serp.text, 'html.parser' )
```

I parametri passati nella funzione sono il documento ottenuto con Requests, restituito con il metodo `.text` e il tipo di *parser* che si vuole utilizzare.

4.7.3 Individuazione dei risultati di ricerca

Adesso, si tratta di navigare l'albero creato dall'oggetto `BeautifulSoup` e trovare tutti i risultati di ricerca presenti nel testo. Come verificato durante l'ispezione del *markup* della SERP, ogni risultato della colonna centrale è all'interno di un `div` con attributo `class="g"`, è dunque necessario utilizzare un metodo in grado di raccogliere tutti i `div class="g"` presenti nell'albero.

Il metodo in questione è `.find_all()`, che recupera tutti i *tag* che corrispondono ai tipi di filtro passati come parametri, come per esempio nome del *tag* e attributo, e li restituisce in una lista.

L'istruzione `serp_soup.find_all('div', class_='g')` trova tutti i `div class="g"` presenti nel documento. Tuttavia, i `div class="g"` nella SERP si trovano anche nella colonna di destra: se si recuperano tutti indiscriminatamente si otterrebbe una lista di risultati della colonna principale e quella di destra mischiati fra di loro senza quindi avere la giusta percezione della loro disposizione e ordine reale della SERP. I `div class="g"` della colonna centrale sono all'interno di un `div` con attributo `id="search"`, sono quelli che vanno raccolti. Prima di individuare i `div class="g"` è stato quindi recuperato il `div` con attributo `id="search"` attraverso il metodo `.find()`, in tutto e per tutto simile a`.find_all()` ma con la differenza che si limita a recuperare solo la prima occorrenza dell'elemento con i requisiti richiesti. Dopodiché vengono raccolti tutti i `div class="g"` e inseriti in una lista chiamata `div_g`.

```
div_search = serp_soup.find( id='search' )
div_g = div_search.find_all('div', class_='g')
```

4.7.4 Ottenimento *rank*, titolo, URL e *snippet*

Una volta collezionati tutti i risultati, è il momento di cominciare a raccogliere i dati, a partire quelli recuperabili direttamente sul *markup* della

SERP. In seguito si tratterà di ottenere la risorsa Web a cui si riferiscono per raccogliere il resto.

Il recupero dei dati deve essere effettuato, ovviamente, per ognuno dei risultati presenti nella lista `div_g` precedentemente creata. Tutte le istruzioni sono perciò inserite all'interno di un ciclo `for` che le esegue per ogni elemento presente in `div_g`.

Prima di iniziare il ciclo `for` viene dichiarata una variabile `rank` è stato assegnato valore 0. Questa variabile è utilizzata per contare, aumentando di 1 ad ogni ciclo, i risultati nell'ordine in cui sono stati inseriti della lista `div_g` e, di conseguenza, nell'ordine in cui sono stati trovati nella SERP. Il valore che assume questa variabile per ogni risultato di ricerca corrisponde dunque al *rank* dello stesso.

```
rank = 0
for g in div_g:
    rank += 1
    titolo = g.h3
    try:
        title = titolo.get_text()
    except:
        title = 'non trovato'
    try:
        url = re.search('url\?q=(.+?)\&sa', titolo.a['href']).group(1)
    except:
        url = 'non trovato'
    try:
        snippet = g.find('span', class_='st').get_text()
    except:
        snippet = 'non trovato'
    row = [query, rank, title, url, snippet]
```

Il titolo del risultato è all'interno di un *tag* `h3`, evitando di definirne l'attributo `class` si può estrarre il titolo anche di risultati diversi da quelli standard, come immagini e *news*. Beautiful Soup consente di ottenere un elemento HTML con una serie di metodi che portano il nome del *tag* che si vuole raggiungere. Ogni risultato è rappresentato dalla variabile `g` dichiarata

nell'istruzione `for`, applicandovi il metodo `.h3` si ottiene l'elemento `h3` da cui è possibile, tramite il metodo `.get_text()` estrarne il contenuto sotto forma di stringa e assegnarla alla variabile `title`(nel caso il tag contenga altri tag annidati, il metodo `.get_text()` restituisce solo il testo eliminando qualsiasi traccia del *markup*. Un'eccezione assegna alla variabile `title` la stringa `non trovato` nel caso Beautiful Soup non riesca a trovare il *tag*.

Il *tag* `h3` contiene anche il URL della risorsa *web*, all'interno di un tag `a` che è quello che rende il titolo del risultato "cliccabile" per collegarsi nell'immediato alla pagina web relativa. L'URL si trova anche, in forma più pulita, all'interno del `div class="s"` insieme allo *snippet*. Tuttavia, in caso di indirizzi particolarmente lunghi, questi appaiono in forma abbreviata (con tre puntini che sostituiscono la parte mancante) rendendosi quindi inutili all'ottenimento dell'URL effettivo. È stato scelto quindi da recupere l'URL direttamente dal *tag* `h3`, ripulendolo da una *query string* molto più complessa che lo circonda sfruttando le espressioni regolari utilizzabili col modulo Re. L'operazione è stata un po' più difficile, soprattutto per l'utilizzo delle espressioni regolari, ma in questo modo è stato possibile avere la garanzia di ottenere un indirizzo esistente.

Infine viene trovato e ottenuto lo *snippet* che è racchiuso all'interno di uno `span` con attributo `class_="st"` sfruttando il metodo `.find()` e successivamente il metodo `.get_text()`.

Una volta recuperati questi primi tre dati, viene creata una lista nella quale vengono inseriti. Questa lista (nominata `row`) andrà a far parte, una volta completata con altri dati ancora da estrarre, della matrice che verrà restituita dalla funzione `scrap`.

4.7.5 Ottenimento *status code*

Esauriti i dati da raccogliere direttamente sul *markup* della SERP, è giunto il momento di ottenere un nuovo documento tramite una richiesta HTTP, ossia la pagina relativa al risultato di ricerca. Il primo dato che è possibile collezionare con il solo invio della richiesta HTTP per l'indirizzo URL, già estrapolato nella prima parte della funzione, è proprio lo *status code* della risposta inviata dal *server*, che ci rivela l'effettiva disponibilità e accessibilità della pagina *web*. Entra di nuovo in campo il modulo Requests il quale, alla stregua dell'invio della richiesta inviata a Google Search all'inizio della funzione, utilizza il metodo `.get` per inviare la richiesta e ottenere la risposta

e in seguito, tramite il metodo `.status_code`, restituisce lo *status code* della risposta ottenuta.

```
if url != 'non trovato':
    try:
        page = requests.get(url)
        status_code = page.status_code
    except:
         status_code = 'Errore'
    row.append(status_code)
```

Da notare che le istruzioni, insieme a quelle che seguiranno per l'ottenimento dei dati del proprietario del dominio e l'occorrenza di *keyword*, sono all'interno di un condizionale `if` che verifica se la URL è stata precedentemente ottenuta con successo. In caso contrario (alla variabile `url` è stato assegnato il valore di `'non trovato'` dall'eccezione), tutta la raccolta dati relativa alla risorsa riferita è bypassata. Anche in questo caso, un eccezione assegna il valore di `'non trovato'` alla variabile `status_code` nel caso l'invio della richiesta HTTP non sia andata a buon fine.

La stringa ottenuta, identificata con la variabile `status_code` viene poi aggiunta alla lista `row`.

4.7.6 Ottenimento dati proprietario del dominio

Per ottenere i dati del proprietario del dominio della pagina *web*, è stato usato il modulo Whois che permette di avere tutti i dati necessari con molta semplicità.

I dati che vengono ottenuti sono:

- nome del proprietario,
- indirizzo postale completo (via, città, codice postale, nazione)
- indirizzi di posta elettronica

Ciascuno di questi dati vengono ottenuti uno per uno attraverso metodi dedicati. Vengono poi accorpati in un'unica stringa affinché vengano inseriti nella tabella finale come un unico *record*. Per questo, come separatore viene utilizzato il simbolo "|" anziché la virgola.

```
try:
    w = whois.whois(url)
    resp  = str(w.name) + ' | ' + str(w.address) + ' | ' + str(w.city) + ' | '
+ str(w.zipcode)+' | ' + str(w.country)

    if type(w.emails) is list :
       for mail in w.emails:
           resp = resp + ' | ' + str(mail)
    else:
        resp = resp + ' | ' + w.emails
except:
     resp = 'non trovato'
row.append(resp)
```

4.7.7 Conteggio occorrenze *keyword*

Questa parte del programma effettua la ricerca esatta delle parole di cui si vuole verificare la presenza all'interno della pagina *web*. Prima di tutto viene verificato se l'utente ha effettivamente inserito almeno una parola in *input* quando è stato richiesto di inserire le *keyword* (semplicemente controllando che la lista passata come parametro `keyword` della funzione non sia vuota). Successivamente, la pagina viene parsificata tramite Beautiful Soup.

Per l'individuazione delle `keyword` nel testo è stata usata una combinazione del metodo `.find_all()` di Beautiful Soup e delle espressioni regolari. La ricerca che viene effettuata è di tipo esatto, cioè cerca le occorrenze che corrispondono esattamente alla stringa (se si cerca "processo", non verrà conteggiata la parola "processi"). La ragione è che, oltre a rendere il programma meno complesso, la ricerca esatta consente di fare ricerche più specifiche, in particolare se ciò che si sta cercando non sono tanto nomi di cose, ma nomi e cognomi di persona, di cui è necessario verificare la corrispondenza esatta.

```
if keywords:
    try:
        page_soup = BeautifulSoup(page.text, 'html.parser')
    except:
        pass
    for key in keywords:
```

```
        try:
            trova = page_soup.find_all(string = re.compile(key, re.IGNORECASE))
            occor = len(trova)
        except:
            occor = '-'
        row.append(occor)
```

4.7.8 Stampa dei dati ottenuti e creazione della matrice

Una volta eseguito il recupero di tutti i dati richiesti per un solo risultato di ricerca, quello che abbiamo è una lista, assegnata alla variabile `row`, che si è riempita progressivamente e nel giusto ordine di tutti i dati raccolti.

Per prima cosa, al fine di offrire all'utente un *feedback* in tempo reale dell'avanzamento del programma, la lista viene stampata sulla finestra da cui è stato lanciato il programma (per separare i risultati l'uno dall'altro, viene stampato. Successivamente la lista viene inserita nella lista `matrix`, parametro della funzione inizialmente vuoto, che progressivamente si riempie delle liste che corrispondono ai dati di ciascun risultato di ricerca formando in questo modo la matrice che verrà restituita dalla funzione alla fine dell'operazione.

```
print (row)
print ('-')
matrix.append(row)
```

4.7.9 Ricerche correlate e restituzione della matrice

La funzione `scrap` viene infine invocata ricorsivamente per ogni ricerca correlata che viene trovata in fondo alla pagina (se richiesto). Come parametro `matrix` viene fornita la matrice già riempita con i dati raccolti precedentemente. Tra un invocazione e l'altra della funzione vengono lasciati passare 22 secondi: questo impedisce che Google Search sia "bombardato" da richieste HTTP e blocchi l'indirizzo IP da cui vengono inviate.

Infine, la matrice completa viene restituita con istruzione `return`.

```
if corr > 0:
        corr -= 1
        correlate = serp_soup.find_all('p', class_='_Bmc')
```

```
        for p in correlate:
            sleep(22)
            scrap(p.get_text(), keywords, results, corr, matrix)


return matrix
```

Capitolo 5

Strategia reputazionale sottrattiva

5.1 Cancellare contenuti dal Web è possibile?

Applicare l'oblio significa in via ideale cancellare ricordi digitali dalla memoria del Web, quindi, in altre parole, eliminare fisicamente le informazioni sulla Rete. Le società che si occupano di reputazione online, usano spesso termini come "*identity cleaning*", a volte associando immagini di gomme da cancellare, come se esistesse effettivamente una tecnica per ripulire l'identità digitale raschiando via le informazioni scomode che la imbrattano.

Da un punto di vista pratico, rimuovere un'informazione dal Web significa cancellare il documento che la contiene, che esso sia una pagina HTML (o anche una sola parte di essa), un *file* immagine, un video, un PDF ecc., dal *server* in cui è ospitato. Di per sé il processo è banale, di gran lunga più facile che rimuovere un ricordo "analogico", visto che un modo per annientare un ricordo dal cervello di un essere umano non risulta essere ancora stato trovato.

Il problema non è come cancellare contenuti dal Web, ma chi può farlo. Le informazioni sono contenute nei *server* e solo chi è autorizzato al loro accesso può rimuoverle. Di fatto, il potere esecutivo dell'applicabilità dell'oblio risiede nel proprietario del *server* e di chi ne ha affittato lo spazio, cioè il *webmaster* (con cui si intende il proprietario del sito o chi è incaricato di gestirlo, due figure che possono coincidere). Il primo passo da fare per rimuovere un contenuto dal Web è chiedere al *webmaster* di eliminarlo dal

suo sito. Questo può essere fatto sostanzialmente i tre modi, ognuno dei quali ha una variabile probabilità di successo:

a in via amichevole e informale,

b per vie legali,

c tramite estorsione.

Il metodo c) è semplicemente illegale, indipendentemente dalle aspettative di successo della strategia. Il metodo a) è sicuramente il più veloce ed economico. Basta una telefonata o un *e-mail*, al massimo può valere il costo di un caffè o di una cena nell'improbabile eventualità che si incontri personalmente il *webmaster* e si cerchi di farselo amico. Le probabilità di successo dipendono molto dalla propensione del *webmaster* ad accettare la richiesta e dalla sua comprensione del problema. Se pensa che quel contenuto abbia tutti i diritti di stare sul suo sito e che la rimozione limiterebbe la sua libertà di espressione o quella dei suoi utenti non c'è niente da fare, il contenuto resterà lì dov'è.

Diverso è il caso del metodo b), in cui la decisione del *webmaster* è subordinata alla Legge del Paese in cui vive e si rifà alle responsabilità dell'*internet content provider*. Il successo dipende da che tipo di azione legale è stata fatta: denunciare il sito o fargli causa tramite avvocati può essere sufficiente a persuadere il *webmaster* ad accogliere la richiesta, ma può anche innescare una battaglia legale dagli esiti incerti, un'ingiunzione del tribunale non lascerebbe scampo al *content provider*, salvo eventuali ricorsi. L'applicabilità della legge dipende inoltre anche dalla posizione geografica della parte offesa, del *webmaster* e persino del *server* che ospita il contenuto: se siamo in Italia e il proprietario del sito è americano (e il *server* anche) non si può pensare di far valere la legge italiana per un soggetto che vive in America; se il *webmaster* vive e opera in Italia sottostà alle leggi italiane e quindi, se deciso dalle istituzioni giuridiche, dovrà rimuovere il contenuto. Tuttavia, se il sito si trova su un *server* in America il suo proprietario potrebbe non essere tenuto a eliminare alcun contenuto qualora vi restasse per qualche motivo.

Vanno inoltre considerati i tempi e i costi dell'azione legale: processi che durano anni e le spese per eventuali avvocati.

Per ultimo, è bene non sottovalutare le possibili reazioni del *webmaster* nel momento in cui si trovasse costretto a rimuovere controvoglia il contenuto,

magari dopo aver pagato una multa e delle spese legali: potrebbe vendicarsi, trovare il modo di diffondere nuovamente notizie lesive e magari in modo del tutto legale, scatenare una tempesta mediatica che metterebbe sotto i riflettori proprio quello che si cercava di celare. Nell'eventualità (remota) che il contenuto interessato venga rimosso, la sua presenza nel Web potrebbe non essersi ancora estinta del tutto. Una parte o l'intero contenuto potrebbe essere stato quotato in un altro sito, la pagina potrebbe trovarsi nella *cache* di qualche motore ricerca, o potrebbe rimanere indicizzata per molto tempo dopo la sua eliminazione, comparendo comunque nei motori di ricerca.

Rimuovere un'informazione alla sua fonte è dunque un'operazione non banale, con scarsissime possibilità di successo e a volte non del tutto risolutiva per l'oblio. Se un contenuto non può essere rimosso è sempre vero che limitare la sua accessibilità agli utenti potrebbe ridurre notevolmente la possibilità che questo venga trovato. Gli utenti difficilmente visitano pagine *web* conoscendo le *directory* in cui sono allocate a memoria, ma lo fanno attraverso i *link*. Se non è possibile rimuovere *link* da una pagina *web*, perché ricondurrebbe agli stessi problemi che sono stati descritti per la rimozione di un contenuto, è invece possibile fare richiesta di rimozione di un *link* dai risultati di un motore di ricerca. Ottenere la deindicizzazione di una pagina *web* in relazione a *keyword* specifiche.

