Gli open data pubblici a supporto e validazione della ricerca

Diego GIORIO

Abstract

Into the information era, the immense wealth of data held in public offices, even if collected sometimes in a messy, redundant, uncoordinated manner, can be published and made available to all: citizens, scholars, other public entities and researchers. Demographic data, births, deaths with related causes, topographic surveys, museums and library catalogs, information on industrial and artisanal activities, traffic analysis... Just to mention a few examples that come to mind out from an almost infinite list. Data contained in public archives may be imperfect or incomplete, but is nevertheless official data which is used to determine state policy. Policy that is increasingly based on economic algorithms and statistics rather than on real and perceived social needs. Starting with available data, researchers are able to build on standard, verifiable and not easily falsifiable elements. Otherwise, whether intentionally or in good faith, conclusions and policies might be based on information which is incorrect and/or difficult to verify. Laws on the matter are already in force; however, the diffusion of open data is far away to take off. This is due to several reasons: from the shortage of time and reduced staff in public Italian offices to the poor attitude of employees; from software not ready to manage the task to the bad habit of printing and scanning tables, so it becomes impossible to read texts automatically. Moreover, in many offices there is the ingrained habit of not sharing information and data except on very rare occasions, as if publishing data would represent a loss in prestige and power. With appropriate information campaigns, and with the desirable turnover inside public administrations, it is nevertheless possible to overcome these problems. Certainly, there are costs associated with training, and what is now ironically known in all public offices as the "last-subsection syndrome", namely the financial invariance clause that closes most of the rules dedicated to PA innovation, does not help. However, the availability of courses on the internet can reach a wide audience at a very low cost, especially if they are centrally organized by the state and not entrusted to a plethora of private companies. Moreover, open data and transparency could be included as a mandatory part of any new law relating to PA innovation. A second question to address is not to underestimate anonymization: data must be made available in sufficiently detailed form to be useful and usable, but adequately aggregated to avoid de-anonymization. This is a huge risk and a rather slippery question in the big data era; many recent studies show that, in a large majority of

cases, apparently anonymous data can be related back to the owner by providing sufficient information and computing power. Such risk is significant in Italy, due to the specificity of Italian jurisdiction, comprising almost 8000 municipalities, often of very small sizes. However, this is a problem that can be easily overcome by moving aggregation to a higher level, such as counties or a group of municipalities, in order to include a sufficiently large numbers. In any event, assuming that open data is available and properly managed, this huge public asset can have positive effects on many types of research. First, researchers may draw from an open and complete set of basic data. In addition, it may be easier to verify the results, by a peer-review or other type of verification. Also, considering that data of public administration is not always correct and complete, a reverse check may also occur, correcting errors and anomalies on revealed discrepancies between research results and basic data, or aligning different databases when comparison produces contradictory results. Lets consider a few examples such as medical research, which is usually joined with demographic statistics, perhaps accompanied by data related to the ethnicity and/or profession of the subjects. Transport study, which can not be separated to the territory's orography, the distribution of the population, the typology and consistency of the productive apparatus. Or historical research, facilitated by the availability of maps and online archives. Above all, consider all this data being put together, and then linked into a wider network of information that allows extensive, perhaps original, correlations, by finding interconnections that would, at first sight, seem meaningless. Of course, any data obtained through public service must be verified, properly analyzed and appropriately located in a more general context. However, to find an entire knowledge base directly on the WEB, certainly facilitates the researcher's activity, avoid the frustration of ignored requests, uncooperative offices, data scattered in a thousand different formats, and perhaps printed on unreadable continuous forms. And consider, from the other side, public offices: seeing that publication of open data is not a sterile compliance with a useless law, and a waste of time, but represents a useful tool, will certainly be encouraged to publish good quality data on regular base. Of course, if such incentive would not be just moral but would be part of the evaluation, also in order to introduce really - meritocracy into PA, it would be a further boost. As a conclusion, the above idea represents a potential virtuous circle that is not easy to trigger, but, once powered up, it can only bring benefits to society as a whole.

