Cosimo Comella

Indici, sommari, ricerche e aspetti tecnici della 'de-indicizzazione'

> *«Lo recuerdo (yo no tengo derecho a pronunciar ese verbo sagrado, sólo un hombre en la tierra tuvo derecho y ese hombre ha muerto) con una oscura pasionaria en la mano, viéndola como nadie la ha visto, aunque la mirara desde el crepsculo del día hasta el de la noche, toda una vida entera»*. (*Funes el Memorioso,* J.L. Borges)

1. Introduzione

Avere a disposizione strumenti efficienti per l'individuazione e il reperimento di informazioni sulla rete è stata sempre una necessità avvertita dalla comunità informatica e dagli sviluppatori di Internet anche prima dello sviluppo e della diffusione della tecnologia e dei servizi World Wide Web. Oggi questi strumenti, che spaziano dai semplici metodi di accesso ai più sofisticati sistemi di *information retrieval*, conoscono il loro apice di successo, anche commerciale, nella tecnologia dei motori di ricerca, cui si affiancano altri sistemi più pecialistici di *business intelligence* e di analisi dei dati secondo il paradigma dei *big data.*

La loro applicazione a informazioni di carattere personale, insieme alla ubiquità e pervasività di Internet e alla semplicità d'uso degli strumenti di

ricerca accessibili al pubblico, espongono tuttavia l'individuo alla perdita di controllo sui propri dati personali, che possono entrare a far parte di una sorta di memoria permanente della rete contro la volontà e all'insaputa degli interessati a cui sono riferiti (il cosiddetto *inconscio digitale*). In questi casi tentare di sottrarsi all'invadenza di una continua proiezione e propagazione in rete di informazioni per effetto dei *search engines* può avere perfino effetti paradossali, contribuendo all'ulteriore diffusione di notizie o dati di cui si vorrebbe limitare la conoscibilità. Nel seguito verranno esposti i principali strumenti disponibili, con le attuali tecnologie, per esercitare una qualche forma di controllo, anche *ex post*, sui dati pubblicati con la tecnologia *web* in Internet, discutendone l'efficacia e le limitazioni.

2. Searching and sorting

Le tecniche di ricerca e di indicizzazione delle informazioni hanno sempre costituito un primario campo di ricerca per i *computer scientist* fin dagli albori dell'informatica, soprattutto nella forma di metodi di accesso a informazioni strutturate organizzate in grandi *data base* oppure in quella di algoritmi di ricerca in strutture dati, con applicazioni e ricadute tecnologiche su tutta l'informatica, dall'elettronica dei circuiti logici ai sistemi operativi, dal *software* di base fino alle applicazioni, alle reti distribuite e ai servizi di *cloud computing*. Non a caso Donald E. Knuth dedicò al tema degli algoritmi di ricerca, nel 1973, il terzo volume della sua opera fondamentale *The Art of Computer programming*[1], iniziata nel 1962: gli algoritmi di ricerca, al pari di quelli di ordinamento e di altri algoritmi fondamentali, sono alla base non solo della formazione degli studiosi ma anche del funzionamento di quasi tutto il *software* che quotidianamente adoperiamo o che viene utilizzato per far funzionare i servizi della società dell'informazione. Il problema della ricerca efficiente di dati in grandi archivi, siano essi strutturati in *database* ovvero non strutturati e consistenti in raccolte di informazioni di vario tipo e formato o in informazione sparsa, è quindi centrale ell'informatica e non nasce con Internet o con il *web*, anche se con questi si trasforma da problema meramente tecnologico in fenomeno sociale ed economico[2].

[1] D. E. Knuth, *The Art of Computer programming*, vol. 3, *Sorting and Searching*, Addison-Wesley, 1973.
[2] Hal R. Varian, *The Economics of Internet Search*, Rivista di Politica Economica, November-December 2006.

3. Indici e sommari della rete

Problema comune a tutte le tecniche di *information retrieval* è il descrivere opportunamente l'informazione da cercare e, sulla base di una metrica predefinita, che consenta di calcolare la distanza dei possibili risultati dal *target* di ricerca, individuare la collocazione delle informazioni corrispondenti o correlate a quella cercata nell'ambito di raccolte dati di elevata dimensione.

Al crescere del volume di informazioni disponibili per la consultazione in rete, equiparabili ormai per vastità a quella *Biblioteca di Babele* immaginata da Jorge Luis Borges[3], l'importanza degli indici e dei metodi di catalogazione e recupero selettivo delle informazioni è determinante per la fruizione della rete come serbatoio di conoscenza a disposizione dell'individuo.

Nello sviluppo di Internet già la fruizione efficiente di alcuni servizi di rete poneva dei problemi di indicizzazione e di efficiente recupero di informazioni senza un controllo centrale: in questa fase nasce il *Domain Name System* (DNS)[4], sistema di *naming* distribuito che consente tuttora, con estrema efficienza, di cercare e recuperare informazioni (indirizzi IP, nomi a dominio, informazioni tecniche) necessarie alla comunicazione in rete tra utenti e tra sistemi.

Più vicini ai *search engines* oggi utilizzati sono i primi strumenti di indicizzazione di file e documenti accessibili nella rete Internet prima dell'avvento del *World Wide Web*, che conobbero la loro massima espansione nei primi anni '90 con i servizi *Archie* e *Veronica*.

Mentre il primo[5] consentiva di selezionare gli *FTP* (*File Transfer Protocol*) *server* che mettevano a disposizione per il *download*, con la 'modalità anonima' di trasferimento dei file (*anonymous ftp*), i documenti digitali o codici programmativi di vario tipo cercati in base al nome del rispettivo file o alla eventuale sua descrizione, con Veronica[6] era possibile effettuare ricerche nel cosiddetto *GopherSpace*, ovvero tra i documenti multimediali accessibili sui *Gopher server* sparsi in Internet e che costituirono un primo esempio funzionante di rete ipertestuale di dimensione

[3] J.L. Borges, *La biblioteca de Babel*, in *Ficciones*, SUR, 1944.
[4] P. Mockapetris, Internet Protocol Suite RFC 882, *Domain Names – Concepts and Facilities*, November 1983.
[5] P. Deutsch, *Archie - An Electronic Directory Service for the Internet*, Computing Centre, McGill University, 1990.
[6] S. Foster - F. Barrie, *Very Easy Rodent Oriented Netwide Index to Computerized Archives*, University of Nevada, Reno, 1992.

globale che precedette il *World Wide Web.* Analogamente, il problema della ricerca si era posto già rispetto all'enorme mole di informazioni contenuta nella rete di notizie *Usenet*, sviluppatasi a partire dal 1979 indipendentemente da Internet/ARPAnet e basata inizialmente su protocolli UUCP (*Unix-to-Unix Copy Program*). Con il *porting* di Usenet sui protocolli TCP/IP di Internet la richiesta di più ampie funzionalità di ricerca nelle *news* produsse diversi sistemi, il più famoso dei quali è stato il servizio *DejaNews* accessibile tramite *web.*

«Remember - your current or future employers may be reading your articles. So might your spouse, neighbors, children, and others who will long-remember your gaffes».