5.2 Rimuovere *link* dai motori di ricerca: le linee guida di Google

Il primo a fare luce sui processi di rimozione di un contenuto sul Web o dai risultati del motore di ricerca è Google stessa, che in una pagina di supporto di Google Search Console, intitolata non a caso *Rimozione di informazioni da Google*[1], offre una panoramica discretamente dettagliata sul tema in questione. La guida si divide in due parti che corrispondo a due fasi del processo di rimozione:

- Fase 1: identificare il problema
- Fase 2: intervenire

Nella Fase 1 Google chiede di rispondere a due domande fondamentali:

[1]Google, *Rimozione di informazioni da Google*, https://support.google.com/webmasters/answer/6332384?hl=it, 19/4/2017.

1. *Vuoi rimuovere qualcosa solo dai risultati di Google Search oppure dall'intero Web?*

2. *Chi controlla la pagina che ospita il contenuto?*

La prima domanda prevede una risposta pressoché scontata. Se si vuole eliminare un *link* dai risultati di Google Search è ovvio che saremmo ancora più felici se fosse possibile cancellare il contenuto a cui esso rimanda direttamente alla fonte. L'unico caso in cui ci si può occontentare unicamente della rimozione del risultato di Google è che si tratti di un *link* obsoleto, che punta cioè ad una pagina che è già stata rimossa.

La domanda 1 in realtà è un modo di Google per cogliere l'occasione per ricordare che egli stesso non è il Web, ma un semplice motore di ricerca:

> Ricorda che Google non è il Web, ma soltanto un motore di ricerca che consente di trovarvi le informazioni. Rimuovere contenuti dalla pagina dei risultati di ricerca di Google non significa rimuoverli dal Web. Anche se riesci a far rimuovere informazioni dai risultati della Ricerca Google, è possibile che tu debba svolgere altri passaggi per farle rimuoverle dal Web.

La domanda 2 costituisce uno dei temi fondamentali della tutela dell'identità digitale: quale controllo si ha sulle proprie informazioni presenti in Rete? Le informazioni nei risultati di Google Search sono reperite da una fonte: spesso, anche se non sempre, da una pagina *web*. È necessario individuare chi controlla questa fonte prima di capire come bloccare o rimuovere l'informazione.

Tipiche pagine sotto il nostro controllo sono: i nostri profili *social*, il proprio sito *web* o *blog*, una pagina Google my Business della nostra azienda.

Esempi di pagine fuori dal nostro controllo sono invece: un articolo su una testata *online*, il *blog* di qualcun altro, una voce di Wikipedia che parla di noi oppure il profilo *social* che non sia il nostro. Ovviamente, se i contenuti sono ospitati su pagine di nostra proprietà rimuoverle o correggerle è semplice, anche se questo non è sempre assicurato, in quanto l'intervento su un *forum*, il caricamento di una foto su un *social network* o il commento a un *post* possono a volte sfuggire al nostro controllo, persistere all'interno della piattaforma ospitante anche dopo un nostro intervento di rimozione. Se l'informazione non si trova su un canale gestito da noi le cose si fanno molto più difficili, come è facile immaginare.

Il processo di rimozione di *link* da Google Search dipende, ribadisce Google nel momento in cui inizia ad esporre la fase 2, dal controllo che si ha sulla pagina che ospita l'informazione. Se il contenuto è su un sito o un *account* che controlliamo o si è il proprietario verificato del sito (cioè la proprietà di un sito è stata registrata su Google Search Console) è possibile:

1. nascondere temporaneamente l'informazione da Google Search rimuovendo in via non definitiva l'URL in questione dai risultati di ricerca attraverso apposita richiesta tramite Google Search Console.[2]

2. Rimuovere permanentemente l'informazione da Google Search utilizzando il *tag* `<meta name="robots" content="noindex">`o `<meta name="googlebot" content="noindex">` (quest'ultimo vale solo per i *crawler* di Google) oppure il *file* `robots.txt`.

Se invece il contenuto non è sotto il nostro controllo, si aprono due scenari:

1. Il contenuto non esiste più sulla pagina *web*: è possibile fare richiesta a Google di rimuovere il contenuto non più esistente dai propri risultati di ricerca tramite apposita funzione "Rimuovi contenuti obsoleti".[3]

2. Il contenuto esiste sulla pagina *web*: se l'informazione è presente su una piattaforma di proprietà di Google (come YouTube o Blogger) è possibile fare richiesta di rimozione a Google stessa. Nel caso invece si trovi su una proprietà non di Google viene consigliato di contattare il proprietario della pagina *web* e chiedere di rimuovere il contenuto. Una volta rimosso il contenuto fare richiesta a Google di rimuoverlo dai risultati di ricerca come descritto al punto 1. Aver rimosso l'informazione dai risultati di ricerca di Google non significa certo averla resa del tutto irreperibile: il *link* alla pagina che la contiene potrebbe essere presente da un'altra parte, come un'altra pagina *web*, un *post* di un *forum* o *social*, o semplicemente la cronologia di qualche utente. Se davvero si vuole nascondere o rimuovere informazioni è necessario intervenire sulla fonte. Anche in questo caso, la prima cosa da definire è chi controlla la fonte dei contenuti che si vogliono eliminare. Se la

[2]Google, *Search Console*, https://www.google.com/webmasters/tools/url-removal, 19/4/2017.

[3]Google, *Rimuovi contenuti obsoleti*, https://www.google.com/webmasters/tools/removals, 19/4/2017.

pagina è in nostro possesso è possibile rimuovere o modificare il contenuto autonomamente, in caso contrario sarà necessario contattare direttamente il proprietario del sito e chiederne la rimozione.

Google dichiara che è possibile fare richiesta di rimozione di alcuni dei suoi risultati, tuttavia resta sua discrezione accogliere o meno la richiesta. Solitamente, vengono rimossi contenuti che vanno contro la *policy* di Google o in seguito a una valida richiesta legale. Per quanto riguarda informazioni personali, quelle che di solito è disposta a eliminare, stando a quanto riporta sulla sua pagina di "Removal Policies"[4] sono numeri di identificazione, numeri di conto bancario e carta di credito, immagini della propria firma o immagini di nudo o di sesso esplicito che sono state caricate senza il proprio consenso.

Per il resto è probabile che le possibilità di successo della propria richiesta di rimozione riguardino esclusivamente pagine che violano le linee guida di Google, per le quali mette a disposizione *form* specifici per ogni tipologia di infrazione:

- spam (http://www.google.com/webmasters/tools/spamreport.)
- vendita e acquisto di *link* (https://www.google.com/webmasters/tools/paidlinks)
- phishing (https://www.google.com/safebrowsing/report_phish/?hl=it)
- *spam* nei *rich snippets* (https://support.google.com/webmasters/contact/rich_snippets_spam?l

Per tutte le richieste di rimozione che non riguardano violazioni esplicite delle *policy* di Google esiste invece il seguente *form* "Rimozione di contenuti da Google"[5], che attraverso una procedura guidata reindirizza verso la specifica pagina rimozione di cui si ha bisogno.

Per prima cosa viene richiesto quale è il servizio Google che contiene l'informazione da rimuovere (probabilmente, si tratta di Google Search, ma nella lista sono presenti tutti i servizi, come YouTube, Google Images, inserzioni AdWords e molto altro). Nella seconda fase viene richiesto che tipo di contenuto si vuole segnalare. Se appunto non si tratta di violazioni esplicite o servizi quali Google My Business o Knowledge Graph, per le quali ci sono

[4]Google, *Removal Policies*, https://support.google.com/websearch/answer/2744324?hl=en, 19/4/2017.

[5]Google, *Rimozione di contenuti da Google*, https://support.google.com/legal/troubleshooter/1114905?hl=it/ 19/4/2017.

procedure apposite, la voce da selezionare è "Vorrei rimuovere le mie informazioni personali dai risultati di ricerca di Google" e di seguito indicare la motivazione della rischiesta di rimozione tra una serie di opzioni:

- *Vorrei chiedere al webmaster di una pagina dei risultati di ricerca contenente informazioni non corrette o imprecise di rimuovere completamente la pagina dai risultati di ricerca di Google.* Da utilizzare quando si vuole sapere come contattare il *webmaster* di un sito per richiedere la rimozione del contenuto. L'opzione rimanda alla guida Google su come trovare le informazione di contatto di un sito.

- *Il mio nome e cognome o il nome della mia azienda sono visualizzati in un sito per adulti che genera spam nei risultati di ricerca di Google.* Da utilizzare se le informazioni personali sono associate a siti con contenuti per adulti.

- *Vorrei rimuovere le mie informazioni personali riservate dai risultati di ricerca di Google.* Da utilizzare quando informazione pubblicate riguardano dati sensibili. Gli altri casi di richiesta di rimozione dei contenuti si dividono in due tipologie:

- Violazioni di natura legale. Contenuti che riguardano la persona o azienda interessata e per i quali esista già un provvedimento di natura legale (in corso o terminato).

- Violazioni di natura personale. Questo tipo di richieste fa capo alla legge europea per la protezione dei dati personali (Diritto all'Oblio) che prevede che i singoli individui possono chiedere ai motori di ricerca di rimuovere specifici risultati che appaiono effettuando una ricerca con il proprio nome, qualora tali risultati siano relativi all'interessato e risultino obsoleti.

5.3 Il modulo di richiesta di rimozione di Google

5.3.1 Un tentativo concreto di rimozione

A seguito della sentenza del maggio 2014, Google ha implementato un *form* per gestire la *Richiesta di rimozione di risultati di ricerca ai sensi della legge europea per la protezione dei dati personali.* Per poter essere rimossi, la

richiesta deve provenire da un cittadino europeo e i risultati associati alla ricerca del proprio nome devono essere giudicati da Google come «obsoleti, inadeguati, irrilevanti o non più rilevanti».

Per fare richiesta basta fornire le proprie credenziali personali e compilare il *form* specificando:

- Tutti gli URL che di desidera rimuovere dall'elenco dei risultati di ricerca
- Le motivazioni per cui le informazioni da rimuovere riguardano personalmente il soggetto interessato
- Il motivo per cui si ritiene che l'inserimento nei risultati di ricerca sia irrilevante, obsoleto o comunque discutibile. Questo per ciascun URL indicato nel modulo

Le richieste di rimozione ricevute da Google che rispettano i requisiti minimi vengono messe in una coda di produzione e valutate singolarmente. Alla fine l'individuo/agenzia di ORM[6] riceve un'*e-mail* con notifica dell'esito.

Viene riportato ora un esempio di invio di una richiesta e relativa risposta di Google.

La richiesta consiste nella rimozione di un *link* che compare tra la terza e la quarta posizione della prima pagina della SERP corrispondente al nome e cognome del soggetto richiedente. Il *link* riguarda un verbale ufficiale di un richiamo disciplinare relativo alla professione del soggetto, evento che si era svolto e concluso otto anni fa e non aveva avuto alcuna conseguenza. Tra la pubblicazione della notizia e oggi, il soggetto non esercita più la stessa professione.

La risposta di Google è arrivata molto velocemente, solo due giorni dopo la data di sottomissione del modulo di richiesta:

> Gentile xxxxx,
> La ringraziamo per il suo messaggio.
> In merito ai seguenti URL:
> xxxxxxxxxxxxxxxxxxxxxxxx
> I contenuti in questione riguardano la mansione o professione che potrebbe esercitare attualmente. Per consentirci di valutare in maggiore

[6] Agenzie di Online Reputation Management, il modulo infatti può essere compilato e inviato sia dal soggetto richiedente sia da una persona delegata, come un legale o un consulente reputazionale.

dettaglio la sua richiesta, la invitiamo a rispondere indicando la sua professione attuale. Se i contenuti riguardano la sua esperienza professinaale precedente, la preghiamo di confermare la data in cui si è conclusa. Sarebbe utile se potesse fornire documentazione in merito.

Coridali saluti,
Il team di Google

Una volta forniti con un'altra e-mail i dati richiesti, la seconda risposta di Google è stata questa:

Gentile xxxxx,
La ringraziamo per il suo messaggio.
In merito ai seguenti URL:
xxxxxxxxxxxxxxxxxxxxxxxx

In questo caso sembra che gli URL da lei identificati includano informazioni su di lei pertinenti e aggiornate. Di conseguenza, siamo giunti alla conclusione che la menzione del materiale in questione nei nostri risultati di ricerca sia giustificata dall'interesse del pubblico di avervi accesso.

Abbiamo quindi deciso di non prendere alcun provvedimento in merito all'URL in questione.

Qualora non fosse soddisfatto della decisione presa da Google, potrebbe avere il diritto di sottoporre la questione all'autorità per la protezione dei dati del suo paese. Può includere nella comunicazione all'autorità il numero di riferimento xxxxxxxxxxxxx e una copia della conferma di invio del modulo relativo alla richiesta presentata a Google.

Può inviare la richiesta di rimozione direttamente al webmaster che gestisce il sito in questione. Il webmaster ha la possibilità di rimuovere i contenuti in questione dal Web o di impedirne la visualizzazione nei motori di ricerca. Può visitare il sito all'indirizzo https://support.google.com/websearch/answer/9109?hl=it per sapere come contattare il webmaster di un sito.

Se nei risultati della Ricerca Google vengono ancora visualizzati contenuti obsoleti di un sito, può chiedere a Google di aggiornare o rimuovere la pagina utilizzando lo strumento per le richieste di rimozione di pagine web all'indirizzo: http://www.google.com/webmasters/tools/removals

Cordiali saluti,
Il team di Google

Alla richiesta da parte del soggetto richiedente di ulteriori informazioni sulle ragioni della decisione di Google, questa è stata la risposta:

> Gentile xxxxx,
>
> La ringraziamo per il suo messaggio.
>
> Abbiamo ricevuto ed esaminato il suo reclamo. Al momento, Google ha deciso di non prendere provvedimenti in base alle norme relative alla rimozione dei contenuti. Come sempre, la invitiamo a risolvere eventuali controversie direttamente con il proprietario del sito web in questione.
>
> Se intenta contro questo sito un'azione legale che comporta la rimozione del materiale incriminato, i nostri risultati di ricerca rispecchieranno la modifica alla nostra successiva scansione del sito. Se il webmaster apporta tali modifiche e lei desidera sollecitare la rimozione della copia cache, invii la sua richiesta utilizzando il nostro strumento per la richiesta di rimozione di pagine web all'indirizzo: http://www.google.com/webmasters/tools/removals.
>
> Il posizionamento dei siti nei nostri risultati di ricerca è determinato automaticamente in base a una serie di fattori, descritti in dettaglio alla pagina http://www.google.com/insidesearch/howsearchworks/index.html. L'associazione di parole chiave a siti non è un'operazione manuale, e non modifichiamo il posizionamento di nessun sito nelle pagine dei risultati di ricerca di Google.
>
> In genere, i webmaster possono migliorare il posizionamento aumentando il numero di siti di qualità che si collegano al proprio sito. Per maggiori informazioni su come migliorare la visibilità del suo sito nelle pagine dei risultati di ricerca di Google, ti consigliamo di prendere visione delle nostre Istruzioni per webmaster all'indirizzo: http://www.google.it/support/webmasters/bin/answer.py?answer=35769. In questa pagina sono descritti i criteri chiave per la creazione e la gestione di un sito Web conforme alle indicazioni di Google.
>
> Cordiali saluti,
> Il team di Google

Al momento dell'invio della modulo di rimozione, la reazione di Google si è sviluppata in 3 fasi:

1. Richiesta di maggiori informazioni sulla correlazione tra la notizia e lo stato attuale del soggetto

2. La comunicazione della decisione presa dal *team*

3. Ulteriori spiegazioni su sollecitazione del soggetto richiedente

La risposta 1 è chiaramente precompilata, lo si evince dal linguaggio generico di tutto il messaggio e dal fatto che si riferisca a più URL, quando era stato chiesto di eliminarne uno solo. Tuttavia, dimostra che il *team* responsabile della valutazione delle domande esamina ogni richiesta singolarmente ed è disposta a documentarsi sulla situazione del richiedente in relazione a quanto è contenuto nella notizia lesiva.