Introduzione

Viviamo nell'era dell'informazione ed i dati stanno assumendo un valore economico importante, spesso superiore a quello di beni materiali una volta considerati essenziali; pensiamo ad esempio al cibo, che nelle società ricche e tecnologicamente evolute è dato spesso per scontato, oppure all'energia, percepita come talmente naturale che ci accorgiamo della sua importanza solo quando viene a mancare. Già agli albori della Rete Bill Gates, fondatore della Microsoft

e persona certamente visionaria, immaginava un mondo dove l'informazione è una delle ricchezze principali:

Immaginai conversazioni senza senso accanto alla macchinetta del caffè in un ufficio del futuro: "Quanta informazione possiedi?"; "La Svizzera è un grande paese grazie alla quantità di informazione che ha"; "Ho sentito che l'indice dell'informazione sta salendo".[1]

In questa era dell'informazione, l'immenso patrimonio di dati detenuti negli uffici pubblici può essere messo a disposizione di tutti: cittadini, studiosi, altre entità pubbliche e di ricerca. Le anagrafi raccolgono informazioni sui movimenti della popolazione e sulla sua composizione, lo stato civile cataloga nascite, morti con relative cause, matrimoni, unioni civili, cittadinanze, fornendo uno spaccato della società reale, magari non studiata sui libri di storia, ma non per questo meno importante o meno utile ai fini della ricerca. È sufficiente scorrere i vecchi registri per comprendere come sono cambiati la mentalità ed il modo di vivere: a fine ottocento sono comuni gli atti di morte a coppie, quello del piccolo Mario "di giorni zero ed ore due" e della mamma di 16 anni che lo ha seguito poco dopo. Si scopre che una donna di 32 anni si sposa "col consenso del padre". Si vedono i passaggi di epidemie e carestie, con tassi di mortalità che da un anno all'altro variano anche del 300%, mentre ai giorni nostri sono sostanzialmente costanti in funzione del numero di abitanti. Registri apparentemente aridi e senza scopo consentono invece di ricavare dati preziosi per comprendere l'evolvere della società.[2]

L'insieme dei dati detenuti dagli Enti Statali e locali, formato e mantenuto con risorse pubbliche, non deve rimanere relegato negli uffici e messo a disposizione solo saltuariamente, a discrezione dei funzionari, ma dev'essere reso fruibile da chiunque abbia un interesse ad utilizzarlo. Naturalmente evitando di divulgare dati personali, ovvero fornendo dati aperti resi sufficientemente anonimi e aggregati da non poter risalire nominativamente ai soggetti, il che, come vedremo, è meno immediato di quanto possa sembrare.

I dati pubblici

Enti locali e centrali, a partire dai Comuni per arrivare alle Istituzioni europee, raccolgono dati di tutti i tipi: rilievi topografici, cataloghi di musei e biblioteche, informazioni sulle attività industriali ed artigianali, flussi di traffico, dati sanitari, informazioni sull'assistenza sociale e sulla protezione dei minori, sui reati, sugli animali da compagnia e così via. Solo per citare i primi esempi che vengono in mente di una lista quasi infinita, così come il loro utilizzo è limitato solo dalla fantasia: in Danimarca esiste una lista dei bagni pubblici ricavata dai dati aperti, in Germania e UK ci sono siti che consentono di decidere la

[1] Bill Gates, La strada che porta al domani, Ed. Mondadori. 1994

[2] Spero mi si voglia perdonare l'auto-citazione, ma non ho trovato lavori equivalenti che ripercorrano la storia d'Italia attraverso i servizi demografici: Diego Giorio, Un lungo viaggio in compagnia della Rivista attraverso i servizi demografici, SEPEL Editrice, Ottobre 2011 http://issuu.com/statocivile/docs/speciale110?mode=window&backgroundColor=%23222222

zona dove abitare in funzione dei tempi di percorrenza dalla casa all'ufficio, ottenuti dal confronto di dati catastali e flussi di traffico; nella stessa Danimarca si può pianificare la ristrutturazione della casa per il risparmio energetico attraverso un sito che incrocia dati catastali, incentivi governativi, e dati delle imprese locali.[3] I dati contenuti negli archivi pubblici possono essere imperfetti, incompleti, ma sono comunque dati ufficiali, sui quali si fondano le scelte dello Stato, le quali oggi si basano sempre più su algoritmi economici fondati sulla statistica[4] e sempre meno sulle necessità sociali[5]. Nell'articolo citato in nota, Supinot osserva che lo Stato, nato ed inteso come una "superpersona" che si occupa del benessere dei cittadini, è oggi superato da una gestione politica basata essenzialmente sull'economia, a sua volta fondata e decisa dagli algoritmi anziché dalle necessità umane. A titolo provocatorio immagina persino la fine dei governi, almeno come li intendiamo oggi, dato che il loro potere decisionale è subordinato alle necessità dell'economia - oggi assolutamente sovranazionale - più che orientato alle necessità sociali. Sullo stesso filone il testo di Rahnema - Robert, che però pone l'accento sui rischi di questa impostazione, dato che i poveri sono molti e le spinte sociali da essi create, pur non essendo ricomprese negli algoritmi di borsa, non potranno che riportare il focus dell'azione politica sui problemi sociali. Oggi è comunque un dato di fatto - non è questa la sede per discutere se sia giusto o sbagliato - che la statistica e gli algoritmi sono sempre più pervasivi; pensiamo ad esempio alle borse, dove le decisioni sono sempre più affidate ai software e sempre meno agli operatori umani[6], oppure ai prestiti, che attraverso software di gestione diretta possono essere gestiti senza l'intermediazione delle banche[7], o ancora alle monete virtuali, che prosperano senza uno Stato che le garantisca o una zecca che provveda a coniarle[8]. Questi algoritmi, per poter funzionare efficacemente, necessitano di essere continuamente alimentati con dati quanto più possibile precisi, estesi ed aggiornati. I dati pubblici, anche se spesso raccolti in modo disordinato, ridondante, non coordinato, rappresentano un insieme di informazioni fondamentali che costituisce una base di partenza ufficiale per qualunque ricerca e progetto. Certo, i dati forniti dagli Enti non sono perfetti, ed in alcuni casi proprio la loro natura ufficiale li rende oggetto di mistificazione; pensiamo ad esempio ai dati ufficiali sugli immigrati, che, per definizione, non censiscono i clandestini, oppure i dati dell'Agenzia delle Entrate, che non possono includere l'economia sommersa. A volte i dati ufficiali possono essere falsati da alcuni fattori economici, quali i benefici fiscali concessi per le prime case, che spingono i cittadini a dichiarare