(da «Hints on Writing Style For UseNet», *post* permanente nel gruppo di discussione Usenet *news.announce.newusers*, 1991)

Proprio la formazione di questo primo grande archivio *online* dei c.d. *post* circolati sulla rete Usenet dal 1981 in poi[7] pose su larga scala i primi problemi di *privacy* (così come di diritto d'autore relativamente alla diffusione di contenuti protetti da *copyright* attraverso le cosiddette gerarchie *alt.binaries.*)* relativi alla ripubblicazione e all'accresciuta diffusione di messaggi i cui autori, nonostante i *caveat* presenti già nei *tutorial* per l'uso della rete, non si rendevano conto dei fattori moltiplicativi legati alla circolazione globale delle notizie. Men che meno essi potevano immaginare l'ulteriore diffusione, via *web*, di messaggi affidati anni prima al sistema dei *newsgroups*, quando Usenet era utilizzata da una relativamente ristretta cerchia di utenti e prevedeva l'utilizzo di *software* specifico per la consultazione e la creazione delle *news.*

4. I motori di ricerca web

4.1. Sviluppo dei motori di ricerca

La diffusione del *World Wide Web*, divenuto fenomeno di massa dal 1994-1995, ben presto amplificò le esigenze di recupero di informazioni, sia per l'ampiezza della rete ipertestuale che si andava sviluppando, sia per la varietà di contenuti e di formati digitali in essa reperibile: testi, docu-

[7] Tale servizio è stato acquisito da Google nel 2001 ed è entrato a far parte, come *Google Groups*, dei servizi offerti agli utenti del motore di ricerca.

menti, immagini, filmati, registrazioni sonore, file di ogni tipo e natura. Da qui parte il forte sviluppo della tecnologia dei motori di ricerca, basato sui risultati scientifici conseguiti nei decenni precedenti nel campo degli algoritmi dell'*information retrieval* e dei *data base management systems.* I motori di ricerca via via resi accessibili al pubblico di Internet vengono utilizzati tramite l'inserimento di parole chiave testuali, che possono anche essere composte tra loro tramite operatori logici (perché derivati dalla logica booleana), per formare delle complesse *query strings.* Al di là di quale sia l'oggetto della ricerca, sia esso quindi un contenuto documentale testuale o multimediale, lo strumento di ricerca è di tipo *text-based*, perché prevede un'interazione basata sull'inserimento di stringhe di caratteri e sulla loro successiva ricerca all'interno di contenuti pubblicati.

Strumenti più avanzati consentono di cercare un'informazione a partire da chiavi non soltanto testuali: per la ricerca di immagini in rete può infatti essere più adatto uno strumento che permetta di inserire come chiave una fotografia o un altro contenuto digitale (filmati, registrazioni sonore...) per cui si vogliano trovare delle corrispondenze, piuttosto che inserire delle parole da confrontare con etichette di testo (*tags*) che talvolta arricchiscono la descrizione dei contenuti digitali presenti in rete. Ai motori di ricerca testuali si affiancano dei sistemi di ricerca più evoluta, che consentono di specificare l'oggetto della ricerca in linguaggio naturale o prossimo al linguaggio naturale. Nell'esperienza della maggior parte degli utenti il concetto di motore di ricerca è però attualmente associato all'utilizzo di funzionalità di ricerca testuale, per parole-chiave, proposto in rete da servizi *online* quali Excite, Altavista, Yahoo!, Bing e altri, fino a Google, maggiore operatore del settore. Quest'ultimo, abbinando la raccolta pubblicitaria (caratterizzata da originali sistemi di fissazione del prezzo per i contratti di *advertising*)[8] a una notevole efficienza nel fornire i risultati delle ricerche, ha assunto una tale posizione dominante, soprattutto nel contesto europeo, da poter essere assunto come riferimento per una discussione sulle implicazioni di *data protection* che scaturiscono, in generale, dalla tecnologia dei *search engines* in Internet. Anche a causa di queste caratteristiche di efficienza tecnica e di forza di mercato, il servizio Google Search è stato al centro di diversi casi giudiziari, fino alla recente decisione della Corte di giustizia dell'Unione Europea in tema di diritto all'oblìo nel c.d. *caso Google Spain.*

[8] M.Naldi, G. D'Acquisto, G. F. Italiano, *The value of location in keyword auctions, in Electron. Commer. Rec. Appl.* 9, 2 (March 2010), pp. 160-170.

4.2. Indici e sommari nel modello di Google Search

Il funzionamento di Google Search, come quello di analoghi servizi, è basato sull'utilizzo di un agente automatico, detto *spider* o *crowler*, per la raccolta delle pagine *web* accessibili *online*, esplorate tramite algoritmi di visita su grafi che consentono di percorrere tutte le diramazioni della rete di riferimenti ipertestuali in modo efficiente. Il buon esito di una ricerca presuppone quindi che il contenuto cercato sia stato oggetto di 'visita' da parte dello *spider* affinché i dati contenuti nelle pagine ispezionate possano popolare la gigantesca *cache*[9] del motore di ricerca ed essere lì indicizzati e resi disponibili per la consultazione. Essenziale per l'efficacia del servizio di ricerca è la determinazione della *frequenza di visita* dei diversi siti: stime della stessa Google fanno ritenere che, in assenza di *input* specifici, occorrano diverse settimane o anche mesi per ottenere un completo *refresh* della *cache*, ottenibile con una visita esaustiva della *websphere* tramite gli *spider*. Va da sé che i siti più visitati dagli utenti, più ricchi di contenuti in continuo aggiornamento e più citati tramite *link* ipertestuali (HREF o *hypertext reference*) sono quelli più frequentemente ispezionati dalle sonde di indicizzazione.