La risposta 2 è quella più importante, in questa è contenuta la decisione di Google di accogliere o meno la richiesta di rimozione. Il messaggio è composto da due blocchi fondamentali:

a Le ragioni della non accettazione della richiesta

b Quali sono le possibili mosse alternative per ottenere la rimozione del contenuto.

Anche in questo caso la genericità degli argomenti fa pensare che si tratti di un testo precompilato. Va considerato che se già la prima risposta era stata velocissima, la risposta 2 è arrivata il giorno dopo aver inviato i dati richiesti. Di fatto Google ha stabilito in meno di 24 ore quanto il soggetto richiedente "meritasse" di essere dimenticato per un evento complesso avvenuto otto anni prima. Da notare nel blocco a) come l'argomentazione della decisione di Google marchi l'accento sul diritto dell'informazione, concludendo che la presenza del *link* nei risultati di ricerca è giustificato "dall'interesse pubblico ad avervi accesso".

Nel blocco b) Google, dopo aver giustificato l'inaccettabilità della richiesta offre alcuni consigli per la risoluzione del problema indicando i due canali più naturali per la questione: l'autorità locale della *privacy* e il *webmaster* del dominio che ospita il contenuto. In poche parole Google ci rimanda al punto di partenza: convincere il *webmaster* a rimuovere il contenuto, o contattandolo in via informale, o attraverso un'azione legale. La responsabilità dell'oblio risiede nell'editore dell'informazione e nella legge, la persona interessata e il motore di ricerca hanno un ruolo marginale nella questione.

In pratica Google si libera della patata bollente e la reindirizza a coloro che ritiene i veri responsabili della questione, ma allo stesso tempo assolve il

suo mandato attribuitogli dalla Corte Europea offrendo delle soluzioni (per quanto di fatto non siano risolutive).

Il concetto viene ribadito nella risposta 3, che nel secondo paragrafo dice, parafrasando, «riprenderemo in considerazione il suo caso solo se sarà la Legge a chiedercelo esplicitamente con un atto del tribunale». In alternativa, solo dopo la rimozione della pagina da parte del *webmaster* Google sarà disposta ad eliminare il *link* dai propri risultati di ricerca, dal momento che quel *link* è obsoleto visto che rimanda ad una pagina ormai inesistente. In questo caso è possibile sollecitare la rimozione del *link* attraverso la segnalazione di indirizzi obsoleti tramite il modulo fornito alla pagina indicata.

Davvero molto interessante è il seguito del messaggio: prima, Google chiarisce che il *team* non può intervenire manualmente sul posizionamento dei risultati sulla pagina del proprio motore di ricerca, ma poi, nel secondo paragrafo, suggerisce che i *webmaster* possono fare in modo che i propri siti si posizionino meglio (e cita sommariamente l'esempio del *link building*, accennando alla possibilità di aumento dei collegamenti del proprio sito da siti di alta qualità). Questo ultimo paragrafo è piuttosto singolare, perché sembra non c'entrare molto con il tema in questione. Cosa c'entra il posizionamento di un sito con la rimozione di un *link* di una pagina che, per giunta, non appartiene a quel sito?

Intanto è evidente che questa volta non si tratta di una risposta precompilata, lo dimostra il fatto che ci sia un errore grammaticale: se si legge attentamente la frase

> Per maggiori informazioni su come migliorare la visibilità del suo sito nelle pagine dei risultati di ricerca di Google, ti consigliamo di prendere visione delle nostre Istruzioni per webmaster all'indirizzo: http://www.google.it/support/webmasters/bin/answer.py?answer=35769.

si può notare come nello stesso periodo chi scrive si riferisca al destinatario prima dandogli del lei («il suo sito») e poi gli dia del tu («ti consigliamo»). Questo dettaglio fa pensare che il messaggio sia stato scritto in tempo reale. La ragione è che probabilmente questa *e-mail* è una risposta ad una domanda non standard. Il soggetto aveva infatti chiesto se fosse possibile, visto che la richiesta di rimozione era stata negata, almeno ottenere un posizionamento meno evidente nel *link* nella SERP.

Chi ha risposto, ha dato un'indicazione piuttosto approssimativa. Non chiarisce infatti quale sia la correlazione tra la deindicizzazione di un conte-

nuto indesiderato e il miglioramento della qualità di un sito per gli standard di Google.

La realtà è che, non si sa se volutamente o no, chi scrivendo ha fatto le veci di Google ha fornito al soggetto la chiave per contrastare il contenuto lesivo con le sue stesse forze. In pratica il messaggio complessivo di Google è questo: «io non accetto di cancellare il risultato per garantire il diritto di cronaca, non giudico la pagina *web* in virtù del suo contenuto ma in base alla sua qualità stabilita dal mio *crawler*, se tu saprai fornirmi un contenuto di miglior qualità rispetto a quello che vuoi rimuovere è possibile che quest'ultimo sia meno visibile».

5.3.2 Perché così tanti rifiuti?

Il fatto che la richiesta non sia stata accettata dal motore di ricerca, malgrado l'evidente obsolescenza del contenuto in questione, non deve sorprendere: in Italia ben il 70,5% delle richieste non vengono accettate.

Stando ai dati di fine 2015, in Europa erano state fino ad allora ricevute da Google 333.450[7] richieste di rimozione e sono state esaminate 931.223 pagine *web*, per una media di 3,4 pagine per richiesta. Questo significa che nel giro di un anno e mezzo dopo la sentenza della Corte Europea Google si è trovata a dover esaminare singolarmente più di 600 casi al giorno, valutando quotidianamente più di 1700 pagine *web*. Una domanda nasce subito spontanea: sono in grado di gestire questi numeri, soprattutto se si pensa che Google si è trovata costretta a fare questo tipo di valutazioni?

Delle 333.450 richieste ne sono state accolte il 41,9%, meno della metà. Va inoltre precisato che all'interno di una stessa richiesta alcuni *link* potrebbero essere rimossi e altri no: non è la richiesta che non viene accettata in sé, ma viene valutato ogni singolo contenuto e per ognuno di questi presa una decisione. Una domanda di rimozione può dunque ottenere l'eliminazione di tutti gli URL richiesti, una parte di questi oppure nessuno di questi.

In Italia sono 25.118 le richieste inviate (sempre dati del 2015) e solo il 29,5% sono state accettate. La differenza rispetto alla percentuale europea è davvero molto alta, ben tre casi su quattro vendono respinti.

[7]Dati riportati da Barchiesi, *La tentazione dell'oblio*, cit. Un articolo più recente del New York Times parla di 418.000 richieste nell'aprile 2016 (Scott, *Europe Tried to Rein In Google. It Backfired*, cit.) in Europa, ma non riferisce quante ne siano state ricevute da utenti italiani.

Purtroppo, l'idea che l'utente, a partire dalla sentenza Google Spain, avrebbe potuto rimuovere con un semplice *form* le informazioni indesiderate, si è rivelata una falsa promessa. L'esistenza del modulo di rimozione e la sua ampia accessibilità hanno creato una falsa aspettativa di facilità dell'operazioe e certezza del risultato.

Secondo Barchiesi, è probabile che buona parte dei rifiuti sia imputabile a richieste non del tutto accurate accurate o magari del tutto fuori luogo: spesso, infatti, le persone non hanno ben chiaro il raggio d'azione del diritto all'oblio. Il modulo di Google contiene diversi campi da compilare, di cui non tutti sono di immediata comprensione per chi non conosce il linguaggio informatico (non è scontato, per esempio, sapere cos'è un URL). Già i primi campi non sono chiari: è necessario fornire sia il nome che ha prodotto i risultati di ricerca dai quali si vuole rimuovere il contenuto, sia il proprio nome e cognome anagrafico. Certamente nella maggior parte dei casi i due campi coincidono, ma è possibile che la stringa di ricerca contenga più termini rispetto al nome. Cioè se Mario Rossi digita "Mario Rossi Pisa" appare un elenco di risultati in cui compare un *link* lesivo che vorrebbe rimuovere non è detto che succeda anche se digita semplicemente "Mario Rossi". È necessario quindi che Mario Rossi sappia distinguere tra la *query*, che nel modulo è ambiguamente indicata con "Nome", e il suo nome vero e proprio. Se ciò non accadesse e scrivesse solo il suo nome e cognome anche nel campo relativo alla ricerca, invierebbe una richiesta nulla perché il *link* che indica non comparirebbe nei risultati di ricerca associati.

Inoltre, chiarire i motivi per cui si vorrebbe rimuovere un contenuto non è banale, soprattutto per chi non ha ben chiaro in quali casi sia effettivamente applicabile il diritto all'oblio.

La richiesta sopra riportata in ogni caso è stata compilata in maniera corretta ed esaustiva. È probabile che la sua non accettazione dipenda da altri fattori. Bisogna capire per esempio quali siano le difficoltà di Google nel rispondere alle richieste e valutarle in modo opportuno.

Intanto, il modulo viene compilato nella lingua nazionale, in questo caso l'italiano. Questo presuppone che ci sia un *team* per ogni nazione che sappia bene la lingua e al tempo stesso sia competente in materia. Come si è visto, ogni caso viene valutato singolarmente attraverso l'analisi e l'approfondimento di documentazione riguardante la sfera professionale o personale del soggetto richiedente e il contesto in cui si colloca il contenuto che intende

rimuovere. La documentazione sarà molto probabilmente in italiano e richiede di avere una comprensione e consapevolezza della cronaca e della cultura del paese. Le verifiche possono perciò essere molto complesse. A questo si aggiungono l'eterogeneità delle richieste, che impedisce la standardizzazione del processo complicandone e rallentendone la gestione, e la frequente incompletezza o inadeguatezza dei moduli inviati che affaticano ulteriormente il lavoro del *team* responsabile.

5.4 Le ragioni per cui l'oblio non funziona

Il fatto che una richiesta di rimozione venga accolta non decreta necessariamente il successo dell'operazione di cancellazione dell'evento dalla memoria digitale (e di conseguenza, dalla memoria collettiva). Una cancellazione di un *link* dalla pagina di Google potrebbe non risolvere il problema o in certi casi persino peggiorarlo.

In primo luogo, la rimozione di un *link* per esercizio del diritto all'oblio vale soltanto per l'Unione Europea. Lo stesso concetto di diritto all'oblio si è sviluppato solo in Europa, che ripongono nel concetto di *privacy* la salvaguardia della dignità dell'individuo. In America, che è la patria di Google, non c'è un "diritto ad essere dimenticati", piuttosto un "*right to be let alone*"[8] : l'accento si sposta molto più sulla libertà dell'individuo e di conseguenza sulla sua libertà di esprimersi e di essere informato, mentre è messa da parte la tutela della riservatezza del singolo. Il diritto all'oblio non esiste negli Stati Uniti e probabilmente mai esisterà nel futuro. Un *link* rimosso da Google.it probabilmente rimarrà su Google.com.

Il dominio riceve una notifica nel caso una sua pagina venga deindicizzata di proposito. Questa notifica comunica esplicitamente che la pagina in questione è stata eliminata dai suoi indici in seguito ad una richiesta di qualcuno che ha esercitato il suo diritto all'oblio. La reazione dell'editore potrebbe scatenare una tempesta mediatica dall'impatto ben maggiore della pagina eliminata. È successo che in alcuni casi la pagina venisse in seguito reindicizzata perché finita sotto i riflettori e dunque molto visitata e condivisa.

Va ricordato che il contenuto viene solo deindicizzato, non eliminato. Nessuno impedisce ad un utente di visitare quella pagina o di trovarla attra-

[8]Warren, Brandeis, *The right to privacy*, Harvard Law Review, IV, 5, 2015.

verso una ricerca più approfondita. La deindicizzazione inoltre è solo parziale. Viene effettuata solo per quella determinata *query* indicata nel modulo di rimozione. La pagina potrebbe essere visibile su altre *query*, magari poco differenti da quella per cui è avvenuta la deindicizzazione.

I servizi di Autocomplete, Ricerche Correlate e Istant Search potrebbero, di conseguenza, suggerire *query* per cui compare il *link* alla pagina indesiderata, o addirittura contenere parole che suggeriscono il contenuto stesso della pagina.

Per ultima cosa, bisogna considerare che esistono siti che raccolgono *directory* dei *link* rimossi. Alcuni "militanti" della libertà di espressione e di informazione sono molto contrari alla cancellazione di contenuti *online*, soprattutto se riguardanti personaggi d'interesse pubblico, perciò raccolgono i *link* delle pagine deindicizzate da Google e li rendono disponibili su siti dedicati. Se un *link* che ci riguarda finisce su uno di questi siti, lo svantaggio non sta solo nel suo recupero e conseguente vanificazione dell'operazione di rimozione, ma significa anche essere inseriti in una sorta di lista nera in compagnia di migliaia di altri *link* di scandali, reati e contenuti imbarazzanti.

Tutte queste ragioni non sono affatto trascurabili nel momento in cui si decide di cancellare una propria informazione dai motori di ricerca. L'oblio, così come Google lo applica, costituisce nella maggior parte dei casi più un palliativo che la vera soluzione ad una crisi reputazionale. Certamente, il rischio di peggiorare la situazione in seguito ad una rimozione riguarda maggiormente personaggi di spicco, sui quali basta poco per accendere i riflettori su eventi recenti così come passati da molto tempo. Una persona non famosa può trarre più beneficio da un'operazione sottrattiva, in quanto ad una rilevanza pubblica equivalente della notizia bassa e la non appartenenza a gruppi sensibili corrisponde un rischio di reazione particolarmente moderato.

5.5 "Sparire da Internet": l'identità digitale nulla non è la soluzione

Se cancellare contenuti dal Web fosse più semplice di quello che effettivamente è, se l'oblio avesse un campo di applicazione ben chiaro da un punto di vista giuridico a tal punto da costituire una garanzia per l'individuo, si potrebbe contemplare una radicalizzazione del metodo sottrattivo, ovvero

la totale cancellazione di ogni propria traccia sul Web, una vera e propria uccisione dell'Io *online*.

L'idea ha un suo fascino, perché significherebbe vanificare il controllo che l'Io *online* ha sull'Io *offline*, mettersi al sicuro da spionaggi governativi, da *hacker* ma anche semplicemente impedire che in qualsiasi modo avvenga la stretta di mano digitale, che come si è visto può avere conseguenze molto negative. Uccidere l'Io *online* significa scollegarsi definitivamente dalla realtà virtuale per vivere solo ed esclusivamente nella vita reale, che in una visione romanzata assomiglia molto a liberarsi dalla schiavitù delle macchine scollegandosi dal Matrix.

"Sparire da internet" non significa solo applicare l'oblio con la rimozione di *link* dai motori di ricerca o di contenuti direttamente dalle loro fonti, ma anche chiudere ogni *account* di *social network*, *forum* e *blog*, siti *e-commerce* e altri servizi che sfruttano la rete eliminando pian piano la nostra presenza da ogni piattaforma *web* (il che comporta anche la rinuncia al loro utilizzo). Alcuni *tool* possono aiutare nell'operazione, come per esempio il sito Deseat.me[9] che attraverso la ricerca in Google scova tutti gli *account* registrati a proprio nome, anche quelli dimenticati, e facilita l'accesso alle pagine di disiscrizione, che spesso sono tenute ben nascoste.

L'esercizio del diritto all'oblio unito ad una approfondita pulizia di tutti nostri profili *web* può andare molto vicino ad un'identità digitale nulla, anche se, come si è visto, gli attuali mezzi di applicazione dell'oblio non sempre sono risolutivi e una lunga presenza sui *social media*, magari non completamente responsabile e consapevole, può comportare la persistenza di video e foto a anche dopo la cancellazione definitiva del profilo. Ammesso che l'orizzonte dell'identità digitale nulla sia davvero possibile, non deve comunque far pensare che sia la miglior soluzione per metterci al sicuro da possibili crisi reputazionali digitali e che la nostra identità digitale non possa prendere o riprendere vita anche senza la nostra collaborazione.

L'idea che alla nostra identità digitale corrisponda un foglio bianco, un "404 *not found*", è allettante, ma occore capire cosa avviene, soprattutto sulla SERP di un motore di ricerca, quando si agisce per sola sottrazione.