[3] Open data: l'innovazione della pubblica amministrazione a servizio della nuova imprenditorialità nell'ICT - Centro Studi Consiglio Nazionale Ingegneri - Roma, 23 gennaio 2013.

[4] cfr Alain Supiot, La gouvernance par les nombres Broché 2015 e Jacky Fayolle, La gouvernance par les nombres est-elle la fin de l'histoire de la statistique? - http://www.luxstat.lu/telechargements/JFayolle_ConfLux.pdf

[5] cfr Majid Rahnema - Jean Robert, La puissance des pauvres - Poche, 2012

[6] http://www.huffingtonpost.it/2017/02/01/algoritmi-borsa_n_14545156.html

[7] http://www.lastampa.it/2017/08/29/economia/da-wall-street-a-singapore-il-boom-dei-prestiti-online-i7GNSwlYjeJOnA0DI7rzaL/pagina.html

[8] http://www.diritto24.ilsole24ore.com/art/avvocatoAffari/mercatiImpresa/2017-07-14/partnership-professionisti-e-imprese-reti-miste-113003.php#

residenze fittizie per godere delle agevolazioni. Tuttavia, pur con la consapevolezza di alcuni limiti intrinseci e di errori possibili, basandosi sui dati aperti ufficiali i ricercatori partono da un elemento standard, pubblicamente verificabile e non facilmente falsificabile; diversamente, sia in buona fede che per dolo, si potrebbe costruire un castello di tesi e conclusioni poggiato su fondamenta errate o difficilmente verificabili.

Normativa principale

Legge 7 agosto 1990, n. 241 (Nuove norme in materia di procedimento amministrativo e di diritto di accesso ai documenti amministrativi). Anche se non direttamente legata agli open data, non si può non ricordare questa pietra miliare del rinnovamento amministrativo, che ha segnato la svolta da una PA autarchica ed autoritativa ad una PA partecipativa e trasparente. Il cittadino non è più colui che porge rispettosa domanda e poi aspetta paziente che l'amministrazione risponda, ma ha diritto di partecipare al procedimento, di prendere visione degli atti, di presentare osservazioni. Questa legge non prevede ancora la pubblicazione periodica di dati specifici, consente l'accesso solo se motivato da un legittimo interesse e non finalizzato ad un controllo della PA, ma ha rappresentato un cambiamento epocale, peraltro ancora oggi non del tutto recepito da tutti gli uffici.

Decreto Legislativo 7 marzo 2005, n. 82 (Codice dell'Amministrazione Digitale). Tutto il CAD è permeato dalla volontà di favorire l'interscambio dei dati ed i formati aperti. In particolare ricordiamo il capo V (artt. 50 - 57 bis) e l'art. 68 c 3:

3. Agli effetti del ((presente Codice)) legislativo si intende per:

a) formato dei dati di tipo aperto, un formato di dati reso pubblico, documentato esaustivamente e neutro rispetto agli strumenti tecnologici necessari per la fruizione dei dati stessi;

b) dati di tipo aperto, i dati che presentano le seguenti caratteristiche:

1) sono disponibili secondo i termini di una licenza che ne permetta l'utilizzo da parte di chiunque, anche per finalità commerciali, in formato disaggregato;

2) sono accessibili attraverso le tecnologie dell'informazione e della comunicazione, ivi comprese le reti telematiche pubbliche e private, in formati aperti ai sensi della lettera a), sono adatti all'utilizzo automatico da parte di programmi per elaboratori e sono provvisti dei relativi metadati;

3) sono resi disponibili gratuitamente attraverso le tecnologie dell'informazione e della comunicazione, ivi comprese le reti telematiche pubbliche e private, oppure sono resi disponibili ai costi marginali sostenuti per la loro riproduzione e divulgazione.