5. Memoria della rete e controllo dei dati

5.1. Aggiornamento delle notizie e coerenza dell'indice

Oltre a rendere disponibili le informazioni su cui costruire l'indice e far funzionare l'algoritmo di ricerca, la presenza di una memoria *cache* a supporto di un *search engine* permette di presentare all'utente un'anticipazione del risultato completo in forma di riassunto (*summary*). Laddove si verifichi un disallineamento tra il contenuto presente in *cache* e la pagina originaria, la *cache* fornirà una visione non aggiornata del contenuto pubblicato, che potrebbe esser stato modificato o anche rimosso dal sito di originaria pubblicazione. Per questo motivo, è fondamentale minimizzare i tempi di esposizione a consultazione di risultati non aggiornati, associando una sorta di 'misura di freschezza' alle informazioni caricate nell'indice. Le modalità con cui questa misura viene calcolata costituiscono oggetto

[9] Memoria veloce in grado di assicurare l'immediata disponibilità di un dato presente a livello inferiore in una struttura a gerarchia di memoria.

di ricerca nel campo della *computer science*, con diverse differenti proposte rinvenibili in letteratura[10] [11].

Quando una notizia indicizzata da un motore di ricerca riporta informazioni riferibili a un individuo, la misura tecnica della freschezza dell'indice è anche, in qualche modo, misura del corretto trattamento dei dati personali dell'interessato: l'aggiornamento delle informazioni riferibili agli individui è uno dei valori della protezione dei dati personali, anche nell'ordinamento italiano, laddove l'articolo 11 del Codice in materia di protezione dei dati personali (d.lgs. 30 giugno 2003, n. 196) prevede che i dati personali oggetto di trattamento siano «esatti e, se necessario, aggiornati».

Benché la responsabilità di mantenere informazioni esatte e aggiornate gravi *in primis* su chi le ha pubblicate per primo, l'attività di indicizzazione e di *caching,* con la successiva presentazione di risultati in forma di sommario, è molto vicina, o del tutto coincidente, a una vera e propria ripubblicazione delle stesse informazioni, con tutte le implicazioni che da ciò possono derivare in tema di responsabilità del *publisher* e di doveri rispetto all'esercizio dei diritti degli interessati.

Il motore di ricerca Google ha messo a disposizione degli utenti delle apposite sezioni (*Google Webmaster Tools*) per segnalare la presenza nell'indice di *link* a pagine non più esistenti oppure modificate e perciò non coerenti con le chiavi di ricerca o che presentano nel *summary* delle notizie non aggiornate, differenti da quelle correntemente *online* sul sito di originaria pubblicazione. In assenza di segnalazione il motore di ricerca potrebbe impiegare anche alcuni mesi prima di accorgersi che una data pagina sia stata modificata o addirittura rimossa, con ritardo dipendente da molti fattori, tra cui la popolarità della notizia e del sito che la pubblica.

La richiesta di rimozione di pagine non più esistenti o il loro aggiornamento non richiede di essere utenti registrati ed è accessibile in italiano all'indirizzo https://www.google.com/webmasters/tools/removals?pli=1 .

Analoghi strumenti sono messi a disposizione da altri motori di ricerca, come *Yahoo Search* e *Bing* che offrono ai propri utenti i *Bing Webmaster Tools* e i *Content Removal Tools*[12] tramite cui è possibile rimuovere elementi della *cache* scaduti o i c.d. *broken links.*

[10] J. Cho - H. Garcia-Molina, *Synchronizing a database to improve freshness*, in Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, Dallas, Texas, 2000, pp. 117-128.

[11] E. G. Coffman Jr - Z. Liu - R.R. Weber, *Optimal Robot Scheduling for WebSearch Engines*, 1 *Journal of Scheduling* pp. 15-29 (1998).

[12] http://www.bing.com/webmaster/help/bing-content-removal-tool-cb6c294d [consultato il 30/09/2014].

5.2. Controllo delle informazioni indicizzabili.

Già a partire dalla seconda metà degli anni '90 la comunità Internet rese disponibili strumenti che permettevano di affrontare l'esigenza, sentita per lo più dai gestori di siti *web*, di limitare la possibilità da parte dei motori di ricerca di indicizzare i contenuti pubblicati *online* e non protetti con accorgimenti di sicurezza, perché lasciati intenzionalmente disponibili per la consultazione da parte di chiunque. L'esigenza di escludere alcuni contenuti dall'azione invasiva dei motori di ricerca aveva originariamente motivazioni del tutto differenti da quelle che oggi appaiono più rilevanti: in un'era, quella degli anni '90, caratterizzata da un livello di prestazioni dei sistemi informatici di almeno due ordini di grandezza inferiore rispetto a quello dei sistemi odierni (a parità di categoria)[13], era necessario evitare il sovraccarico dei *server* derivante dalle connessioni intensive degli *spider* dei motori di ricerca, soprattutto quando queste non avrebbero poi potuto produrre risultati apprezzabili in termini di arricchimento degli indici, per la presenza di contenuti difficilmente indicizzabili, ovvero poveri di testo (immagini, video, audio...).

Escludere dalla visita degli *spider* alcune porzioni dei *filesystem* che ospitavano le pagine e i *file* accessibili tramite il protocollo http (*Hypertext Transfer Protocol*)[14] diventava allora una questione di *performance* significativa, soprattutto per siti con grandi basi documentali.

La capacità di controllare l'indicizzazione e quindi la diffusione ulteriore di dati pubblicati in rete, generata o amplificata dall'azione dei motori di ricerca, è oggi invece più importante per assicurare quel *right to be forgotten,* che costituisce uno dei punti più importanti della proposta di regolamento europeo sulla protezione dei dati personali presentata dalla Commissione Europea nel gennaio 2012[15]. A questo proposito, la *European Union Agency for Network and Information Security* (ENISA) ha

[13] Una differenza di un fattore 100 è riscontrabile sia nei livelli di integrazione, espressi in numero di transistor per circuito, sia nella frequenza di clock delle CPU su microprocessore, passata dagli 8 MHz del 1988 al GHz del 2001.

[14] Inizialmente pubblicato come HTPP V0.9, il protocollo venne standardizzato da IETF con la pubblicazione dello IPS RFC 1945 «Hypertext Transfer Protocol -HTTP/1.0» nel maggio 1996.