La prima pagina dei risultati di ricerca di Google comprende solitamente 10 *link*. Si parta dal presupposto che un soggetto abbia cancellato ogni suo profilo ed ogni sua informazione su di sé che era sotto il suo diretto

[9]Deseat.me, https://www.deseat.me/, 19/4/2017.

controllo. Significa che in questi 10 risultati ci sono solo informazioni non riferibili al soggetto e informazioni riferibili pubblicate da domini non sotto il suo controllo. Se si suppone che ci sono 7 contenuti non riferibili e 3 contenuti riferibili non desiderati, si può applicare l'oblio su questi ultimi tre.

Nell'eventualità che questi tre contenuti vengano rimossi, si creano 3 lacune nella SERP. Queste vengono subito riempite dal *link* che era successivo nella lista, dando così origine ad uno spostamento verso l'alto di tutti i risultati per cui i primi tre della seconda pagina finiscono in fondo alla prima. Se tra questi ultimi c'è qualche contenuto lesivo la situazione non è migliorata, se non addirittura peggiorata. Il metodo sottrattivo implica inevitabilmente un movimento verso l'alto di tutti i contenuti che è del tutto casuale e fuori controllo.

Anche ammesso che il soggetto riuscisse ad eliminare tutti i risultati a lui riferibili, da tutte le pagine della SERP, nel momento in cui ne venisse creato uno nuovo e venisse indicizzato questo risulterebbe agli occhi di Google come il contenuto più rilevante in base alla *query* (che è il nome e cognome del soggetto) e quindi schizzerebbe immediatamente alle prime posizioni. Questo contenuto sarebbe quindi il primo che gli utenti troverebbero cercando il nome del soggetto, il primo ad effettuare la stretta di mano digitale con chiunque voglia. Se il contenuto è negativo, le conseguenze sono scontate ed essendo nuovo e attuale, l'applicabilità dell'oblio si abbassa notevolmente, per non dire che è completamente nullo.

È evidente che ad un'identità digitale nulla corrisponde una vulnerabilità massima a crisi reputazionali. Un foglio bianco è facile da scrivere e quel poco che verrà scritto sarà al massimo della sua leggibilità.

Inoltre, l'uccisione dell'Io *online* potrebbe stimolare l'indole investigativa degli internauti, che andrebbero alla ricerca del suo cadavere. In altre parole, non avere informazioni sul Web potrebbe creare sospetti, proprio come se si incontrasse qualcuno che ha distrutto tutti i suoi documenti, e potrebbe spingere le persone a fare ricerche più approfondite e magari trovare qualcosa che non dovrebbero.

L'Io *online* non è solo un mostro prodotto della Società dell'Informazione che va combattuto e distrutto, ma è l'interfaccia tramite cui noi interagiamo con altri soggetti, interazione che oggi, necessariamente, passa attraverso il canale digitale. Accettare la nostra presenza *online* (che oggi ormai è indispensabile e irrinunciabile per la maggior parte delle sfere professionali e

relazionali) e comprendere a fondo le sue dinamiche di generazione e azione, ci rende in grado di poter ottimizzare l'interfaccia e arrivare così sempre più vicini al controllo del nostro Io *online*.

Il metodo sottrattivo non può, e non deve, essere applicato se non inserito in un programma ben più ampio che agisce principalmente per addizione, cioè costruzione del proprio Io *online*.

Capitolo 6

Strategia reputazionale additiva

6.1 Il miglior posto dove nascondere un cadavere è la seconda pagina di Google

L'applicazione della strategia sottrattiva si è rivelata avere un campo di azione molto limitato, che si riduce tuttalpiù, una volta esclusa la possibilità di ottenere l'intervento sul sito sorgente del contenuto lesivo, all'invio del modulo di richiesta di rimozione a Google. Le probabilità di successo sono poche e la controllabilità delle conseguenze di un'eventuale rimozione è molto bassa.

Rimuovere informazioni dal Web è difficile, almeno quanto è facile produrle e condividerle. Se questa facilità è sfruttata in modo inconsapevole, è probabile che prima o poi si incorra in qualche danno nei confronti della propria reputazione o di quella altrui. Se invece le informazioni personali vengono immesse nella Rete in modo mirato e strategico, possono contribuire alla costruzione di un'immagine digitale positiva e coerente. Google tratta tutti i contenuti in cui si imbatte durante la sua attività di *crawling* allo stesso modo, ciò che gli interessa è il loro livello di qualità e rilevanza in base ai suoi fattori di valutazione.

Si prenda in considerazione un passo dell'ultimo messaggio inviato da Google in risposta alla richiesta di rimozione esaminata nel capitolo precedente:

> In genere, i webmaster possono migliorare il posizionamento aumentando il numero di siti di qualità che si collegano al proprio sito. Per maggiori informazioni su come **migliorare la visibilità del suo sito** nelle pagine dei risultati di ricerca di Google, ti consigliamo di prendere visione delle nostre Istruzioni per webmaster all'indirizzo: http://www.google.it/support/webmasters/bin/answer.py?answer=35769. In questa pagina sono descritti i criteri chiave per la creazione e la gestione di un sito Web conforme alle indicazioni di Google.

Visibilità: ecco la chiave della soluzione. La visibilità di una notizia è l'aspetto che più influisce sull'immagine digitale di una persona, specialmente sulla pagina dei risultati di ricerca. Se non si può agire per sottrazione su un contenuto negativo, si può agire sulla visibilità di altri contenuti, quelli sotto il controllo del soggetto (o di chi agisce per lui).

Agire sulla visibilità dei contenuti positivi significa ottimizzarli affinché Google li restituisca nelle posizioni più alte della SERP, con la conseguenza che i listati che erano già in alto siano costretti a scalare più in basso. La SERP è come una pila di oggetti con una forza che dal basso li spinge verso l'alto: se si tolgono degli oggetti, quelli sottostanti prenderanno il posto di quelli tolti, ma se se ne aggiungono altri in cima alla pila, gli oggetti già presenti saranno costretti a scendere in basso. Lo stesso risultato si otterrà sfilando degli oggetti dal basso e riaggiungendoli in alto.

C'è un detto, diffuso tra chi si occupa di *web marketing*, specialmente di SEO, che recita:

> Il posto migliore dove nascondere un cadavere è la seconda pagina di Google.

La frase ironizza sul fatto che se il proprio sito *web* compare nella seconda pagina delle ricerche, è come se fosse sepolto in un luogo dove non lo troverà mai nessuno: compito del *web marketer* è quello di "disseppellire" i propri *link* dalle pagine successive alla prima affinché compaiano nelle prime posizioni, diventando economicamente utili. Se tuttavia si legge questo detto dalla prospettiva di chi una pagina *web* la vuole nascondere, e non promuovere, il suo significato diventa ancora più efficace: nel caso la prima pagina della SERP contenga qualche "cadavere", disfarsene è difficile, non resta dunque che "seppellirlo" nella seconda pagina (o la terza, ancora meglio la quarta) ottimizzando contenuti positivi attraverso i criteri della *search engine optimization* (SEO).

L'idea non è nuova, si tratta della strategia più diffusa da parte delle agenzie che si occupano di ORM (*Online Reputation Management*) tanto che viene definita, impropriamente, *reverse SEO*. La definizione è impropria perché non esiste un sistema di deindicizzazione o di de-*ranking* di pagine non direttamente controllate (a parte alcune tecniche scorrette proprie della *negative SEO*[1]), ma esiste la possibilità di ottimizzare al meglio altri contenuti nel tentativo di scalzare quelli indesiderati.

Il manuale *The Art of SEO*, la "Bibbia" della disciplina per chiunque se ne occupa, dedica un piccolo paragrafo alla SEO per la reputazione, che definisce così:

> La SEO per la gestione della reputazione riguarda, in parte, il processo di neutralizzare menzioni negative del vostro nome nelle SERP. In questo tipo di strategia SEO, dovete sforzarvi di occupare posizioni aggiuntive nei primi 10 risultati in modo da spingere il listato critico più in basso e, auspicabilmente, fuori dalla prima pagina.
>
> La SEO permette questo processo attraverso la creazione di contenuti e la loro promozione tramite link, così come attraverso l'ottimizzazione del contenuto su piattaforme di terze parti come Pinterest, Facebook e LinkedIn. [2]

L'unica prospettiva per cui si può considerare la SEO per la reputazione *reverse*, è il fatto che mentre nella SEO "tradizionale" l'obiettivo è di norma il posizionamento di una pagina web su più *keyword*, nella strategia reputazionale si mira a posizionare più pagine web su una sola (o poche) *keyord*, ossia quella per cui compare un risultato compromettente (solitamente, il nome e cognome del soggetto).

«Diversamente dalle altre tattiche SEO - continua *The Art of SEO* - la *reputation management* comporta l'ottimizzazione di pagine su molteplici domini per far retrocedere i listati negativi. Questo comporta l'utilizzo di profili *social* e altre pagine su piattaforme di terze parti, relazioni pubbliche, comunicati stampa e *link* da reti di siti che si possiedono o si controllano, insieme alla classica ottimizzazione di link interni ed elementi *on-page*».

[1] Pratica SEO che consiste nel provocare penalizzazioni da parte dei motori di ricerca nei confronti di siti concorrenti. Vengono creati migliaia di *link* da pagine di pessima qualità che puntano il sito concorrente, il quale viene penalizzato dagli algoritmi *anti-spam* quale Google Penguin.

[2] E. Enge, S. Spencer, J. Stricciola, *The Art of SEO. Mastering Search Engine Optimization*, O'Reilly Media, Sebastopol, 2015, p. 150, trad. mia.

6.2 Applicazione della strategia additiva su un caso reale

6.2.1 Obiettivi dell'esperimento

Viene di seguito presentato un tentativo di correzione della reputazione di un soggetto reale sul motore di ricerca, applicando il metodo additivo per colpire un risultato indesiderato e migliorare complessivamente la prima pagina dei risultati relativa al suo nome e cognome.

L'obiettivo dell'esperimento è stato quello di "spostare" un risultato indesiderato che si trovava in quarta posizione oltre i primi 10 risultati di ricerca, ossia fuori dalla prima pagina della SERP, riducendo quindi la sua visibilità del 91,5%.

La SERP del soggetto in questione (di cui non sarà rivelato il nome né alcun riferimento alla sua identità per ovvie ragioni di *privacy*) è stata tenuta sotto osservazione per 8 settimane raccogliendo i dati necessari con il programma Scrappy. Nel frattempo sono state effettuate varie operazioni di aggiornamento e ottimizzazione di canali *social* e *blog* controllati dal soggetto.

I risultati ottenuti al termine del periodo di test sono più che positivi e sembrano confermare che l'applicazione della strategia abbia funzionato. Tuttavia, è da tenere presente che si tratta di un unico caso di studio, i cui dati sono stati raccolti per un periodo relativamente limitato. I risultati ottenuti non godono del confronto con casi simili e dunque anche i possibili fattori di successo non possono trovare conferma in altri casi di studio. Questo comporta che non è obiettivo di questo esperimento fornire una prova scientifica delle probabilità di questo metodo, ma un esempio di applicazione della strategia additiva per la tutela della reputazione che, a posteriori, sembra aver funzionato e che mostra una possibile correlazione tra operazioni effettuate e comportamento della SERP di Google.

6.2.2 Raccolta dati e anonimizzazione risultati

Durante le 8 settimane di osservazione, è stato lanciato giornalmente il programma Scrappy inserendo come *query* il nome e cognome del soggetto. I dati ottenuti sono poi stati normalizzati e processati attraverso Microsoft Excel (versione Mac 2011) per poter essere analizzati.

Scrappy ha raccolto ogni giorno i primi 50 risultati di ricerca della SERP sotto osservazione. Ai fini di questo esperimento, non è stata necessaria la rilevazione di occorrenze di *keyword* all'interno delle pagine *web* relative ai risultati di ricerca, dal momento che l'obiettivo era quello di colpire un risultato preciso di cui già si conosceva il contenuto e il contesto, mentre il *crawling* delle ricerche correlate non ha restituito alcun dato, in quanto la SERP in questione non restituiva ricerche correlate. *Rank*, titolo e URL sono stati i dati sufficienti per valutare i risultati degli interventi che sono stati applicati: l'obiettivo principale del monitoraggio è stato quello di studiare gli spostamenti dei risultati di ricerca nella prima pagina della SERP, ossia il movimento dei listati nel *range* delle prime 10 posizioni.

In rispetto della *privacy* del soggetto che si è prestato all'esperimento, i risultati sono qui presentati in forma anonima, senza riferimento alcuno alla sua identità né all'effettivo contenuto delle pagine, sia di natura positiva sia negativa, relative ai risultati di ricerca di cui è stato osservato il comportamento.

Sfortunatamente la necessità di mantenere l'anonimato penalizza la descrizione nel dettaglio delle procedure di correzione, ma non impedisce di esporne la natura e la strategia adottata, nonché discutere la possibile correlazione tra le operazioni eseguite e i risultati ottenuti.

6.2.3 L'identità digitale prima dell'intervento

In questo esperimento è stata presa in considerazione l'identità digitale rappresentata dalla prima pagina della SERP del nome e cognome del soggetto. Dal momento che si tratta della pagina visitata dal 91,5% degli utenti, la sua ottimizzazione (e, in questo caso, correzione) può rappresentare un successo già più che soddisfacente.

La prima fase è consistita nell'analisi dei primi 10 risultati di ricerca, i quali sono stati suddivisi in 4 + 1 categorie a cui è stato assegnato un colore:

1. positivi controllati (verde): risultati riferiti a pagine dal contenuto positivo per l'identità digitale che sono sotto il controllo del soggetto o di chi agisce per lui. Un esempio di risultato positivo controllato è il profilo LinkedIn personale, la pagina di un *blog* personale o per il quale il soggetto collabora.

2. positivi non controllati (blu): risultati riferiti a pagine dal contenuto positivo per l'identità digitale che non sono sotto il controllo del soggetto o di chi agisce per lui. Può trattarsi di un articolo sul soggetto scritto da terzi per esempio o un qualche tipo di pubblicazione prodotta dal soggetto ma ripubblicata da terzi.

3. non riferibili (giallo): risultati riferiti a pagine dal contenuto (positivo o neutro) non riferibili al soggetto, come casi di omonimia o altri casi in cui il nome del soggetto risulta associato a contesti che non lo riguardano

4. negativi (rosso): risultati riferiti a contenuti negativi o non desiderati, sono di norma non controllati dal soggetto (altrimenti sarebbero facilmente rimovibili), possono essere riferibili al soggetto come no.

Fuori categoria ci sono i risultati verticali (grigio), i quali contengono al loro interno più risultati che seguono la stessa categorizzazione dei risultati standard. Dal momento che l'obiettivo era colpire un preciso risultato che compariva tra i listati standard, la "striscia" di immagini che compariva nel risultato verticale non è stato tenuto sotto particolare osservazione, anche se si sono potuti osservare alcuni cambiamenti riconducibili anche questi alle attività svolte durante il periodo di correzione.

La Figura 6.1 mostra come appariva la SERP all'inizio della fase di osservazione, precisamente il 6 novembre 2016. I blocchi rappresentano i risultati di ricerca, colorati secondo la suddivisione in categorie sopra esposta. Di lato è definita la tipologia di contenuto (senza, per motivi di *privacy*, rivelare quale effettivamente sia).

La pagina è costituita da una considerevole maggioranza di listati verdi (7 su 10). Si tratta di profili *social* di tipo sostanzialmente professionale (LinkedIn, Google +, Accademia.edu) e pagine di due *blog* per cui il soggetto scrive. Blog 1 è una *webzine* non di proprietà del soggetto ma all'interno della quale è sufficientemente coinvolto da poterne controllare i contenuti (suoi) che vengono pubblicati. Blog 2 è invece gestito direttamente dal soggetto. Da notare come il profilo Google Plus (lo stesso) conti ben due listati: è probabile che Google favorisca la comparsa di risultati relativi al suo *social network* proprietario.