Decreto Legislativo 14 marzo 2013, n. 33 (diritto di accesso civico e gli obblighi di pubblicità, trasparenza e diffusione di informazioni da parte delle pubbliche amministrazioni). Questo decreto è stato il primo intervento dello Stato centrale che ha obbligato i Comuni a strutturare i loro siti, almeno nella sezione dedicata alla trasparenza, secondo uno schema prefissato ed uniforme

in tutta Italia. Questo approccio non è finalizzato solamente a consentire una ricerca più agevole da parte umana, ma soprattutto a rendere i siti *machine readable*. Che poi molte volte vengano caricate delle scansioni, vanificando l'intento, è altro discorso, ma quantomeno la struttura portante del sistema è stata predisposta. Inoltre una macchina può quantomeno verificare automaticamente la presenza o meno del materiale previsto dalla norma, anche senza entrare nel merito della sua completezza o attualità. Il D.Lgs 33 ha inoltre ribaltato il concetto della L. 241/90, ovvero già al primo articolo statuisce che l'obiettivo del decreto è di consentire un controllo dell'operato della PA; di conseguenza l'accesso agli atti - fermo restando il limite della riservatezza e della protezione dei dati personali - non è più limitato a chi può dimostrare un legittimo interesse, ma è consentito senza una motivazione specifica. La pubblicazione dei dati è però volta soprattutto a favorire la trasparenza nella contabilità, nell'assegnazione degli incarichi e degli appalti, nella gestione del patrimonio pubblico, ovvero è complessivamente finalizzata al contrasto della corruzione, anche se l'art. 5-ter è espressamente dedicato alle finalità di ricerca.

Direttiva 2013/37/UE del Parlamento Europeo e del Consiglio del 26 giugno 2013 e suo recepimento **Decreto Legislativo 24 gennaio 2006, n. 36** (Attuazione della direttiva 2003/98/CE relativa al riutilizzo di documenti nel settore pubblico). Questa direttiva, che rinnova una precedente direttiva del 1998, si apre col punto 1 dei considerando affermando che *3. [i] documenti prodotti dagli enti pubblici degli Stati membri costituiscono un ampio bacino di risorse diversificato e prezioso in grado di favorire l'economia della conoscenza.* La Direttiva si sviluppa poi affermando la necessità di rendere disponibili documenti in formato aperto in forma gratuita o con costi marginali di mera riproduzione. Specifica inoltre che i "documenti" debbano essere leggibili "meccanicamente". Queste due scelte lessicali, effettuate peraltro nel 2006, non negli anni '50, sembrano limitare l'efficacia della Direttiva, dato che al termine "documento" normalmente si associa un concetto un po' più ristretto rispetto agli open data come li concepiamo oggi (per quanto una tabella di dati sia effettivamente un documento) e soprattutto il termine "meccanico" fa più pensare alle schede perforate che ai moderni sistemi informatici. Al di là della discutibile scelta terminologica, comunque, il senso e gli scopi della norma sono abbastanza chiari e condivisibili e dovrebbero quindi essere evitati i file in formato immagine, utilizzando solamente formati considerati aperti dalle linee guida dell'AgID[9].

Decreto Legislativo 30 giugno 2003, n. 196 Codice in materia di protezione dei dati personali e **Regolamento (UE) 2016/679 del Parlamento europeo e del Consiglio, del 27 aprile 2016**. Il nostro codice privacy sta per essere superato dal nuovo Regolamento Europeo, che entrerà in vigore il 25 maggio 2018. Entrambe le norme, comunque, impediscono di pubblicare in modo incontrollato dati personali, per cui gli open data devono essere resi anonimi in modo da prevenire sia una correlazione diretta con le persone, sia una connessione indiretta attraverso algoritmi e comparazioni. Entrambi i testi

[9] http://egov.formez.it/sites/all/files/open_data_-_formati_aperti.pdf

prevedono il trattamento di dati ai fini di ricerca e di analisi statistica.

Ulteriori riferimenti:

Linee guida AgID (ed dicembre 2016)

http://www.dati.gov.it/sites/default/files/LG2016_0.pdf[10]

Queste indicazioni dell'AgID si propongono di superare i problemi della digitalizzazione della PA, i cui progressi sono ancora troppo spesso confinati a iniziative virtuose isolate di alcune amministrazioni. Parlano dunque di metadati, di formati aperti, di modelli e processi. A fronte di un lavoro tecnicamente ben fatto e completo, è però mancata la sua diffusione capillare, per cui resta un documento che, di fatto, è circolato solo fra i pochi appassionati della materia.