[15] Proposta di Regolamento del Parlamento Europeo e del Consiglio concernente la tutela delle persone fisiche con riguardo al trattamento dei dati personali e la libera circolazione di tali dati (regolamento generale sulla protezione dei dati). 25/01/2012. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri =COM:2012:0011:FIN:IT:PDF. [consultato il 30/09/2014].

pubblicato già nel 2011 un primo interessante rapporto sugli strumenti tecnici disponibili per implementare quel diritto nel contesto della pubblicazione in rete o dell'uso di dati personali nei sistemi informativi[16].

Tradurre con strumenti tecnici efficaci e accessibili alla maggior parte degli utenti delle rete, e anche agli interessati che non la utilizzano, l'esercizio dei propri diritti in tema di dati personali costituisce comunque una sfida tecnica di grande portata, che solo da qualche anno vede i principali gestori di *search engines* impegnati nello sviluppo di apposite funzionalità da integrare nei propri servizi *online.*

5.3. Il Robots Exclusion Protocol

Lo strumento sviluppato dalla comunità Internet negli anni '90 è stato il *Robots Exclusion Protocol* (REP)[17], ovvero una sorta di convenzione tecnica di cui, da una parte, gli sviluppatori di *software* per servizi *web* e, dall'altra parte, i programmatori delle funzioni di indicizzazione dei principali motori di ricerca possono tener conto nella realizzazione dei rispettivi prodotti, affinché la visibilità dei contenuti di un sito tra i risultati di una ricerca sia modulabile dal gestore tramite la compilazione di un *file* di tipo testuale denominato *robots.txt* e presente nella *root directory* dei *web server* in modo tale da poter essere letto all'indirizzo http://*nome.a.dominî.del.sito*/robots.txt.

Il formato di robots.txt

Il file *robots.txt* contiene dei *record* raggruppati in sezioni introdotte dal campo *User-agent* seguito da uno o più direttive *Disallow.*

Il campo *User*-agent indica a quale *search engine* le direttive successive sono rivolte. La sua sintassi è:

User-agent:<spazio><nome dello spider>

La direttiva *Disallow* serve a indicare allo *spider* indicato nel campo *User-agent* quali *pagine* o *directory* non può acquisire, con la sintassi seguente:

[16] Il rapporto è disponibile sul sito di Enisa all'indirizzo: https://www.enisa.europa.eu/activities/identity-and-trust/library/deliverables/the-right-to-be-forgotten/at_download/fullReport [consultato il 30/09/2014].

[17] M. Koster, *A Standard for Robot Exclusion*, 30/06/1994 (testo originale in *http://www.redhat.com/support/resources/faqs/RH-apache-FAQ/misc/norobots.htmle*) [consultato il 30/09/2014].

Disallow:<spazio><nome file o nome directory>

I nomi dei *file* o delle *directory* sono espressi come indirizzi relativi, nel *filesystem* del *server* che ospita le pagine *web*, alla *home directory* dei contenuti del sito, indicata da un carattere *slash* (/).

Il campo User-agent può contenere un asterisco (*) per significare 'qualunque spider'.

User-agent: *

Disallow: /pagina/da_non_indicizzare.html

Disallow: /pagine/da_non/indicizzare/

Il campo *Disallow* può contenere anche un singolo carattere *slash* a indicare 'qualunque *file* o *directory*', oppure può essere lasciato vuoto, per indicare che non ci sono pagine sottratte all'indicizzazione.

Il REP prevede che ogni richiesta di acquisizione di una pagina *web* da parte di uno *spider* debba essere preceduta dalla verifica di non inclusione dell'indirizzo della pagina stessa nella 'lista di esclusione' del sito, descritta proprio dal contenuto del *file robots.txt.*

Il REP prevede anche altre modalità di esclusione, tra le quali il ricorso ad appositi *metatags noindex* da incorporare nelle pagine Html al fine di interrompere il percorso di visita degli agenti di indicizzazione, impedendo che questi seguano i *link* ipertestuali presenti nelle pagine in forma di *tag* HREF (*Hypertext Reference*).

In molti casi l'efficacia di funzionamento del REP è rafforzata dall'utilizzo congiunto di tutte e due le tecniche.

Azioni preventive alla pubblicazione

Qualora si sia in presenza di dati non ancora pubblicati, e perciò non presenti in nessun indice di ricerca, il file robots.txt può essere predisposto in modo opportuno affinché la notizia o la pagina web che si sta per inserire non venga acquisita dagli agenti di ricerca. L'indicazione delle pagine da non indicizzare non deve essere necessariamente puntuale, ma può comprendere tutti i documenti presenti in una cartella, o *directory*, di *filesystem.*

È evidente che l'utilizzo di indirizzi puntuali può più facilmente generare gli effetti paradossali cui si è fatto cenno prima, prestandosi anche a un'analisi perfino solo visiva del file *robots.txt.*

Lo stesso accade inserendo nel file l'indicazione delle *directory* da escludere laddove queste abbiano nomi immediatamente riferibili al loro contenuto.

Una volta compreso l'indirizzo della pagina tra quelli da non indicizzare, si può procedere alla pubblicazione *online* del contenuto, con ragionevole aspettativa — se la configurazione del *robots.txt* è stata correttamente realizzata — che né Google Search, né Yahoo né gli altri principali servizi di indicizzazione e ricerca web saranno in grado di acquisirlo e indicizzarlo.

Azioni a posteriori

Nel caso in cui un certo contenuto sia stato già indicizzato perché pubblicato e non escluso, in un primo momento, dall'indicizzazione tramite gli strumenti del REP, e lo si voglia rimuovere dai risultati di un motore di ricerca occorrerà agire su più fronti.

In primo luogo occorrerà configurare opportunamente il *robots.txt* o utilizzare i *metatag noindex* per segnalare che una pagina o dei *link* ipertestuali non devono essere visitati e indicizzati.

Successivamente occorrerà agire presso i motori di ricerca affinché i riferimenti alla pagina in questione siano espunti dai rispettivi indici Nel caso di Google, la società ha predisposto, nell'ambito dei Webmaster Tools, oltre alla possibilità di richiedere la rimozione dall'indice di riferimenti a pagine non più esistenti o per le quali sia stata modificata la visibilità da parte degli *spider* di ricerca, anche l'aggiornamento dei riferimenti a seguito di modifiche apportate ai documenti indicizzati.