Il risultato negativo si trova in quarta posizione, si tratta di un articolo che pubblicizza un evento, ormai passato da diversi anni, al cui tema

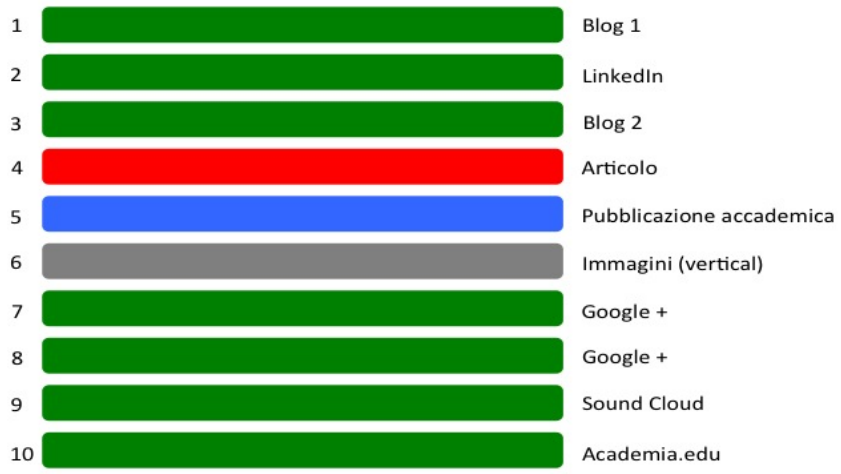

Figura 6.1: SERP del 6/11/2016

il soggetto non vuole più essere associato. Non si tratta di un contenuto che colpisce di proposito la reputazione del soggetto, ma semplicemente lo inserisce in un contesto nel quale egli non si riconosce.

Il listato blu rappresenta un elaborato accademico prodotto dal soggetto di cui non ha autorizzato la diffusione, ma comunque sostanzialmente positivo per la sua immagine pubblica.

La maggioranza di listati verdi delineano una identità digitale già complessivamente positiva, anche se la sola presenza di un risultato negativo o indesiderato può di fatto compromettere tutto l'elenco. La presenza di molti profili *social* era già un buon punto di partenza per poter iniziare il processo di occupazione della SERP con canali proprietari.

Mentre i *blog* erano aggiornati con regolarità e frequenza, buona parte dei profili *social* risultavano invece trascurati e poco utilizzati. Da notare l'assenza di un risultato che rimanda al profilo Facebook: il soggetto aveva già da tempo deindicizzato il proprio profilo attraverso le impostazioni della *privacy* fornite dalla piattaforma stessa.

Al momento non compaiono risultati non riferibili (gialli), anche se se ne vedrà la comparsa durante il periodo di osservazione della SERP.

6.2.4 Strategia di correzione

Cercare di correggere un'identità digitale attraverso la strategia additiva significa riuscire ad ottimizzare una serie di pagine *web* in modo che queste

compaiano nelle prime posizioni, spingendo i contenuti indesiderati verso il basso.

Se la prima pagina dei risultati conta 10 listati, idealmente occorrerebbero 10 contenuti positivi controllati per occuparla per intero, costringendo gli altri risultati, compreso quello negativo, a spostarsi nelle pagine successive. L'obiettivo ideale per un totale controllo della reputazione sarebbe ottenere almeno la prima pagina della SERP completamente "verde", ossia occupata interamente da risultati positivi controllati.

Durante un confronto con il responsabile di "Ingegneria Reputazionale" presso Reputation Manager[3], Renzo Mastromattei, è tuttavia emerso che un'identità digitale composta esclusivamente da pagine *web* del soggetto stesso può nascondere delle controindicazioni. Per ottenere un'identità autentica infatti, occorre non cadere nell'autoreferenzialità: la credibilità di una persona non deriva esclusivamente da ciò che essa dice di sé, ma anche da cosa gli altri dicono di lei.

Una buona identità digitale sul motore di ricerca si compone dunque di una ragionevole "miscela" di risultati verdi e blu, cioè positivi controllati insieme a positivi non controllati (o, ancora meglio, controllati da qualcuno non ricollegabile al soggetto ma con il quale egli ha contatto diretto).

Da "abbattere" insieme ai risultati negativi sono invece i risultati non riferibili: questi sono fuorvianti, creano ambiguità nella costituzione dell'identità digitale e potrebbero anche trasformarsi in risultati negativi. Al momento dell'inizio dell'operazione di correzione non erano presenti listati gialli nelle prime 10 posizioni, ma sono comparsi successivamente. Essendo non controllati come i listati rossi, non si può agire su di essi se non con gli stessi mezzi di addizione di contenuti controllati.

I criteri da utilizzare nel metodo additivo sono, sostanzialmente, quelli del *content marketing*, ossia portare traffico ai propri canali attraverso la pubblicazione di contenuti. In poche parole una strategia di *content marketing* mira ad attirare l'attenzione degli utenti (e di conseguenza dei motori di ricerca) producendo e condividendo risorse editoriali pertinenti con un certo tipo di prodotto. Nel caso della reputazione il prodotto è il nome stesso del soggetto (o dell'azienda, in caso si tratti di un'operazione di *brand reputation*).

[3]Importante agenzia di ORM italiana.

Il problema sta che un'azienda, così come un personaggio pubblico, può contare su un grande numero di canali proprietari su cui condividere e ottimizzare contenuti. In primo luogo, il sito *web*, che se ottimizzato a dovere lato SEO, può facilmente svettare in cima alla SERP. Il sito può inoltre essere affiancato da un *blog* aziendale, il quale può essere aggiornato anche quotidianamente, fornendo sia ad utenti sia ai *crawler* di Google materiale sempre nuovo. Per di più, è probabile che anche i canali *social* possano contare su un considerevole numero di fan o *followers* con cui impiantare una discussione online sulle proprie attività e prodotti, rispondere direttamente in caso di situazioni di crisi reputazionali e condividere contenuti.

Un privato non particolarmente famoso, può avere numericamente tutti gli stessi canali di un'azienda, ma molte meno occasioni per poter produrre e condividere contenuti conservando la sua credibilità: curare ed aggiornare la propria presenza *online* è bene, ma occorre fare attenzione che una iper-attività dei suoi canali digitali non getti un'ombra negativa sulla sua intera immagine digitale. È bene ricordare che è di reputazione che stiamo parlando: pubblicare contenuti non pertinenti o esageratamente "gonfiati" potrebbero dare l'impressione di egocentrismo, esibizionismo e superficialità (oltre che *web*-dipendenza).

Anche in questo caso, si tratta di agire secondo le condizioni specifiche del soggetto per cui si sta operando: egli non potrà rifiutarsi di incrementare la sua presenza sul Web, altrimenti non sarebbe applicabile alcuna operazione additiva, ma potrebbe voler conservare un'immagine moderata e professionale, un giusto equilibrio tra presenza online e riservatezza.

Nel nostro caso, il soggetto, già con una presenza *online* sufficiente e sostanzialmente positiva, ha chiesto di fare giusto il necessario per colpire il risultato negativo, mantenendo il più possibile un "basso profilo" della sua immagine digitale. Di seguito sono esposti i tipi di intervento che sono stati effettuati.

6.2.4.1 Sito *web* e *blog*

Ale Agostini, esperto in SEO e *brand reputation*, sostiene che la prima cosa da fare per stabilire quello che lui chiama «presidio del territorio»[4] è creare un sito *web* acquistando un dominio di primo livello contenente il nome e

[4] Agostini, *La tua reputazione su Google e i Social Media. Prevenire, monitorare, curare*, cit.

cognome del soggetto. La presenza della *keyword* in un dominio infatti, è una delle prime cose che Google considera per determinare la rilevanza di una pagina *web* in base ad una determinata *query*.

Inoltre, essendo un sito *web* totalmente sotto il controllo del suo proprietario (il quale può dunque agire direttamente sul codice delle pagine *web*, a differenza di piattaforme di terze parti come i *social* o siti e *blog* di cui non è il proprietario), si presta all'ottimizzazione SEO *on page*, vale a dire la compilazione di tutti i metadati necessari affinché Google le indicizzi al meglio secondo le direttive dell'autore.

Nasce però un problema di fondo nell'applicabilità di questa strategia: perché una persona abbia un sito *web* che porta il suo nome, deve avere davvero cose molto interessanti da dire, se vuole migliorare la sua reputazione e non peggiorarla drasticamente. La pertinenza dei contenuti è il primo requisito della loro qualità. Creare un sito "tanto per" avrà molto probabilmente degli effetti negativi sulla sua immagine pubblica e non certo migliorativi.

Inoltre, creare un *blog* significa che il soggetto deve essere disposto a curarlo nel tempo: deve egli stesso creare contenuti (o per lo meno firmare quelli prodotti da terzi come se fossero suoi) e pubblicarli con una discreta frequenza. Anche in questo caso, un sito trascurato e poco aggiornato non gioverà alla sua immagine, oltre che a perdere nel tempo la sua funzione di occupare i primi posti della SERP, dal momento che Google tende a premiare pagine web aggiornate e contenuti nuovi, penalizzando di conseguenza contenuti "vecchi".

Le ragioni per le quali creare un sito che porta il nome di una persona devono trascendere le necessità di riparare la sua reputazione. In caso contrario, è preferibile accantonare l'idea per non incorrere in ulteriori problemi legati alla sua immagine *web*.

Nel nostro caso, il soggetto già scrive per due *blog*. Questi non portano il suo nome e non sono di sua proprietà, ma è lui a stabilire quali contenuti pubblicare (quelli firmati da lui, i *blog* prevedono anche altri autori). Cosa più importante, entrambi i *blog* sono su piattaforma Wordpress. Wordpress prevede di *default* una pagina dedicata all'autore, la quale raccoglie tutti i suoi articoli e ospita la sua biografia con foto, per una certa misura personalizzabile dall'autore stesso. Questa corrisponde ad una vera e propria pagina personale, nella quale tag rilevanti per l'indicizzazione quali `h1`, `title` e `meta description` contengono, o possono contenere, il nome e cognome dell'au-

tore. Non a caso, i due listati della SERP segnalati come Blog 1 e Blog 2, corrispondono proprio ai *link* alle *author page* del soggetto sui rispettivi *blog*.

Durante l'ispezione nel codice delle pagine autore è stato verificato che nome e cognome del soggetto comparivano già nel *tag* `title` della `head` del documento HTML e nel `h1` del `body`. Non è stato invece rintracciato alcun `meta description`: lo *snippet* del risultato di Google infatti mostra un frammento dell'ultimo articolo pubblicato dal *blog*, ma non dell'autore. La biografia dell'autore, che sarebbe stata il contenuto perfetto per lo *snippet* di Google, è purtroppo all'interno di un paragrafo `p` nel `body` del documento ed è completamente ignorato da Google.

Il nome era presente anche in altri `meta` *tag* della `head` e altri elementi del `body`, tra i quali un collegamento ipertestuale `a`, il quale compare anche in tutte le pagine articolo dello stesso autore. Il collegamento rimanda alla stessa *author page* dell'autore e contiene un attributo *rel=author*: si tratta di un attributo HTML5 che indica l'*authorship* di una pagina *web*. In altre parole, si tratta di un *tag* che crea una relazione tra un documento e una pagina *web* relativa ad un autore.

Il W3c definisce così questo attributo:

> The author keyword may be used with link, a, and area elements. This keyword creates a hyperlink.
>
> For a and area elements, the author keyword indicates that the referenced document provides further information about the author of the nearest article element ancestor of the element defining the hyperlink, if there is one, or of the page as a whole, otherwise.
>
> For link elements, the author keyword indicates that the referenced document provides further information about the author for the page as a whole.[5]

Non è chiaro quanta sia l'effettiva influenza che questo attributo eserciti sull'indicizzazione delle pagine nel 2017, dal momento che, per quanto se ne sa a livello ufficiale, non è più supportato da Google, almeno non più come avveniva in passato.

Fra il 2012 e il 2014, Google ha fatto un uso molto interessante dell'attributo `author`: se, su una pagina *web*, veniva aggiunto un collegamento ipertestuale che conteneva l'attributo `author` e puntava al profilo Google

[5] W3C, *HTML 5.1 - Links*, 1 novembre 2016, https://www.w3.org/TR/html/links.html, 19/4/2017.

Plus dell'autore della pagina stessa, Google attribuiva "ufficialmente" la Authorship di quella pagina al suo autore. Nella pratica, questo si traduceva Google, oltre a tenere un indice dei contenuti, aveva anche un indice di autori, ai quali attribuiva quei contenuti.

Il beneficio più evidente era che i contenuti con Authorship certificata comparivano con la foto a fianco del loro autore, ma, almeno secondo le dichiarazioni di Google, costituiva anche un fattore di *ranking* rilevante. Il fatto che Google fosse in grado di attribuire un autore a un contenuto nel momento in cui lo indicizzava, lo metteva inoltre nella condizione di riconoscere la paternità di quel contenuto anche come frammento copiato o plagiato altrove. Si trattava di uno strumento interno al motore di ricerca sul controllo del *copyright*.

Sfortunatamente, nel 2014 Google dichiarò la fine di Authorship, in quanto, stando alle dichiarazione del portavoce John Mueller, si trattava di un'informazione non poi così utile per gli utenti[6]. La vera ragione era, almeno secondo uno studio[7] di Eric Enge, che l'attributo era stato mal utilizzato dai gestori dei siti web.

Google Authorship sarebbe stato un ottimo strumento per costruire e controllare l'identità digitale, ma purtroppo non esiste più e non c'è altro che possa sostituirlo. *The Art of SEO* in ogni caso fornisce una serie di consigli per poter coltivare, anche senza l'aiuto di Google, la «Author Authority», ecco quelli più interessanti:

- pubblicare col vero nome: per costruire un'*author authority* i motori di i ricerca devono poter riconoscere che più contenuti sono ricollegabili ad un solo individuo. Usare il proprio nome è, ovviamente, il primo modo per farsi riconoscere.

- *cross linking* delle pagine autore se il soggetto scrive per più siti: si tratta di un altro modo per presentarlo come unica entità

- collegare profili social ai contenuti: altra strategia per per creare un piccolo *network* personale intorno alla propria identità digitale. Anche se Google Authorship non esiste più, un *link* al profilo Google Plus è d'obbligo.

[6]Enge, *The art of SEO*, cit., p. 407.

[7]E. Enge, *Authorship Adoption Fail - Detailed Stats*, Stone temple Consulting, 9 settembre 2014, https://www.stonetemple.com/authorship-adoption-fail-detailed-stats/, 19/4/2017.

6.2.4.2 Profili *social*

Se da un lato non era molto pertinente creare un sito `nome-cognome.it`, l'attività *web* di un utente non famoso poteva essere abbastanza spesa sui *social network*.

Al momento dell'inizio della fase di monitoraggio, il soggetto aveva i seguenti profili:

- Facebook
- Google Plus
- LinkedIn
- Sound Cloud
- Accademia.edu
- Twitter

A dire il vero, il profilo Accademia.edu era stato creato appena due giorni prima dell'inizio dell'esperimento e, come mostra la Figura 6.1, già era entrata nella top 10 dei risultati.

Il profilo Facebook, deindicizzato dall'utente, è stato escluso dal programma di correzione sotto richiesta del soggetto, quindi non è stata riabilitata la sua indicizzazione attraverso le impostazioni della *privacy*.

La maggior parte degli altri profili era sì esistente, ma quasi o del tutto inutilizzata. La strategia adottata è stata non tanto di creare nuovi profili ma di quella di intensificare l'attività su quelli già esistenti, per osservare se questi si sarebbero "mossi" nella SERP. Da notare come il profilo Twitter non risultasse ancora indicizzato, o almeno non presente nelle prime 50 posizioni della SERP (vedremo come comparirà nell'ultima fase del processo), che è il massimo di risultati registrati con il programma Scrappy.

Purtroppo, i *social network* si prestano molto poco ad operazioni di SEO, sia che si tratti di SEO *on page* dei profili stessi sia che questi vengano sfruttati per posizionare risorse esterne (SEO *off page*) sfruttando la loro grande potenzialità di produrre *link* verso l'esterno.

I *social* infatti sono piattaforme chiuse e di terze parti: non è possibile accedere al loro codice e ottimizzare l'HTML in ottica SEO. Le uniche operazioni di indicizzazione possibili sono quelle che alcuni di essi, come per

esempio Facebook, mettono a disposizione nelle impostazioni della *privacy*, in cui si può stabilire il grado di visibilità dei propri contenuti sia all'interno del *social network* sia all'esterno.