Statuto Internazionale degli Open Data Pur trattandosi di un'iniziativa privata, priva di valore cogente a livello italiano o europeo, raccoglie una serie di indicazioni e di dichiarazioni di principio riguardo gli open data; la sua validità è garantita dal fatto che la stessa AgID ha richiamato questa iniziativa nelle linee guida del 2016. Anche questo sito, però, sembra porre l'accento più sul contrasto alla corruzione che sulla diffusione della conoscenza.

Linee guida europee su licenze standard e dataset raccomandati e tariffe da applicare nel riutilizzo di dati pubblici.

http://ec.europa.eu/newsroom/dae/document.cfm?action=display&doc_id=6421

Pubblicate nel 2014, possono essere leggermente superate dal punto di vista tecnico, ma mantengono la loro validità nello stimolare l'apertura delle banche dati pubbliche, in quanto *[o]pening up public sector information (PSI) for reuse brings major socio economic benefits. Data generated by the public sector can be used as raw material for innovative value-added services and products which boost the economy by creating new jobs and encouraging investment in data-driven sectors.*

Guardare avanti

Se tutte queste normative ed indicazioni fossero scrupolosamente applicate, già in Europa ed in Italia ci sarebbero molti più dati liberamente fruibili; tuttavia lo sforzo principale dovrebbe essere indirizzato a creare un atteggiamento mentale aperto, prima ancora di costruire un insieme di norme che finirebbero comunque col dimenticare qualcosa o col non essere applicate. Favorire una mentalità, includere la menzione degli open data in ogni innovazione normativa che riguardi la PA, spiegare gli open data in occasione di corsi di formazione, sviluppare software orientati ai dati aperti può favorire la formazione di una forma mentis diffusa, che porta ad una fruibilità di dati molto maggiore di quanto possa fare un mero vincolo cogente, uno sterile adempimento, magari poco controllato nella sua implementazione. Inoltre occorrerebbe cambiare approccio, dato che gli open data non dovrebbero essere finalizzati principalmente alla prevenzione della corruzione, per quanto lodevole ed importante possa essere questo obiettivo, quanto piuttosto a rendere disponibile il patrimonio informativo, indipendentemente da una finalità specifica.

[10] Del 2011, quindi un po' sorpassato, ma ancora ottimo per diversi argomenti il Vademecum sugli open data: http://trasparenza.formez.it/sites/all/files/VademecumOpenData_0.pdf

Le problematiche

Se le norme già ci sono, la diffusione dei dati stenta però a decollare per tanti motivi, dalla penuria di tempo e di personale negli uffici alla scarsa attitudine mentale degli impiegati, dagli applicativi software non ancora adeguati alla pessima abitudine di stampare e scansionare le tabelle, rendendole impossibili da leggere in modo automatico. Purtroppo la *forma mentis* che permane in molti uffici è quella di tenere sotto chiave qualunque dato ed informazione, facendo cadere dall'alto ogni erogazione graziosamente concessa, quasi che a pubblicare un dato si perda di prestigio e di potere. Con opportune campagne di informazione - e con l'auspicabile svecchiamento della PA - non si tratta tuttavia di un problema insormontabile. Certo, la formazione non è mai completamente gratuita e quella che oramai è ironicamente conosciuta in tutti gli uffici pubblici come "sindrome dell'ultimo comma", ovvero la clausola di invarianza finanziaria che chiude la maggior parte delle norme relative all'innovazione nella PA, assolutamente non aiuta. Tuttavia la diffusione dei videocorsi consente oggi di raggiungere una vasta platea di persone a costi molto contenuti, soprattutto se sono organizzati centralmente dallo Stato e non affidati ad una pletora di diverse società private; inoltre gli open data e la trasparenza potrebbero essere parte integrante di ogni innovazione normativa, in qualunque settore. Ogni PA possiede e produce dati che possono essere interessanti per qualcuno. Un'università, ad esempio, fa un grosso servizio nel mettere on-line le tesi, ma già i dati grezzi sul numero di studenti divisi per corsi, numero di tesi discusse, numero di anni intercorsi fra l'iscrizione e la laurea possono essere utili per analizzare l'evoluzione della cultura e della società. Una biblioteca farebbe un'opera meritoria a scansionare tutti i volumi, quantomeno quelli non più coperti dal diritto d'autore, e renderli fruibili on line. Ma anche se l'impresa fosse troppo costosa, si potrebbe almeno pubblicare il catalogo e fornire i dati sul numero di tessere, numero di accessi, categorie più richieste: in questo modo si permetterebbe di sapere se un libro disponibile è consultabile oppure no e si consentirebbe di analizzare le abitudini di lettura della popolazione. I Comuni raccolgono dati sui movimenti della popolazione, contornati da una serie di dati accessori che comprendono la cittadinanza, il colore dei capelli, la composizione delle famiglie, il grado di istruzione, l'altezza... Gestiscono lo stato civile, dunque nati, morti, matrimoni, unioni civili, acquisto della cittadinanza. Controllano il traffico, l'edilizia privata e industriale, i contributi erogati alle famiglie in difficoltà, la raccolta dei rifiuti. ASL, Questure, INAIL, INPS raccolgono dati a profusione. Pensare di elencare con atto avente forza di legge ogni singola tabella di numeri che potrebbe essere di qualche interesse ha poco senso, si tratterebbe di una battaglia persa in partenza; meglio far sì che ogni Ente si senta responsabilizzato a pubblicare i propri dati, magari puntando sull'automazione, quindi prevedendo che i software includano la funzione di elaborazione e pubblicazione dei dati che servono, e soprattutto che lo facciano in modo automatico, almeno generando una proposta che compaia automaticamente con cadenza periodica e che l'operatore potrà confermare o meno. Il programma d'anagrafe dei Comuni dove lavoro, ad esempio, mette a disposizione una funzione che estrae alcuni dati anagrafici