5.4. Limitazioni del Robots Exclusion Protocol

Come accennato, nonostante il nome, il REP non è un vero protocollo, né *standard* cui l'industria del *software* debba necessariamente attenersi. Non vi è quindi garanzia, pur in presenza di una corretta configurazione del *robots.txt*, che una certa pagina non venga indicizzata da un qualsiasi motore di ricerca, se la si vuole mantenere comunque accessibile via *web* a un pubblico indeterminato.

D'altra parte, anche i più avanzati motori di ricerca non rispettano pienamente il REP, limitando in molti casi l'efficacia dello strumento e sminuendone l'importanza. Per quanto riguarda Google, le azioni per limitare l'indicizzazione dei contenuti di siti *online* si sono basate sulla conformità del motore di ricerca al REP, anche se il livello di *compliance* è

stato soltanto parziale. Probabilmente un'applicazione più rigorosa e più efficace del REP avrebbe permesso una più agevole modulazione della visibilità di alcune informazioni nei risultati di ricerca e avrebbe consentito all'azienda di non subire gli effetti di una decisione molto severa e di difficile attuazione come quella recente della Corte di giustizia UE con cui l'azienda americana, insieme ad altre parimenti interessate, si sta confrontando in queste settimane.

La stessa Google chiarisce, nella *Google's webmaster documentation*, quale sia l'attenzione rivolta a questo sistema di gestione delle indicizzazioni, fornendo le seguenti indicazioni:

«Blocking Google from crawling a page is likely to decrease that page's ranking or cause it to drop out altogether over time. It may also reduce the amount of detail provided to users in the text below the search result. ... However, robots. txt Disallow does not guarantee that a page will not appear in results: Google may still decide, based on external information such as incoming links, that it is relevant.»

Sempre nelle pagine informative del centro di supporto per i *webmaster* di Google si può leggere il seguente avviso:

«While Google won't crawl or index the content of pages blocked by robots.txt, we may still index the URLs if we find them on other pages on the web. As a result, the URL of the page and, potentially, other publicly available information such as anchor text in links to the site, or the title from the Open Directory Project (www. dmoz.org), can appear in Google search results.» (dal Webmaster Tools Help Center di Google)

La parziale efficacia, dovuta all'interpretazione stessa del REP fatta da Google, ha portato evidentemente a risultati insoddisfacenti e che talvolta rivelano proprio l'incorretta implementazione del REP. Basti osservare come a fronte di determinate ricerche con parole chiave alcuni risultati vengano presentati con l'indirizzo URL deprivato del sommario (copia *cache*) con la motivazione seguente:

> *«A description for this result is not available because of this site's robots. txt — learn more.»*

È evidente da questo avviso come, a fronte di un contenuto *online* che, nelle intenzioni del *publisher,* avrebbe dovuto essere sottratto all'indicizzazione da parte dei *crawler* o *spider* dei motori di ricerca (e di ciò è rivelatore l'esplicito riferimento di Google a questa circostanza, anche se in riferimento alla mera assenza del sommario o copia *cache*), questo possa essere nonostante ciò incluso tra i risultati e correlato a determinate parole chiave, rivelando la sua presenza nell'indice di ricerca!

Questo comportamento svela la violazione o la non corretta implementazione del *Robots Exclusion Protocol* da parte di Google. Altro esempio di mancata *compliance* è quello relativo alla riapparizione periodica dei risultati di ricerca che si vorrebbero esclusi: parrebbe infatti che l'efficacia della de-indicizzazione tramite *robots.txt* sia garantita per soli novanta giorni, trascorsi i quali il motore di ricerca riprende a inserire riferimenti alle pagine che l'originario *publisher* vorrebbe in effetti sottratte all'acquisizione dallo *spider*. Non vi è stata finora una chiara posizione di Google rispetto a questa anomalia, che evidenzia una incorretta applicazione del REP, laddove gli indirizzi delle pagine da acquisire dovrebbero previamente essere controllati per verificarne l'eventuale inclusione nel *robots.txt* del sito da visitare. Le pagine escluse tramite *robots.txt* non dovrebbero quindi essere in alcun modo acquisite, mentre Google evidentemente procede alla loro acquisizione salvo provvedere successivamente a gestirne, con propri criteri che non rispecchiano la volontà del *publisher*, la visibilità tra i risultati di ricerca.

Alcuni cenni a questo comportamento si ritrovano nella *knowledge base* del motore di ricerca e suscitano dibattito tra utenti, sviluppatori e l'azienda stessa. Sul piano della *data protection* e delle relazioni con le autorità nazionali europee (DPA) tuttavia Google non ha finora chiarito questo aspetto e non ha fornito spiegazioni soddisfacenti all'evidente anomalia che inficia l'efficacia e quindi la fiducia del pubblico in un meccanismo (il *robots.txt*) già in sé abbastanza debole.

La pubblicazione della sentenza *Google Spain* nel maggio 2014 ha poi fatto momentaneamente diminuire l'interesse per la questione, che però occorrerà chiarire perché lo strumento *robots.txt* continua a essere utilizzato dai siti *web* e consente in qualche misura, se ben realizzato, di modulare la visibilità delle pagine tra i risultati di ricerca a volontà dei *publisher.* Tale possibilità andrebbe mantenuta e meglio realizzata, anche se gli interessati, quantomeno in Europa, hanno adesso nei confronti di Google e degli altri *search engines* sottoposti all'ordinamento comunitario la possibilità concreta di esercitare il proprio diritto alla cancellazione dei dati dall'indice

tramite procedure alternative, del tutto separate da quelle connesse all'uso del *Robots Exclusion Protocol*, e svincolate dallo stato di pubblicazione di una determinata pagina *online.*

5.5. Effetti paradossali

I siti dei maggiori periodici *online* ricorrono da tempo al *robots.txt* inserendovi gli indirizzi delle pagine o delle *directory* che, pur essendo accessibili *online*, non sono da indicizzare. A seconda del tipo di *Content Management System* (CMS) adottato per la gestione dei contenuti e del processo editoriale *online*, l'indicazione della pagine può essere espressiva del loro contenuto. Poiché il file *robots.txt* deve essere leggibile da chiunque, il visitatore curioso può liberamente esaminarne il contenuto e trarne degli spunti per andare a recuperare in seguito notizie che la volontà dell'editore vorrebbe in qualche modo limitate nella conoscibilità.