Per quanto riguarda invece la possibilità di sfruttarli nella *link building*, cioè creare collegamenti ipertestuali pagine web esterne (come per esempio gli articoli dei blog, o altre pagine positive per la reputazione sia controllate che non) al fine di aumentarne l'*authority*, i *social network* sono, tecnicamente, inutili: quasi tutti i *social* infatti, aggiungono automaticamente l'attributo `rel="nofollow"` agli elementi `a` in uscita. Questo significa che qualsiasi indirizzo web pubblicato su una piattaforma *social* non riceverà alcun beneficio da questo *link* in ingresso, dal momento che il *crawler* del motore di ricerca non seguirà quel *link*.

Restano tuttavia una serie di studi[8] che mostrano una correlazione, evidente ma non sufficiente a costituire una prova scientifica, tra i cosiddetti *social signals*, ossia vari eventi tipici dei *social media* come *like*, condivisioni, *retweet* ecc., e posizionamento di contenuti esterni sui motori di ricerca, così come posizionamento di *link* degli stessi profili o pagine *social*.

6.2.4.3 Selezione dei canali controllati

I canali controllati coinvolti nella strategia di correzione sono stati complessivamente 6:

- Blog 1
- Blog 2
- profilo LinkedIn
- profilo Google Plus
- profilo Academia.edu
- profilo Twitter

Sono rimasti fuori dalle operazioni Facebook e Sound Cloud: il primo per scelta di *privacy* da parte del soggetto, il secondo per impossibilità, in quel momento, di poter fornire contenuti pertinenti e soprattutto di carattere

[8] Jacopo Matteuzzi, *Social signals e SEO: i risultati di cinque studi*, "Studio Samo", 20 gennaio 2014, https://www.studiosamo.it/seo/social-signals-e-seo-risultati-di-cinque-studi/, 19/4/2017.

professionale. L'obiettivo generale infatti è stato quello di agire per addizione di contenuti che contribuissero il più possibile a creare un'immagine professionale del soggetto, indipendentemente dalle necessità di correzione. Facebook e Sound Cloud sono state scartate perché facenti parte della sfera personale.

I canali scelti sono inferiori al numero ideale, cioè 10 (10 canali per le prime 10 posizioni della SERP), per una serie di ragioni. La prima, ancora una volta, è la pertinenza dei contenuti: è stata scartata, per esempio, la possibilità di creare profili *social* dedicati alla condivisione di contenuti multimediali quali foto e video (come Istagram, Youtube, Flikr) per il semplice fatto che il soggetto non avrebbe potuto fornire una quantità sufficiente di contenuti pertinenti e di qualità.

In secondo luogo, è bene ricordare che la SERP ideale per una reputazione *online* credibile è composta da un giusto equilibrio di listati positivi controllati e positivi non controllati. Contando che un listato sarà sicuramente occupato da Google Image e che alcuni canali tendono a occupare due posizioni (è il caso di Google Plus, Academia.edu, Blog 1 ecc.) è bene non esagerare con i canali attivi onde evitare la sovraottimizzazione dell'identità digitale (anche se l'esito dell'esperimento sarà proprio una SERP sovraottimizzata, cioè completamente verde).

6.2.4.4 Ottimizzazione SEO *on page* e *off page* dei canali controllati

Una volta stabiliti su quali canali organizzare la strategia di correzione per addizione, la prima operazione è stata la loro ottimizzazione lato SEO.

Una campagna SEO si svolge solitamente in due fasi: scelta delle parole chiave su cui posizionarsi e ottimizzazione delle pagine *web* per le parole chiave selezionate.

Nella SEO finalizzata alla correzione reputazionale, la prima fase è pressoché inesistente: le parole chiave su cui posizionarsi sono quelle in cui compare il risultato che si vuole colpire.

L'ottimizzazione di una pagina web si svolge invece su due piani:

- ***on page***: attività di ottimizzazione applicate all'interno della pagina *web*. Si suddivide a sua volta in due aree distinte: l'ottimizzazione del codice HTML e della struttura del sito (nel caso si voglia posizionare

un intero sito e non una pagina specifica) e l'ottimizzazione dei contenuti testuali e multimediali (prevalentemente immagini) all'interno delle pagine.

- ***off page***: la gestione dei *link* presenti su altri siti che puntano alla pagina web da ottimizzare, al fine di aumentarne il PageRank[9], oltre che visibilità.

Entrambe le forme di ottimizzazione hanno presentato, nell'applicazione ai 6 canali controllati, alcune criticità.

Ottimizzazione *on page* Per quanto riguarda la prima area dell'ottimizzazione *on page*, cioè quella circoscritta all'ottimizzazione dell'HTML, le operazioni applicabili possono essere riassunte nei seguenti punti fondamentali:

1. **URL ottimizzati**: l'URL della pagina web deve contenere la parola chiave sulla quale si vuole posizionare la risorsa. Nel nostro caso si tratta ovviamente del nome e cognome del soggetto. Le parole devono essere separate da trattino "-" e non da altri segni di separazione come "_" e assolutamente non "/". Gli URL devono essere statici e non dinamici, un problema che si può riscontrare con i CMS che mostrano la chiamata PHP al *server*. L'URL ideale sarà `www.sito.it/nome-cognome/` e non `www.sito.it/nomecognome/`, `www.sito.it/nome_cognome/` o, ancora peggio, `www.sito.it/php?post=8907&action=view/`.

2. **Tag `title`**: l'importanza di questo tag è già stato discusso più volte in precedenza. Il tag `title` è infatti il testo cliccabile che appare nei risultati di ricerca, oltre ad essere uno dei fattori di *ranking* più importanti.

3. `meta description`: altro "classico" della SEO di base, non è in realtà un fattore di *ranking* particolarmente importante, ma dal momento che compare nello *snippet* risultati di ricerca, la sua cura è fondamentale.

[9]Celebre algoritmo usato da Google per attribuire un punteggio di autorità alle pagine *web*. Questo punteggio si basa sulla quantità e qualità dei *link* in entrata di un sito *web*. Il PageRank va da 0 (bassa autorità) a 10 (massima autorità). Solo siti come Google, Twitter o Facebook raggiungono un punteggio alto (lo stesso Google non riesce a superare un PR 9). Siti con un PR fra i 3 e 6 sono da considerarsi già adeguatamente autorevoli.

4. **Tag Heading**: tag quali `h1`, `h2`, `h3`, ecc., ossia gli elementi HTML che servono a definire il titolo principale e i vari sottotitoli delle pagine web. Anche in questi elementi, soprattutto quelli di livello più alto (`h1` e `h2`) è consigliabile inserire le parole chiave, purché sia logico e naturale farlo: le forzature giocheranno sempre e solo a sfavore.

5. **immagini ottimizzate**: per quanto Google stia diventando sempre più sofisticato nell'analisi e riconoscimento delle immagini, apprezza ancora di buon grado che sia l'autore a suggerigli cosa un'immagine rappresenti. Questo avviene in due modi: compilando l'attributo `alt` degli elementi `img` e assegnando al file immagine stesso un nome significativo, come `mario-rossi.jpg` anziché sigle prive di senso come `IMG00032.JPG` (molto tipico di file nominati in automatico dalle videocamere digitali). La separazione delle parole composte segue gli stessi criteri dell'ottimizzazione degli URL.

L'ottimizzazione dell'HTML di pagine *web* statiche è cosa piuttosto semplice, a patto di poter avere accesso ed avere sufficiente conoscenza dell'HTML. Lo stesso non si può dire per pagine generate dinamicamente come quelle dei CMS per il fatto che non è possibile modificare direttamente il codice HTML. Sfortunatamente, tutti e sei i canali controllati erano CMS.

Tuttavia, sia WordPress, con il quale sono realizzati i due *blog*, sia le piattaforme *social* consentono, tramite apposite interfacce utente, di poter modificare alcune impostazioni che vanno a influenzare proprio quegli elementi HTML coinvolti nell'ottimizzazione *on page*.

Per quanto riguarda punto 1, cioè l'ottimizzazione degli URL, le *author page* dei due *blog* presentavano già una configurazione ottimale, cioè `www.blog.it/author/nome-cognome`. I profili *social* impiegati invece, consentono nella sezione impostazioni, di assegnare un *permalink*[10] alla propria pagina profilo. L'unico canale per cui non è stato possibile modificare l'URL è stato Google Plus. Per personalizzare l'URL di un profilo personale infatti, Google Plus pretende che il profilo abbia almeno 10 *followers*. Per quanto possa sembrare strano, Google Plus è talmente poco usato che in due mesi di tempo non è stato possibile raggiungere questa cifra, anche se davvero minima, di *followers* il profilo del soggetto. Gli URL di LinkedIn, Twitter e Academia.edu sono stati invece personalizzati con successo.

[10]URL statico assegnato a una pagina dinamica.

Il *tag* `title` (punto 2) non è modificabile su nessuna di queste piattaforme, ma risulta in realtà già ottimizzato in automatico. Tutti i *tag* `title` dei sei canali contenevano già il nome e il cognome del soggetto. Stessa cosa vale per tag `h1`, ma purtroppo non per `meta description`, i quali purtroppo non sono presenti su queste pagine e non c'è modi di aggiungerli.

Alle foto dei profili, compresi quelle delle biografie dei *blog*, è stato assegnato un nome *file* significativo, `come nome-cognome.jpg`. Purtroppo non è stato possibile impostare l'attributo `alt`, tuttavia, dopo un'ispezione del codice è stato verificato che i CMS avevano già inserito nome e cognome all'interno di questo attributo.

Ottimizzazione *off page* Purtroppo, si è trattato del tipo di operazione più critico da concretizzare e le ragioni del successo dell'esperimento non sono certo imputabili all'ottimizzazione *off page* dei contenuti controllati.

Questa infatti consiste, di base, nel fare in modo che altri siti ospitino dei collegamenti ipertestuali che puntano le pagine *web* che si vogliono ottimizzare. Questo può avvenire in modo spontaneo (cioè utenti che spontaneamente linkano le nostre pagine *web* su altri siti, *blog* e *forum*) oppure prendendo accordi con i gestori degli altri siti, che in cambio di denaro o altri tipi di risorse possono essere disposti a ospitare un *link*.

Con la diffusione dei *social network*, i *link* spontanei sono diventati una rarità. Gli utenti infatti sono sempre più soliti postare contenuti che reputano interessanti sulle piattaforme *social*, i quali *link* in uscita sono del tutto inutili ad aumentare il PageRank delle pagine *web*, e sempre meno da *blog* e *forum*.

Le campagne di *link building* sono dunque praticabili con facilità nelle agenzie di comunicazione o da parte di personaggi rilevanti, ma molto più complesse e meno efficaci a livello privato, per semplice mancanza di contatti e polarità sufficienti per poter intrecciare collaborazioni con altri siti *web* che siano, come sempre, pertinenti e non forzate.

6.2.4.5 Incremento delle attività

La strategia generale adottata è stata quella di incrementare l'attività sui *social*, promuovendo quanto più possibile quello che veniva pubblicato sui due *blog*, i quali, per ragioni editoriali interne, non potevano essere aggiornati più di quanto già si stava facendo.

I canali che sono stati ritenuti più adeguati per la pubblicazione dei contenuti sono stati Google Plus, LinkedIn e Twitter e in alcuni casi Academia.edu. Ottenendo un totale, insieme ai due blog, di sei canali attivi.

Come già discusso in merito alla possibilità di creare un sito personale, la pertinenza e l'adeguatezza dei contenuti prevale sulle necessità di incremento delle pubblicazioni. Il metodo additivo gioca sulla quantità, ma non deve in alcun modo prescindere dalla qualità. Sono stati pubblicati i contenuti solo se ritenuti all'altezza dell'immagine che il soggetto voleva crearsi, anche se a volte questo compromesso la regolarità della frequenza delle pubblicazioni e non ha reso sempre possibile agire in modo sistematico. Resta il fatto che, visto il successo dell'esito dell'operazione, l'intensità con cui sono stati aggiornati i vari canali è stata sufficiente per raggiungere l'obiettivo preposto, con il beneficio aggiunto che tutto ciò che è stato fatto appare dall'esterno come attività naturale, senza lasciar trasparire alcun tipo di "forzatura", perché ogni azione è stata applicata solo se ritenuta pertinente e in linea con le abitudini e la personalità del soggetto.

Nella Figura 6.2 è rappresentato l'incremento di attività sui sei canali controllati, in termini di frequenza giornaliera e intensità, a partire dal mese di novembre, cioè dal momento in cui è iniziata la fase di monitoraggio, rispetto ai due mesi precedenti. Ogni `1` rappresentato nella tabella sottostante al grafico a colonne rappresenta un'azione compiuta sul canale corrispondente: per azione è stata considerata la pubblicazione di un contenuto o un aggiornamento abbastanza consistente del profilo (come per esempio l'aggiornamento dell'immagine del profilo di un canale *social* o l'aggiunta della biografia sulla *author page* del *blog*).

6.2.5 L'identità digitale dopo l'intervento

Dopo 8 settimane di attività sui canali controllati, il contenuto indesiderato risultava in dodicesima posizione, cioè nella seconda pagina della SERP: l'obiettivo era stato raggiunto. Il 7 gennaio 2017, due mesi dopo la SERP registrata in Figura 6.1, risultava come in Figura 6.3, ossia interamente occupata da listati positivi controllati (ad eccezione, ovviamente, del risultato verticale relativo a immagini, il quale presentava comunque nelle prime posizioni immagini provenienti dai canali controllati).

L'identità digitale risultava "ripulita", almeno nella prima pagina, da risultati indesiderati e addirittura da qualsiasi altro contenuto non controllato

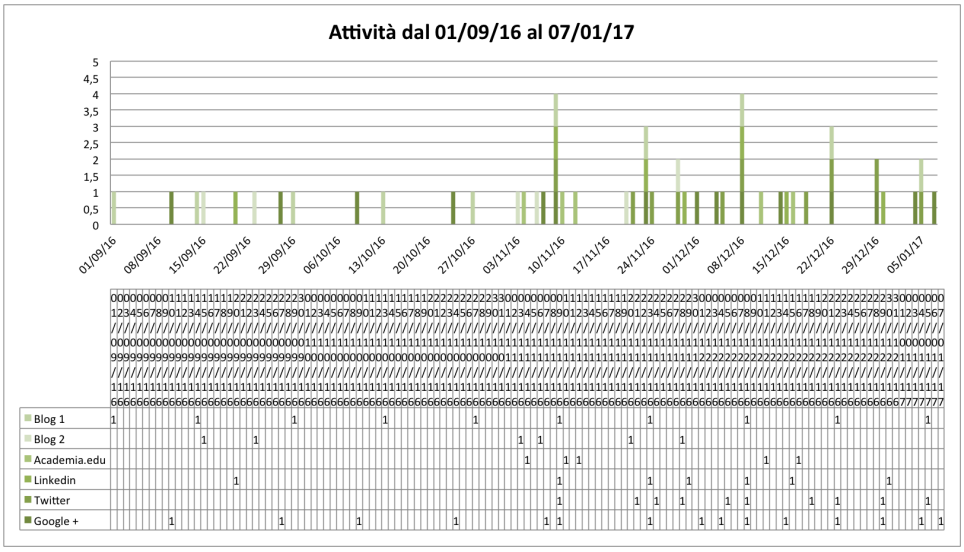

Figura 6.2: Frequenza e intensità dell'attività sui canali controllati da settembre a gennaio

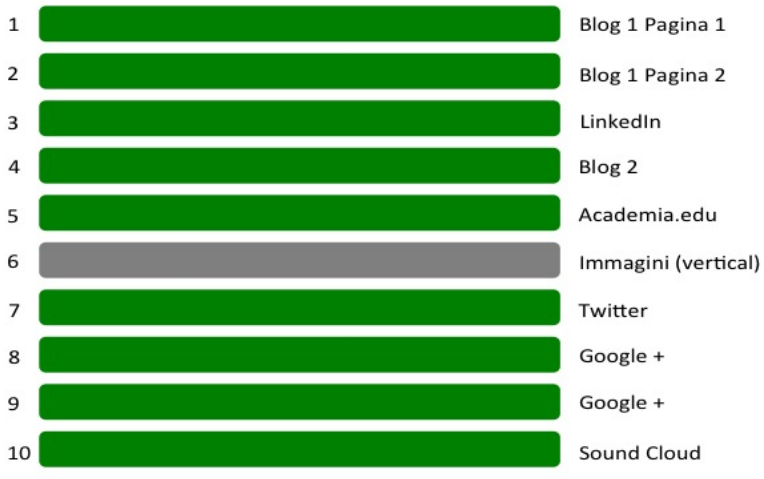

Figura 6.3: SERP del 7/11/2017

o non riferibile. La SERP era addirittura sovraottimizzata: non erano presenti infatti risultati positivi non controllati, col rischio che l'immagine personale risultasse eccessivamente autoreferenziale. Il profilo generale restituito dalla SERP aveva inoltre assunto un aspetto più professionale, in quanto buona parte dei risultati erano *link* ai canali che rappresentano la sfera professionale e accademica del soggetto, riducendo quella personale soltanto al profilo Sound Cloud il quale, per quanto non coinvolto nelle operazioni di correzione, è rimasto in prima pagina ma nella posizione più bassa.