e li pubblica automaticamente sul WEB. Occorrono un paio di click e pochi minuti, non è un problema di tempo o di complessità tecnica. Il problema è che si tratta di una funzione che ho scoperto per caso, perché nessuno si è preoccupato di fare formazione, e che viene attivata manualmente sulla base della buona volontà e della memoria dell'impiegato. Se almeno una volta l'anno uscisse un pop-up di invito all'avvio della routine, probabilmente ci sarebbero molti più dati anagrafici disponibili. Se poi il ragionamento venisse esteso a tutte le Pubbliche Amministrazioni, si renderebbe disponibile una quantità inimmaginabile di dati, consentendo correlazioni di ogni tipo, anche originali ed inaspettate.

L'anonimizzazione

Un secondo problema da non sottovalutare nella pubblicazione degli open data è l'anonimizzazione dei dati, che devono essere resi disponibili in forma sufficientemente dettagliata da essere utili e fruibili, ma abbastanza aggregata da non poter risalire all'interessato neppure per via indiretta, questione piuttosto scivolosa nell'era dei *big data*, dato che molti studi recenti[11] dimostrano come, in una larga maggioranza di casi, dati apparentemente anonimi possono essere resi nominativi disponendo di sufficienti informazioni e potenza di calcolo. Risorse che certo non mancano a quei colossi del WEB che hanno fatto della profilazione degli utenti e della conseguente pubblicità mirata il loro *core business*: grandi Società come Google, Yahoo, Facebook, Amazon possono reclutare le migliori menti del pianeta fornendo loro risorse praticamente illimitate. Estremizzando e semplificando, se un'ASL comunica che c'è un cinese con l'AIDS ed in anagrafe è registrato un solo cinese, chiaramente un dato molto sensibile diviene *de facto* nominativo. Ma possono essere attuate analisi più fini ed ingegnose, se si dispone di dati sufficienti: voglio ad esempio inviare un opuscolo ai bambini del primo anno di elementari; monitorando sui social il gruppo "il primo giorno di scuola" osservo che un padre, oltre alla scuola del figlio, è molto interessato alla pesca sportiva, poi scopro che in quel Comune è attivo un abbonamento alla rivista "Pesca sportiva oggi" e posso immaginare di aver trovato un primo indirizzo! E' questione di disponibilità di dati, di intelligenza artificiale e di potenza di calcolo, non di limitazioni intrinseche alla possibilità di analisi, dunque gli open data, prima di essere pubblicati, devono essere accuratamente valutati ed opportunamente aggregati. Moltissimi articoli[12] riportano casi in cui da dati

[11] Ex multis, vedasi http://ieeexplore.ieee.org/abstract/document/4531148/?reload=true

[12] Tra i tanti si segnala:

- The Security of our Secrets: A History of Privacy and Confidentiality in Law and Statistical Practice - https://papers.ssrn.com/sol3/papers.cfm?abstract_id=886165
- Anonymisation: managing data protection risk code of practice - https://ico.org.uk/media/1061/anonymisation-code.pdf
- De-Identification of Personal Information, Simson L. Garfinkel - http://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf
- De-anonymizing Web Browsing Data with Social Networks, Jessica Su - Ansh Shukla - Sharad Goel - Arvind Narayanan - http://randomwalker.info/publications/browsing-history-deanonymization.pdf

sanitari resi anonimi si è risaliti facilmente all'interessato in modo nominativo, oppure da pochi dati apparentemente anonimi di Netflix si è risaliti in modo diretto all'autore del post; in generale, la de-anonimizzazione dei dati sta diventando un problema, perché pubblicare dati anonimi con finalità statistica e poi scoprire che vengono usati per profilare individualmente le persone è quantomeno imbarazzante, oltre che costituire un problema morale e giuridico. Proprio per questo, però, anche per l'anonimizzazione si stanno effettuando studi e si stanno sviluppando metodologie specifiche, su base matematica[13]. Nel caso citato dei dati sanitari, ad esempio, l'anonimizzazione era stata effettuata per semplice soppressione di alcuni campi del data base, ovvero cognome, nome e Social Security Number, l'equivalente del nostro codice fiscale, lasciando però codice postale, data di nascita completa, sesso e razza. E' stato calcolato che con questi dati si può identificare almeno il 63% della popolazione americana, mentre altri studi riportano valori ancora più alti, dell'ordine dell'87%[14]. In realtà, ai fini medici, l'anno di nascita - o una fascia di età - è più che sufficiente per effettuare delle valutazioni significative, tranne nel caso dei piccolissimi, per cui eliminare giorno e mese, oppure raggruppare l'anno di nascita all'interno di un lustro, permette la piena utilizzazione dei dati senza consentire di risalire al soggetto. Senza pretesa di valore staticamente significativo, ho provato a verificare nei Comuni che seguo ed a chiedere ad alcuni Comuni con cui ho dei contatti di verificare se abbiano fra i residenti qualcuno con la mia stessa data di nascita o con quella di mia moglie. In quattro Comuni di 9000, 19.000, 17.000 e 50.000 abitanti non ne risulta neppure uno. Su 9000 abitanti c'è una persona nata lo stesso giorno di mia madre, ma si tratta di un maschio, quindi non c'è possibilità di confusione. Purtroppo la peculiarità del territorio italiano, diviso in quasi 8000 Comuni, spesso di dimensioni molto piccole, non favorisce certo la possibilità di pubblicare dati statistici puntuali ma non de-anonimizzabili, problema superabile abbastanza facilmente spostando l'aggregazione ad un livello superiore, come una Provincia oppure un gruppo di Comuni che comprenda una popolazione numericamente sufficiente. In una grande metropoli già un rione può contenere abbastanza abitanti da poter fornire dati in forma anonima. Nel mio Comune ci sono vie con due abitazioni e tre famiglie, con tre abitazioni e tre famiglie, per cui un dato statistico a livello toponomastico diventerebbe di fatto nominativo. Per contro, un raggruppamento di dati eccessivamente centralizzato può perdere di significato: un'indagine epidemiologica, ad esempio, dev'essere localizzata, anche per poter prevenire la diffusione di un focolaio. I Comuni italiani, assieme agli altri Enti locali, consentono raccolte dati puntu-

[13] Vedasi, ad esempio:

- Robust De-anonymization of Large Sparse Datasets, Arvind Narayanan - Vitaly Shmatikov - http://ieeexplore.ieee.org/abstract/document/4531148/?reload=true
- A Systematic Review of Re-Identification Attacks on Health Data, El Emam K, Jonker E, Arbuckle L, Malin B (2011) - https://doi.org/10.1371/journal.pone.0028071

[14] Broken Promises Of Privacy: Responding To The Surprising Failure Of Anonymization, Paul Ohm - https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1450006

ali ed estremamente focalizzate sul territorio, raccogliendo dati demografici, del piccolo commercio ed artigianato, dell'edilizia, della salute, dell'istruzione, dei servizi sociali in modo più dettagliato ed aggiornato di quanto possa fare la Regione o lo Stato. Normalmente per un utilizzatore è decisamente più facile raggruppare i dati ad un livello superiore piuttosto che separare dati troppo aggregati, per cui è importante raggiungere un buon equilibro fra la granularità dell'informazione e l'impossibilità di risalire alle persone fisiche.

I beneficiari

Se i palinsesti proposti dalle TV o i titoli dei numerosi giornali di gossip potrebbero non far guardare con ottimismo al livello culturale medio della popolazione, non bisogna dimenticare che sono disponibili programmi e riviste di divulgazione scientifica apprezzati e seguiti da molte persone; quella che in inglese viene definita *citizen science* è un fenomeno diffuso in tutto il mondo ed ovunque non mancano dilettanti che studiano vari argomenti o collaborano con i professionisti in qualità di paleontologi, archeologi, astronomi o quant'altro[15].