Da questo punto di vista, l'utilizzo di indirizzi 'brevi' o 'non parlanti', da cui non è possibile desumere il contenuto anche solo indicativo della pagina cui fanno riferimento, evita di generare effetti paradossali di accresciuta attenzione verso ciò che si vuole in qualche modo proteggere. Altri fenomeni di indicizzazione non voluta possono derivare dalla citazione della stessa notizia o del corrispondente *link* ipertestuale in un'altra pagina, esterna al sito di originaria pubblicazione: se il *link* è 'parlante' l'effetto di inclusione nei risultati di ricerca sarà enfatizzato.

L'efficacia del REP sarà poi vanificata dalla ripubblicazione automatica della notizia su altri siti, tramite funzioni di *mirroring* e altre forme di importazione automatica di contenuti *online.* Infine, il ricorso al *robots.txt* può generare specifica attenzione da parte di utenti della rete proprio verso quei contenuti che si era cercato di sottrarre all'indicizzazione, per il solo fatto che essi siano citati in quel *file.*

Esistono, infatti, in rete siti e servizi dedicati a effettuare le ricerche proprio nella parte 'nascosta' del *web*: in questo senso si collocano iniziative che comportano la costruzione di indici specificamente volti a rendere accessibili proprio quelle informazioni a cui i *webmaster* avevano riservato, nelle intenzioni, un minor grado di conoscibilità.

Tra i diversi siti che offrono servizi *online* di questo tipo ha guadagnato recentemente una certa notorietà '*Hidden from Google*' [18], che però è specificamente volto a elencare e permettere di consultare i siti i cui contenuti pubblicati hanno comportato l'attivazione di Google con gli strumenti

[18] http://hiddenfromgoogle.com/.

approntati dalla società americana per adempiere alle prescrizioni della Corte di Giustizia UE nel noto caso «Google Spain». Presentato come n'iniziativa contro le azioni censorie sul *web*, permette a chiunque di segnalare notizie che, per effetto della sentenza europea, sono state sottratte alla visibilità tra i risultati di ricerca.

5.6. Controllo di accesso con autenticazione o con liste ACL

Le tecniche per la limitazione dell'azione dei motori di ricerca sono molteplici, ciascuna dotata di maggiore o minore efficacia e impatto sulla fruibilità delle informazioni. In primo luogo, chi non voglia conferire alle proprie pubblicazioni *web* una indesiderata pubblicità dovrebbe riflettere seriamente sull'opportunità di mettere *online* quel determinato contenuto in modalità accessibile a chiunque. Il protocollo Http permette infatti la definizione di controlli di accesso tramite la configurazione di un file *htaccess* con cui si può attivare fase di autenticazione con *username* e *password* e si possono definire delle *access control lists* (ACL) con cui abilitare o negare l'accesso a *file* e *directory* del sito in base agli indirizzi di rete di provenienza delle connessioni.

La fase di autenticazione impedisce agli *spider* dei motori di ricerca di proseguire nella visita del sito, mentre la configurazione di ACL selettive impedisce del tutto la connessione di *client* non graditi. Tuttavia effetti avversi si manifestano sull'usabilità del sito e sulla fruibilità da parte del pubblico privo di credenziali di accesso.

Come espediente senza scopi di vera protezione delle informazioni, ma idoneo a sbarrare il percorso di visita di un sito agli agenti di ricerca, viene talvolta utilizzato un *form* per l'autenticazione le cui credenziali sono intenzionalmente pubblicate e conoscibili da chiunque, permettendo l'accesso ai contenuti a un utente ma non a un normale *spider* di un motore di ricerca.

5.7. Uso dei Captcha o di fasi interattive

Effetti interdittivi dell'indicizzazione automatica si ottengono proteggendo le pagine *online* con il ricorso a strumenti Captcha (*Completely Automated Public Turing Test to Tell Computers and Humans Apart*) volti specificamente a discriminare tra utenti umani e programmi automatici di raccolta dei dati interagenti con un'applicazione informatica, secondo il

modello del *test di Turing*[19]. Pur tenendo conto dell'evoluzione degli strumenti informatici che consentono a volte di superare in modo automatico dei controlli *captcha*, questi possono rappresentare una efficace barriera contro l'utilizzo di programmi automatici e, in particolare, degli *spider* di indicizzazione. Analogamente, anche l'interposizione di una qualsiasi fase interattiva, che può essere rappresentata dalla compilazione di un *form* per la ricerca locale delle informazioni o dal *click* su un tasto *submit* per proseguire la navigazione tra i contenuti di un sito, può permettere di proteggere i contenuti *online* dall'invasività di agenti automatici non graditi.

5.8. Il web semantico e il controllo delle informazioni

Con il termine *web semantico*, coniato da Tim Berners-Lee, ideatore del World Wide Web[20], ci si riferisce alle tecnologie e ai metodi per arricchire il web trasformandolo in un ambiente in cui i documenti pubblicati vengono collegati a informazioni e dati (*metadati*) che ne descrivono il 'contesto semantico' con formati idonei all'interrogazione e l'interpretazione automatica. Lo sviluppo del web semantico, pur non progredendo secondo le aspettative dei suoi ideatori iniziali[21], è oggi guidato dal *World Wide Web Consortium* (W3C) e si basa sulla promozione di formati di dato comuni sul *web* in modo da trasformarlo in chiave *datocentrica*, grazie a linguaggi e *framework* descrittivi quali il *Resource Description Framework* (RDF), il *Web Ontology Language* (OWL) e l'*Extensible Markup Language* (XML). Le *ontologie* e i *linked data* costituiscono la base concettuale per la creazione di un nuovo tipo di *web* che renda possibile, tra l'altro, effettuare ricerche molto più sofisticate rispetto ai *search engines* convenzionali, agenti sul livello lessicale del linguaggio, e di creare collegamenti automatici tra documenti e dati andando molto oltre il semplice riferimento ipertestuale, sfruttando la possibilità di penetrare la struttura più profonda, sintattica e semantica, dei dati e dei documenti.

La disponibilità di un livello semantico permette di trattare le differenti porzioni di informazione di cui si compone un documento con flessibilità, disciplinandone il regime di visibilità e protezione dei dati personali, trasformando radicalmente il modo di agire dei motori di ricerca che attualmente indicizzano, in modo che può essere considerato rudimentale,

[19] A.M. Turing, *Computing machinery and intelligence*, in 59 *Mind* pp. 433-460 (1950).
[20] T. Berners-Lee, *L'architettura del nuovo Web*, Feltrinelli, Milano, 2002.
[21] N. Shadbolt, W. Hall, T. Berners-Lee, *The Semantic Web Revisited*, IEEE Intelligent Systems, May/June 2006.

tutto ciò che sono in grado di raggiungere *online*.