La Figura 6.4 rappresenta tutti gli spostamenti fatti dai risultati che sono comparsi nelle prime 10 posizioni della SERP nell'arco del periodo di monitoraggio. Purtroppo il grafico non è di semplicissima interpretazione, dal momento che nelle 8 settimane sono ben 20 i *link* che sono comparsi nella prima pagina e, per quanto siano ognuno distinto da un colore diverso, non è sempre facile distinguerli.

Le categorie di risultato sono distinte per colore: tutte le tonalità di verde sono per i contenuti positivi controllati, blu per i positivi non controllati, giallo per i contenuti non riferibili mentre il rosso indica il risultato che si è cercato di rimuovere. Alcuni risultati sono rappresentati con più colori (per esempio Sound Cloud compare in tre colorazioni, Google +, Academia.edu e Blog1 in due) perché sono comparsi con URL differenti che rimandavano comunque al medesimo profilo. Sono rappresentati in colorazioni distinte perché alcuni di essi sono, per un certo periodo, stati co-presenti sulla SERP.

Nel grafico salta subito all'occhio come la SERP, dopo un periodo di stabilità e immobilità di circa 10 giorni, subisca a partire dal 18 novembre variazioni anche piuttosto drastiche, di cui la più evidente è proprio la posizione del listato negativo, che il 22 novembre sparisce improvvisamente dalla prima pagina della SERP, per ricomparire pochi giorni dopo faticando a riguadagnare posizione, fino a sparire definitivamente circa un mese dopo.

Da notare anche il totale "immobilismo" della prima pagina autore del Blog 1 che rimane indisturbata in prima posizione per tutta la fase di monitoraggio, così come la comparsa a intermittenza di contenuti non riferibili che sono soliti durare pochi giorni per poi scomparire e ricomparire dopo qualche tempo.

Il fatto che, una volta "abbattuto" il listato negativo, le prime 4 posizioni della SERP sono state sempre occupate da risultati positivi controllati, è da considerarsi molto positivo.

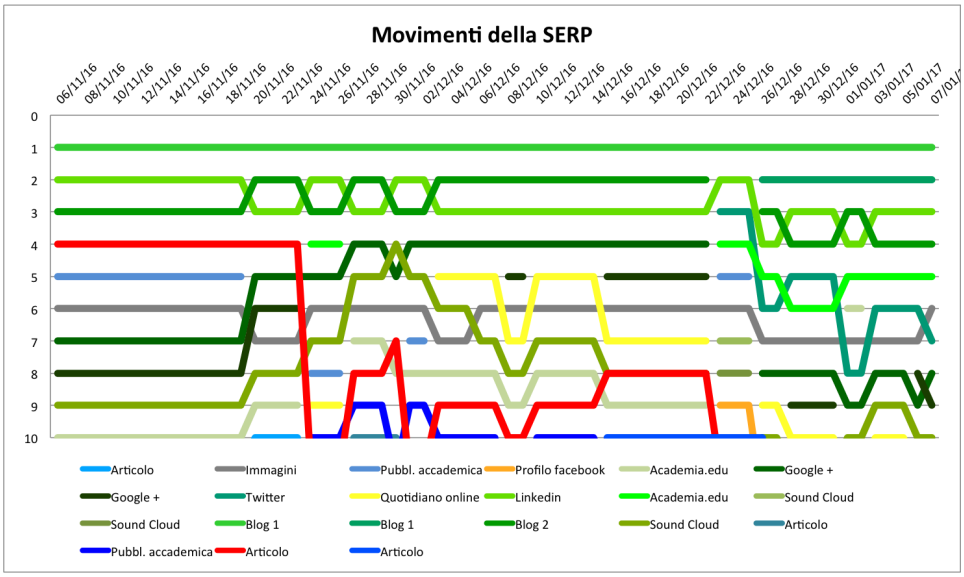

Figura 6.4: Evoluzione della SERP durante il periodo di monitoraggio. Gli elementi di tonalità verde sono i risultati positivi controllati, di tonalità blu i positivi non controllati, di tonalità gialla i non riferibili e l'elemento rosso è il risultato non desiderato.

In Figura 6.4 è possibile osservare nel dettaglio quale sia stato il percorso del risultato negativo in relazione all'attività svolta sui canali controllati (la stessa rapprestata in Figura 6.2). Dopo un'improvvisa e drastica perdita di posizione, il risultato riesce a risalire senza però recuperare la posizione di partenza. Resta definitivamente fuori dai primi 10 risultati a partire dal 22 dicembre, oscillando successivamente tra la undicesima e tredicesima posizione.

Alcune perdite di pozione sembrano corrispondere a un'attività sui canali controllati più intensa del solito, come quella del 23 novembre, 8 dicembre, 22 dicembre e 5 gennaio. In particolare, sembra che il listato negativo si sposti il giorno successivo a quello in cui più canali controllati sono stati aggiornati.

Non è facile stabilire se questa corrispondenza grafica sia da considerarsi significativa o meno. La "decisione" di Google di restituire alcuni risultati più in alto di quanto lo fossero prima (facendo calare di conseguenza il risultati che già occupavano quelle posizioni) dovrebbe, a regola, avvenire almeno dopo il passaggio del *crawler* che "legge" e valuta gli aggiornamenti. Che il passaggio del *crawler* sia più di una volta avvenuto in tempi così rapidi non

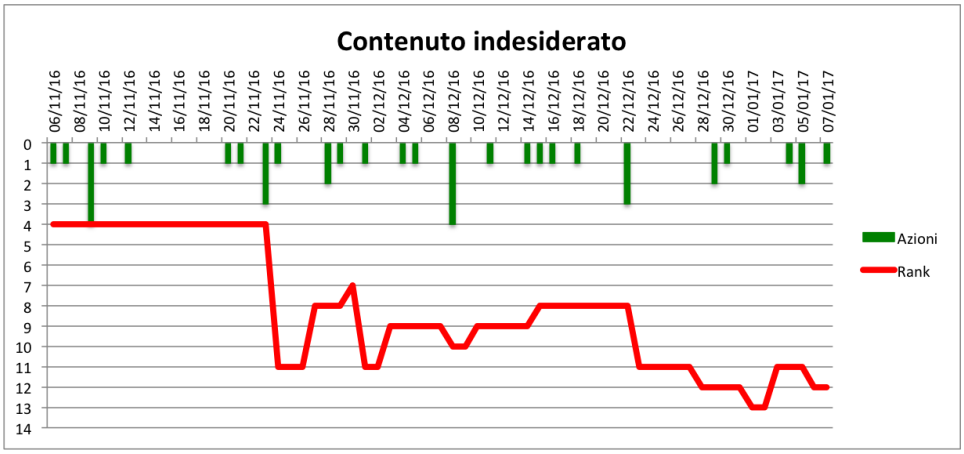

Figura 6.5: Spostamenti del listato negativo

sembra essere molto verosimile, anche se alcuni studi non ufficiali[11] hanno constatato che lo spostamento di alcuni risultati nella SERP sia avvenuto anche dopo pochi minuti il loro aggiornamento.

Il calcolo del coefficiente di correlazione tra intensità dell'attività (quanti canali controllati sono stati aggiornati in un giorno) e *rank* del listato negativo restituisce comunque un valore molto basso (0,07).

Le ragioni della perdita di *rank* del contenuto è da attribuirsi verosimilmente alla generale intensificazione nell'attività sui canali controllati contro la totale assenza di aggiornamenti del contenuto indesiderato. In altre parole, Google ha considerato più "fresche" le pagine aggiornate e le ha valorizzate a scapito di una pagina che non viene aggiornata da anni e probabilmente anche poco visitata. È da prendere invece con più cautela, senza però escluderla, l'ipotesi che uno o pochi giorni di intensa attività, corrisponda immediatamente uno spostamento verso il basso del listato negativo.

Comportamenti di altri listati degni di nota sono i due *blog*. Il Blog 1 (Figura 6.6) è stato aggiornato regolarmente ogni due settimane. La sua posizione, la prima, rimane invariata per tutte le otto settimane, fino a conquistare, nell'ultima fase, persino un secondo listato in seconda posizione.

Il Blog 2 (Figura 6.7) invece, ha interrotto le pubblicazioni a circa metà del periodo di monitoraggio. Questo sembra comportare una perdita di posizione e addirittura non compare tra le ricerche per un breve periodo di

[11]Si tratta di alcuni esperimenti condotti dall'agenzia di *web marketing* Sudio Samo e pubblicati, informalmente, sul suo gruppo di discussione Facebook "Da Zero a Seo"

Figura 6.6: Spostamenti del Blog 1

tempo.

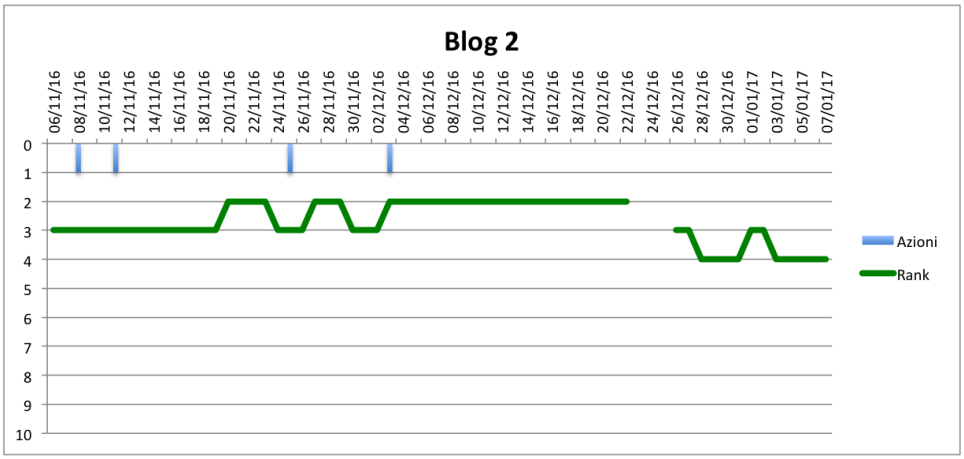

Figura 6.7: Spostamenti del Blog 2

Tra i fattori di resistenza in prima posizione del Blog 1 e di conquista della seconda con un altro *link*, non va però considerata solo la frequenza regolare e sistematica della pubblicazione di articoli. Rispetto al Blog 2, Blog 1 è, nel suo complesso, un sito *web* che esiste da più tempo, molto più grande e molto più visitato: questo significa che la sua *Authority* è decisamente più grande rispetto a Blog 2, più giovane e visitato da una piccola nicchia (si parla di 5000 visite al mese per Blog 1 contro le 20 visite per al mese di Blog 2). Il Page Rank del Blog 1 è infatti 2 (valore discreto) contro 0 del Blog 2 (valore minimo).

Resta il fatto che le pagine (sia Blog 1 e Blog 2) in cui il soggetto risulta essere "autore" di contenuti *web* (ricordiamo la presenza dell'attributo `rel="author"` all'interno delle pagine autore di Wordpress) si giochino a livello costante le prime posizioni della SERP corrispondente al nome e cognome del soggetto stesso, può far pensare che Google tenda a tenere in grande considerazioni i contenuti originali di cui la persona è autrice.

Per ultimo, è interessante osservare la comparsa improvvisa (Figura 6.8) del *link* al profilo Twitter. Questo profilo era stato creato nel mese di settembre dal soggetto, che però non lo aveva mai utilizzato. Questo profilo non risultava indicizzato.

Con l'inizio della fase di correzione, il soggetto ha iniziato a utilizzare il profilo pubblicando *tweet* con una certa regolarità. Dopo circa un mese dal primo *tweet*, il *link* al profilo è comparso improvvisamente in terza posizione, per poi calare nelle posizioni più basse ma sempre nel *range* delle prime 10.

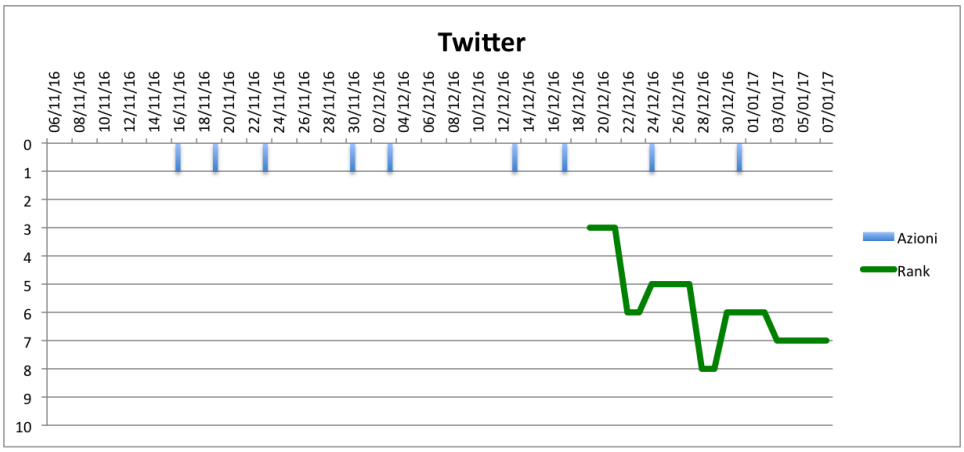

Figura 6.8: Spostamenti profilo Twitter

Le ragioni di spostamento di un listato all'interno della SERP possono essere molte e non tutte conosciute, dal momento che nessuno sa esattamente come funzionino gli algoritmi del motore di ricerca. Calcolare matematicamente un indice di correlazione tra attività su una pagina *web* e la sua rispettiva posizione all'interno della SERP potrebbe restituire dati poco affidabili, soprattutto su un periodo di tempo così breve e, soprattutto, considerando i dati di un solo caso di studio.

In primo luogo, perché tra aggiornamento di una pagina web e un eventuale cambiamento di posizione nei risultati restituiti dal motore di ricerca

c'è un evento necessario e non controllabile: il passaggio del *crawler*. Il passaggio può avvenire il giorno stesso, così come diversi mesi dopo e non c'è modo di prevederlo né tantomeno sapere con precisione quando è passato. Questo significa che se il listato relativo alla pagina *web* aggiornata si sposta nella SERP prima che il *crawler* sia passato, le ragioni di questo spostamento non sono riconducibili all'aggiornamento della pagina ma a qualche fattore esterno. Dal momento che non potremo mai sapere se il *crawler* è passato prima o dopo lo spostamento, stabilire con certezza le cause di questo evento è molto difficile.

Per esempio, c'è buona ragione di credere che il profilo Twitter non sia stato indicizzato da Google prima di novembre, per quanto fosse stato creato nel mese di settembre, perché fino a metà novembre questo non compare nemmeno nella SERP. Se era stato indicizzato prima, allora significa che fino al momento della sua comparsa in SERP non era stato ritenuto rilevante per la *keyword* nome + cognome del soggetto. Dopo alcuni *tweet* pubblicati, il *link* del profilo compare improvvisamente in terza posizione: è facile pensare che l'incremento di attività del profilo abbia "attirato l'attenzione" di Google che l'ha indicizzato e ritenuto molto rilevante per la *keyword*. Tuttavia non si può sapere se il profilo sarebbe stato comunque indicizzato anche in assenza di quell'attività e che i mesi trascorsi tra la creazione del profilo e la sua indicizzazione e conseguente comparsa in SERP non corrisponda semplicemente ad una naturale mancanza di passaggio del *crawler*.