I *search engines* d'altra parte ricaverebbero enormi vantaggi in termini di efficienza e di accuratezza dei risultati, avvicinando i risultati di ricerca alle aspettative degli utenti ma permettendo, tra le altre cose, di incorporare nei documenti *online* le regole di trattamento nel senso della *data protection*, compresa la validità temporale di un'informazione, la sua riproducibilità o visibilità in indici e sommari.

6. Diritto all'oblio e motori di ricerca nell'esperienza del Garante

6.1. Reputazione online e diritto all'oblio

Il tema del diritto all'oblio è sempre stato all'attenzione del Garante per la protezione dei dati personali fin dalla sua istituzione. All'esito della trattazione di un ricorso relativo alla persistenza in rete, amplificata dai motori di ricerca, di notizie pregiudizievoli per l'attività di un soggetto imprenditoriale che era stato destinatario di una sanzione per pubblicità ingannevole, il Garante nel 2004 prescrisse al titolare del trattamento l'adozione di accorgimenti tecnici per mitigare gli effetti della pubblicazione *online* di notizie sulle sanzioni comminate[22] quando la finalità originaria della pubblicazione *online*, ammonitoria e informativa, era venuta meno.

Scopo del provvedimento del Garante era evitare che si realizzasse una sorta di gogna permanente, laddove la pubblicazione per estratto di certe decisioni o sentenze viene tuttora prevista, nel nostro ordinamento, quale pena accessoria, in presenza di un lungo periodo di tempo trascorso dalla sanzione e dal ravvedimento dell'operatore economico.

In base al principio del diritto all'oblio sono stati poi trattati dall'Autorità italiana numerosi casi relativi alla ripubblicazione di notizie negli «archivi storici *online*» dei maggiori quotidiani nazionali. Nei provvedimenti in tema di diritto all'oblio qui citati, e in altri nel tempo adottati, il Garante ha prescritto quale strumento tecnico per adempiere il ricorso al *Robots Exclusion Protocol,* misura idonea, nella maggior parte dei casi, a sottrarre all'indicizzazione da parte dei motori di ricerca le informazioni riferibili a interessati la cui diffusione doveva essere limitata. Sono stati trattati così dei casi di pubblicazione di archivi storici di quotidiani e rivi-

[22] http://www.gpdp.it/web/guest/home/docweb/-/docweb-display/docweb/1116068 [consultato il 30/09/2014].

ste, di diffusione di informazioni personali da parte di soggetti pubblici sulla base di interpretazioni estensive delle norme in tema di trasparenza in ambito pubblico, di pubblicazione di notizie inaccurate o non aggiornate, e perciò lesive della dignità di persone, su quotidiani o riviste *online*.

L'incertezza di giurisdizione e la possibilità di sovrapposizione e conflitto tra differenti giurisdizioni hanno fatto considerare per anni poco praticabile l'imposizione in capo ai gestori dei motori di ricerca di obblighi di espungere dai rispettivi indici i riferimenti a notizie che permangano *online,* pur ledendo diritti degli interessati cui si riferiscono, su sistemi di pubblicazione sottratti alla capacità di intervento delle DPA europee. D'altra parte l'orientamento delle autorità europee di protezione dei dati personali, riunite nel *Working Party* istituito dall'Art. 29 della direttiva 46/95/EU (WP29), rispetto alla titolarità del trattamento dei dati personali è stato costantemente in favore dell'assenza di responsabilità del motore di ricerca, essendo la titolarità del trattamento incardinata sul soggetto che pubblica il contenuto *online* nella sua forma originaria[23]. Tale posizione è andata incontro a un importante cambiamento a seguito della sentenza *Google Spain*, cui si fa più compiutamente cenno più avanti, che ha innovato radicalmente la posizione europea in tema di ruoli dei *search engines* nel sistema di protezione dati europeo, di stabilimento del *data controller*, di legge applicabile.

6.2. Google e la sentenza Google Spain

Il motore di ricerca americano si è impegnato nelle scorse settimane per adeguare i propri servizi alla sentenza della Corte di giustizia dell'Unione Europea pronunciata il 13 maggio 2014 riguardo alla causa «Google Spain SL, Google Inc. / Agencia de Protección de Datos (AEPD), Mario Costeja González» (Causa C-131/12)[24]. La corte europea ha infatti individuato nell'attività di elaborazione delle informazioni svolta dal motore di ricerca un trattamento autonomo, rispetto al quale la società riveste il ruolo di titolare (*data controller*). La sentenza europea costringe oggi Google (e conseguentemente altri operatori che si ritrovino nelle stesse condizioni individuate nella decisione dei giudici europei) ad adottare procedure per cancellare dai propri indici informazioni che possono continuare a essere presenti sul sito di originaria pubblicazione o su altri siti che provvedo-

[23] Si veda in tal senso la «Opinion 1/2008 on data protection issues related to search engines - WP 148» del 4 aprile 2008.
[24] In Appendice a questo Volume

no autonomamente alla loro ripubblicazione o citazione. A tal fine, per consentire agli utenti di richiedere la rimozione di talune notizie che li riguardino dai propri indici, Google ha messo a disposizione un modulo *online* nella pagina intitolata *«Search removal request under data protection law in Europe»* raggiungibile all'indirizzo seguente: https://support.google.com/legal/contact/lr_eudpa?product=websearch

Affinché una richiesta da parte di un utente venga presa in considerazione la società richiede che siano verificate tre condizioni: *a*) siano indicati gli indirizzi URL dei link da rimuovere; *b*) sia giustificata la pertinenza delle pagine a cui viene fatto riferimento nell'indice alla persona che presenta la richiesta; *c*) sia spiegato perché quei collegamenti nei risultati di ricerca siano irrilevanti, non aggiornati o in qualsiasi modo inappropriati. Dal punto di vista tecnico si evidenzia la particolare difficoltà di riconoscere in un richiedente la cancellazione un legittimo interessato cui facciano riferimento dati personali pubblicati *online* e indicizzati dal motore di ricerca. A questo proposito Google chiede che il richiedente trasmetta copia di un documento idoneo al riconoscimento di sé o dell'interessato per conto del quale viene presentata la richiesta.

L'adeguamento alla sentenza *Google Spain* pone problemi che vanno ben oltre il mero fatto tecnico, al punto che la società ha annunciato di essersi dotata di un *team* di esperti tecnici e legali di comprovata competenza, che dovrebbe esaminare le richieste di cancellazione da parte di cittadini europei. La decisione della Corte per un verso, relativo alle responsabilità del motore di ricerca, va molto oltre quanto disposto in passato dal Garante italiano con propri provvedimenti che indubbiamente tenevano conto dell'incertezza sulla giurisdizione, oggi risolta con la sentenza europea. Ma le decisioni del Garante riflettevano anche un bilanciamento di interessi tra il diritto all'informazione e alla manifestazione del pensiero, da una parte, e il diritto alla protezione dei dati personali, dall'altra, avendo effetti anche sui siti di originaria pubblicazione. Gli effetti della cancellazione ai sensi della sentenza *Google Spain* appaiono comunque, al momento, parziali: questa è limitata alle versioni del *search engine* rivolte a utenti europei, accessibili attraverso l'uso dei domini Google nazionali (google.it, google.fr...). Alcune informazioni potrebbero quindi essere rimosse o, meglio, non presentate tra i risultati di ricerca a un utente europeo, ma continuare a essere visualizzabili nella versione internazionale del servizio, accessibile nel dominio *google.com.*

La modalità di adempimento di Google appare quindi giocata sulla presentazione dei risultati, e non su radicali interventi sui *database* su cui

è costruito l'indice del motore di ricerca.

Il ristretto intervallo di tempo trascorso dalla decisione della Corte di giustizia UE non consente al momento di valutare pienamente la soluzione che Google (ed eventualmente altri *search engines*) ha messo in atto, poiché è presumibile che questa debba essere ancora affinata e migliorata, sia relativamente all'efficacia della rimozione sia nei criteri di riconoscimento degli interessati e della loro connessione a determinate notizie o dati personali pubblicati ma possibilmente riferibili a una pluralità di differenti soggetti.

Analogamente si presta a controversie la decisione di Google di notificare agli amministratori o *webmaster* le richieste di rimozione di *link* e notizie che si riferiscano a contenuti pubblicati su siti da questi controllati. La scelta è dichiaratamente volta a garantire un equilibrato esercizio dei diritti degli interessati, ai sensi della sentenza *Google Spain*, migliorando il processo decisionale interno alla Società e la resa per gli utenti, ma viene criticata per la possibile violazione della *privacy* di chi si rivolge a Google per la rimozione di notizie dall'indice di ricerca.

7. Conclusioni

Le tecniche di de-indicizzazione *ex post* o di prevenzione dell'indicizzazione sono al momento rudimentali, e tali permangono dalla metà degli anni '90. Strumenti più efficaci del *Robots Exclusion Protocol* sono quelli che prevedono l'utilizzo di fasi di interazione umana con i contenuti *online* prima della loro fruizione da parte degli utenti, in modo da disincentivare, se non escludere del tutto, l'azione di agenti automatizzati di ricerca.

Tuttavia tali accorgimenti non hanno una diffusa accettazione in quanto incidono negativamente sulla *esperienza d'uso* della rete, oggi caratterizzata da una diffusione di dispositivi mobili di ridotte dimensioni e dalla conseguente necessità di garantire una maggiore semplicità d'uso attraverso una semplificazione delle interfacce uomo-macchina.

Altro possibile approccio è quello basato sul paradigma del *web semantico* per la tipizzazione delle informazioni e la definizione di una disciplina del loro utilizzo, anche con riferimento agli aspetti di protezione dati. Tale approccio richiederebbe però un radicale mutamento delle tecniche di pubblicazione *online*, dei formati dei documenti e una conseguente *compliance* da parte dei *search engines.*

Tutto ciò lascia pensare che, qualora si intenda limitare la conoscenza di informazioni l'unico strumento davvero efficace sia la radicale misura della rinuncia alla loro pubblicazione. La presenza sul *web* di una notizia la candida comunque, nonostante le possibili azioni tecniche mitigatrici, a conoscere in misura più o meno rilevante l'indesiderata diffusione, nonostante interventi di legislatori e decisioni giudiziarie che hanno comunque efficacia limitatamente alla giurisdizione esercitabile, in un contesto di fenomeni tecnologici di tipo transnazionale.

Ci si interroga quindi se occorra rassegnarsi a un mondo tecnologico in cui il passato di ogni individuo possa riemergere con poche o nessuna possibilità di controllo, la sua vita privata possa essere registrata in una sorte di memoria permanente della rete sempre di più somigliante a quella di Funes el Memorioso, personaggio nato dalla fantasia di Borges e oggi quanto mai identificabile non in un individuo dalla memoria prodigiosa ma nell'insieme di strumenti di ricerca e sistemi di memorizzazione, *sharing* di dati o *social networks* disponibili *online* sulla rete. Segnali incoraggianti vengono dallo sviluppo di tecnologie di ricerca di tipo semantico che, oltre a essere più accurate ed efficaci, presuppongono tuttavia modifiche sostanziali alla modalità di presentazione e di elaborazione delle informazioni.

Il successo di qualsiasi iniziativa volta a concretizzare anche in prodotti tecnologici i principi della protezione dei dati dipenderà dalla capacità di affermarli e promuoverli tra la vasta platea costituita dalla comunità di utenti e sviluppatori della rete, in un'era in cui i valori della *privacy* e della protezione dei dati personali vanno incontro a un'altalena di percezioni e considerazioni anche da parte del pubblico, dei legislatori e, più in generale, dei *policy makers*, in un conflittuale rapporto tra libertà di espressione, diritto di conoscere da parte del pubblico e diritto dell'individuo di sottrarsi all'azione invasiva delle tecnologie della rete.

Abstract

This paper describes the fundamental choices available to programmers, software developers and web designers for limiting the spread of data published on a World Wide Web site. Features and limitations of different methods are discussed, together with a brief historical introduction to the notion of infor-

mation indexing related to Internet services, with a special focus on the Robot Exclusion Protocol adopted by web publishers around the Net in order to allow or deny a crawler agent the access to web contents.

The well-known Google Spain case decided by the European Court of Justice in May 2014 is also shortly debated for its implications related to the technical means required for the compliance.