Una delle prime cause esterne (cioè non dipendenti dall'attività su una pagina *web*) di spostamento di un listato è banale, ma è un fattore non trascurabile che rischia di confondere spesso le acque: lo spostamento degli altri listati. Se si guarda la Figura 6.2, si può notare come spesso gli spostamenti dei risultati avvengano in maniera sincronizzata. Questo però non significa che improvvisamente l'attività su ciascun canale abbia avuto un effetto sui rispettivi risultati in SERP: semplicemente, soltanto uno di quei listati si è spostato per un fattore interno, tutti gli altri si sono spostati di conseguenza (come è ovvio, se un listato sale di posizione tutti quelli sotto di lui lo seguiranno per ricoprire il posto che ha lasciato), solo che non possiamo sapere quale sia fra tutti quelli che si sono mossi.

Tuttavia, nel loro comportamento d'insieme, è abbastanza evidente come l'intensificazione dell'attività dei canali controllati, insieme alla loro, per quanto possibile, ottimizzazione lato SEO, abbia comportato uno sposta-

mento verso posizioni più alte della SERP o un mantenimento delle stesse, ai danni del risultato negativo che, già dopo pochi giorni dai primi interventi di ottimizzazione, ha subito una perdita di "quota" che non sarà più in grado di recuperare.

Si tratta di un risultato significativo per questo tipo di esperimento e per la conferma dell'efficacia della strategia additiva, ma anche, molto probabilmente, di un caso fortunato.

Il contenuto indesiderato infatti risulta trovarsi su una pagina per nulla ottimizzata, mancante di tutti gli elementi e attributi per una SEO di base (nemmeno tag `title` e `meta description` e con un PageRank pari a 0. A parte che nel punto esatto (soltanto uno) in cui compare il nome del soggetto, questo non è presente in nessun elemento HTML rilevante per l'indicizzazione. Inoltre, la sua pubblicazione risale a sei anni prima e, trattandosi della pubblicizzazione di un evento, molto difficilmente sarà stata visitata né tantomeno aggiornata negli anni successivi.

Viene quasi da chiedersi perché Google le abbia dato una rilevanza tale da posizionarla al quarto posto della SERP, ma questo sta a dimostrare che ciò che davvero possiamo sapere e prevedere dei criteri di *ranking* di Google è solo una vaga idea.

Un successo così rapido e drastico dell'intervento di correzione è imputabile con buona probabilità alla scarsa "resistenza" del contenuto negativo di fronte all'aggiornamento di pagine molto più ottimizzate e rilevanti per quella parola chiave. Probabilmente, se la notizia fosse stata riportata da una testata giornalistica autorevole, il risultato di ricerca relativo sarebbe stato di gran lunga più difficile da abbattere, almeno con i mezzi, in definitiva abbastanza scarsi, che sono stati impiegati in questo esperimento.

6.2.6 Criticità della strategia additiva

Il concetto alla base della strategia additiva può sembrare banale: scrivere qualche articolo, curare un po' di profili *social* e sperare che questi compaiano nelle posizioni più alte della SERP.

All'atto pratico invece, un lavoro organizzato di produzione e pubblicazione di contenuti per migliorare la reputazione di una persona ha presentato una serie di difficoltà. La maggior parte di queste, legate alla pertinenza e qualità dei contenuti stessi: questi devono rappresentare al meglio l'immagine della persona e allo stesso tempo non risultare autocelebrativi. La strategia

additiva, se condotta su una persona "comune" e non pubblica, si gioca su un difficile equilibrio tra pertinenza e quantità di contenti che mira a fare di questa persona un *brand*, ma senza che nessuno se ne accorga.

Certamente, attività di rilevanza pubblica del soggetto, come pubblicazioni, partecipazione a eventi e svolgimento di professioni qualificate possono aiutare moltissimo a costruire la sua identità digitale. In altre parole, più una persona ha "qualcosa da dire" (purché sia pertinente) più la strategia additiva sarà facile da applicare e non soltanto sarà più efficace in termini di correzione, ma risulterà più naturale in tutto il processo di messa in pratica e nei suoi esiti finali.

Altro fattore da non sottovalutare è la partecipazione del soggetto. Tutto ciò che viene pubblicato costruisce la sua identità, quindi deve essere un suo prodotto o deve potersi riconoscere a pieno. Tutte le attività sui canali controllati devono essere condotte con frequenza e costanza per un periodo di tempo che può anche essere lungo e in tutta la fase di correzione e costruzione dell'identità digitale il soggetto deve essere presente e partecipe.

Prima di portare a termine l'esperimento riportato in questo capitolo, sono stati più di uno i tentativi di applicazione della strategia ad altri casi, ma le operazioni si sono interrotte, o non sono iniziate proprio, proprio perché i soggetti interessati non hanno potuto, o non hanno voluto per "avversione" al mondo digitale, garantire assidua partecipazione al progetto.

Conclusioni

Rimuovere o ridurre di visibilità di informazioni personali sul motore di ricerca si espone ad una critica per nulla trascurabile: si potrebbe dire (e parlando delle mie ricerche con amici e conoscenti, qualcuno l'ha detto davvero) che "annebbiare" il passato di una persona non va a favore soltanto di chi è ingiustamente vittima del proprio passato, ma offre la possibilità anche a chi non se lo merita di essere "perdonato" dalla Rete.

Tutto questo avviene per di più sfruttando, almeno per quanto riguarda la strategia additiva, nient'altro che trucchetti per influenzare a proprio favore uno strumento di pubblica informazione come Google, i cui criteri di selezione dei contenuti dovrebbero essere quanto più neutri e mirati a fornire la migliore informazione possibile, piuttosto che assecondare necessità personali.

La strategia reputazionale additiva, in effetti, non mira ad affrontare direttamente e con trasparenza il problema della presenza di un'informazione compromettente, come avviene invece quando si ricorre ad un'operazione sottrattiva: chiedere al *webmaster* o a Google di considerare la rimozione di un contenuto, così appellarsi alla Giustizia per far valere i propri diritti di riservatezza, implicano che la natura lesiva dell'informazione sia affrontata pubblicamente e solo quando la comunità (rappresentata da chi ha emesso la notizia, il motore di ricerca o la Legge) riconosce il suo grado di sbilanciamento tra *privacy* e diritto di cronaca è possibile ottenere la sua eliminazione.

La strategia additiva si limita, al contrario, ad aggirare il problema: agisce sulla visibilità del contenuto per vie indirette e senza necessariamente verificare la liceità dell'azione correttiva, trattando l'informazione compromettente non come violazione di un diritto da denunciare, ma piuttosto come un cadavere da occultare.

La legittimità morale della pratica di riduzione di visibilità di informazioni personali ricade, ancora una volta, nel problema di equilibrio fra diritti della persona e diritti della comunità, ossia fra diritto di *privacy* e diritto di *cronaca*. È chiaro che ogni caso specifico andrà collocato tra questi due poli prima di considerare accettabile l'applicazione delle strategie correttive e che nessuno potrà mai ostacolare l'uso improprio delle tecniche coinvolte. Tuttavia, è bene fare alcune considerazioni per scagionare la strategia additiva dall'accusa di essere semplice trucco dissimulatore.

Per quanto finalizzata a nascondere notizie, la strategia additiva agisce in piena sintonia con la filosofia dei motori di ricerca e più in generale del Web: essa infatti si basa sulla partecipazione attiva e pertinente di una persona all'interno dell'intera comunità del Web con contenuti di qualità.

Questa persona smette di essere semplice utente, ma diventa autore del Web fornendo informazioni di valore sia da un punto di vista contenutistico, quindi lato uomo, sia di leggibilità da parte dei motori di ricerca, quindi lato macchina. Il Web si arricchisce di contenuti di qualità, la persona migliora e solidifica la sua identità digitale mentre agli utenti non è assolutamente negata la possibilità di informarsi, anzi, vengono offerte loro informazioni nuove e aggiornate.

Mentre rimuovere un'informazione richiede necessariamente un'autorizzazione da parte della comunità, la quale deve rinunciare ad avervi accesso, l'aggiunta di nuovi contenuti asseconda semplicemente il naturale processo per cui l'informazione vecchia cede la propria visibilità a quella nuova e più rilevante, senza in alcun modo impedire che i vecchi contenuti vengano comunque trovati e consultati.

Inoltre, la funzione correttiva della strategia additiva si colloca all'interno di un obiettivo molto più ampio del semplice insabbiamento di informazioni scomode: diventare parte attiva del Web con prodotti di qualità, significa assumere progressivamente il controllo sulla propria identità digitale, nonché costruire l'immagine di sé secondo le proprie necessità e aspettative.

Agire per addizione, nei fatti, significa occupare la prima pagina dei risultati di ricerca del nostro nome e cognome con contenuti positivi e proprietari. Questo non si traduce esclusivamente nella possibilità di seppellire eventuali contenuti indesiderati, ma anche, e soprattutto, nell'assunzione del controllo della SERP che ci rappresenta.

Costruire una solida identità digitale, ben controllata, positiva e perti-

nente, garantisce non soltanto la barriera più sicura contro eventuali attacchi reputazionali, anche se per il momento non se ne corre il rischio, ma consente di espandere e sfruttare al meglio le potenzialità della Rete a nostro favore: non avere una cattiva reputazione non implica averne una buona, ma ci colloca in un limbo che nel migliore dei casi avrà un effetto neutrale sull'opinione di chi si informerà su di noi in Rete e nel peggiore susciterà scarsa considerazione o addirittura sospetto. Certamente, un'identità digitale nulla o poco curata non eserciterà mai una funzione positiva nei nostri confronti.

Se si sceglie di entrare a far parte della comunità del Web, scelta ormai obbligata per la maggior parte delle persone e che in ogni caso molti di noi ha già preso irreversibilmente, comporta accettare la riproducibilità tecnica del proprio Io. L'Io *online* esiste già, resta compito nostro la cura proattiva della propria immagine sui motori di ricerca, per costruirsi il giusto contesto, prima che siano gli altri a crearlo, nella quale inserire e sviluppare la propria immagine digitale.

Lo sviluppo della reputazione sul Web deve fondarsi su un'efficace strategia basata sui contenuti necessariamente preceduta, e successivamente affiancata, da una dettagliata mappatura della nostra presenza nella Rete, dal tracciamento dell'evoluzione del nostro Io *online* e del suo comportamento in relazione alle nostre azioni e a quelle degli altri.

Strumenti di monitoraggio e raccolta dati, di cui quello presentato nei capitoli precedenti vuole essere solo una proposta, sono la chiave per un approccio sistematico e strutturato all'analisi dell'identità digitale, che non deve ridursi alla mera sfera intuitiva (anche se, in materia di reputazione, resta un fattore mai eliminabile del tutto) ma avvicinarsi quanto più possibile al metodo scientifico. Se l'Io *online* si costituisce di frammenti della nostra identità riprodotti tecnicamente e sparsi per il Web, la possibilità di rintracciare e raccogliere questi ultimi in modo automatico, sistematico e omogeneo diventa occasione di aumentare la capacità di controllo sulla nostra immagine nel mondo virtuale.

Bibliografia

Identità digitale e Società dell'Informazione

Ale Agostini, Antonio de Nardis, *La tua reputazione su Google e i Social Media. Prevenire, monitorare, curare*, Hoepli, Milano, 2013.

Andrea Barchiesi, *La tentazione dell'oblio*, Franco Angeli, Milano, 2016.

Daniel Bell, *The Coming Of Post-industrial Society: A Venture in Social Forecasting*, Basic Books, New York, 1973.

Daniel Bell, *The social framework of information society*, in T. Forester(a cura di),*Microelectronics revolution*, Oxford, 1980, pp. 501-549.

Michael Fertik, *Reputation Economy. How to Optimize Your Digital Footprint in a World Where Your Reputation Is Your Most Valuable Asset*, Crown Business, New York, 2015.

Henry Jenkins, *Convergence culture: where old and new media collide*, New York University Press, New York, 2006.

Viktor Mayer-Schönberger, *Delete. The Virtue of Forgetting in the Digital Age*, Princeton University Press, Princeton, 2015.

Jenny Rayner, *Managing Reputational Risk. Managing Reputational Risk Leveraging opportunities, Curbing Threats*, John Wiley & Sons, Indianapolis, 2003.

Daniel J. Solove, *The Digital Person: Technology and Privacy in the Information Age*, NYU Press, New York, 2004.

Laura Sartori, *La società dell'informazione*, Il Mulino, Bologna, 2012.

Daniel J. Solove, *The Digital Person: Technology and Privacy in the Information Age*, NYU Press, New York, 2004.

David Weinberger, *Everything is miscellaneous: the power of the new digital disorder*, Henry Holt and Company, New York, 2007.

Diritto all'oblio e responsabilità giuridiche del motore di ricerca

Tommaso Auletta, *Diritto alla riservatezza e "droit à l'uobli"*, in G. Alpa, M. Bessone, L. Bonechi, G. Caiazza (a cura di), *L'informazione e i diritti della persona*, Napoli, 1983, pp. 127 e ss.

Corte di giustizia dell'Unione europea, *Il gestore di un motore di ricerca su Internet è responsabile del trattamento da esso effettuato dei dati personali che appaiono si pagine web pubblicati da terzi*, Comunicato Stampa, n. 70/14, Lussemburgo, 13 maggio 2014.

Luciana De Grazia, *La libertà di stampa e il diritto all'oblio nei casi di diffusione di articoli attraverso Internet: argomenti comparativi*, "AIC", 4/2013, 29 ottobre 2013, http://www.rivistaaic.it/la-libert-di-stampa-e-il-diritto-all-oblio-nei-casi-di-diffusione-di-articoli-attraverso-internet-argomenti-comparativi.html, 19/4/2017.

Gisella Finocchiaro, *Il diritto all'oblio ne quadro dei diritti della personalità*, "Il diritto dell'informazione e dell'informatica", XXIX, 4 maggio 2014, pp. 591 - 604.

Michele Nisticò (a cura di), *Internet e costituzione. Atti del convegno Pisa, 21 - 22 novembre 2013*, G. Giappichelli Editore, Torino, 2014.

Sabrina Peron, *Sulla diffamazione commessa tramite motore di ricerca*, "Responsabilità civile e previdenza", 117, n. 6 (2011), pp. 1327 – 1335.

Ruben Razzante, *Manuale di diritto dell'informazione e della comunicazione*, Wolters Kluwer, Italia, 2016.

Jeffrey Rosen, *The Web Means the End of Forgetting*, "New York Times", 20 luglio 2010, http://www.nytimes.com/2010/07/25/magazine/25privacy-t2.html?pagewanted=all, 19/4/2017.

Mark Scott, *Europe Tried to Rein In Google. It Backfired*, "New York Times", 18 aprile 2016, https://www.nytimes.com/2016/04/19/technology/google-europe-privacy-watchdog.html, 19/4/2017.

Web scraping e SEO

Chitika, *The Value of Google Result Positioning*, "Chitika.com", 7 giugno 2013, https://chitika.com/google-positioning-value, 19/4/2017.

Eric Enge, Stephan Spencer, Jessie Stricciola, *The Art of SEO. Mastering Search Engine Optimization*, O'Reilly Media, Sebastopol, 2015.

Eric Enge, *Authorship Adoption Fail - Detailed Stats*, "Stone temple Consulting", 9 settembre 2014, https://www.stonetemple.com/authorship-adoption-fail-detailed-stats/, 19/4/2017.

Mukarram Khalid, *[Python] Making Your Own Web Scraper & Mass Exploiter*, "Mukarram Khalid", 26 agosto 2015, https://mukarramkhalid.com/python-making-your-own-google-scraper-mass-exploiter/, 19/4/2017.

Richard Lawson, *Web Scraping with Python. Scrape Data from any Website with the Power of Python*, Packt Publishing, Birmingham - Mumbai, 2015.

Jacopo Matteuzzi, *Social signals e SEO: i risultati di cinque studi*, "Studio Samo", 20 gennaio 2014, https://www.studiosamo.it/seo/social-signals-e-seo-risultati-di-cinque-studi/, 19/4/2017.

Ryan Mitchel, *Web Scraping with Python. Collecting Data from the Modern Web*, O'Reilly Media, Sebastopol, 2015.

Danilo Petrozzi, *Python: come fare scraping dei risultati di ricerca di Google*, "Eternal Curiosity", 12 settembre 2014, http://eternalcuriosity.it/python-come-fare-scraping-dei-risultati-di-ricerca-di-google, 19/4/2017.

Stephan Spencer, *Google Power Search*, O'Reilly Media, Sebastopol, 2011.