A volte emergono anche soggetti inaspettati, come la bambina di 12 anni che, preoccupata di non poter portare con sé il proprio peluche preferito in caso di intervento chirurgico, ha condotto uno studio sulla sterilizzazione dei giocattoli in vista della loro introduzione in sala operatoria come supporto psicologico per i più piccoli[16]. Certo, il fatto che la madre sia un chirurgo e che fosse in contatto con uno scienziato che lavorava su un problema simile avrà certamente aiutato a rendere lo studio professionale e pubblicabile su una rivista scientifica, ma la base di partenza non è venuta da una scienziata affermata. Non è quindi da considerarsi straordinario o insolito che un dilettante possa effettuare uno studio completo, o quantomeno essere di stimolo e suggerimento ai professionisti con domande, osservazioni e studi di livello anche elevato, pur se a volte da rifinire o approfondire.

Peraltro, se un professionista ha normalmente a disposizione maggiori strumenti tecnici e culturali, è anche condizionato dal diktat *"publish or perish"*, per cui se non pubblica non riesce ad ottenere finanziamenti e proseguire nella carriera, mentre un semplice appassionato può prendersi tutto il tempo di cui necessita per curare e completare il lavoro. Tra lo scienziato professionista che si dedica solamente alla ricerca ed il dilettante che studia per passione nel tempo libero vi è comunque un'infinità di sfumature: molti medici, ingegneri, avvocati, professionisti in generale, oltre a svolgere il lavoro quotidiano, approfondiscono le loro materie, scrivono articoli e libri, partecipano a convegni in qualità di relatori. Se la disponibilità di dati di qualità facilmente usufruibili è comodo per tutti, chiaramente saranno proprio le categorie che fanno ricerca per passione a

[15] Due esempi fra le migliaia possibili: l'Associazione degli Amici del Museo Egizio di Torino - http://acme-museoegizio.it/ oppure il progetto Break Through Initiative, che raccoglie oltre nove milioni di volontari che analizzano i segnali provenienti dallo spazio - https://breakthroughinitiatives.org/

[16] http://salute.ilgiornale.it/news/22682/-gaby-dodicenne-peluche-bambini/1.html

beneficiare maggiormente dei dati aperti ed essere incentivate a proseguire: se un professionista, soprattutto se appoggiato da una grande organizzazione, in qualche modo finisce sempre con l'ottenere i dati che gli occorrono, la difficoltà e la frustrazione di accedere agli uffici per ottenere documentazione può scoraggiare chi non ha una forte motivazione, magari anche economica. Inoltre il fatto che alcuni dati siano pubblici rende molto più semplice la revisione dei lavori e la validazione degli stessi: la valutazione di un articolo prima di mandarlo in stampa, sia ad opera di un comitato scientifico o per *peer review*, molto spesso è resa difficoltosa dal fatto che i dati di partenza e quelli elaborati nel corso della ricerca non sono facilmente verificabili, per cui o ci si basa sulla fiducia e si valuta esclusivamente il metodo, oppure si devono impegnare ingenti risorse per ripercorrere il cammino di raccolta e analisi a partire dai dati di base fino alle conclusioni. Gli open data rappresentano quindi tanto una garanzia rispetto alla qualità dell'informazione, quanto un risparmio di tempo e risorse. Il che può anche giustificare la modesta spesa per la formazione del personale della PA e la pubblicazione delle tabelle che la cultura degli open data comporta.

I privati

Se i dati detenuti dagli Enti Pubblici sono raccolti e gestiti con i soldi della collettività, e dovrebbe essere quindi naturale che rappresentino un patrimonio fruibile da tutti, anche molte Società private possiedono informazioni che potrebbero essere utili per la ricerca, se rese disponibili sotto forma di open data. Sempre con la dovuta attenzione alla privacy degli interessati ed ai segreti industriali delle Società stesse, pensiamo a quanti dati vengono raccolti da banche, assicurazioni, supermercati, catene di abbigliamento ed altre realtà private, oppure da Società dedicate a pubblici servizi come poste, distributori di energia, servizi telefonici, consorzi di raccolta rifiuti. Rendere fruibili sotto forma di open data alcuni dei dati raccolti dai privati aiuterebbe a completare il quadro e fornirebbe dati utili e curiosi. Ad esempio Yandex, il motore di ricerca più famoso in Russia, ha fornito le statistiche riguardo i siti più visitati dai moscoviti sulla metropolitana[17], rivelando che al mattino si cercano siti di preghiera, nella parte centrale della giornata indicazioni riguardo uffici, musei e centri commerciali, all'uscita dal lavoro ricette e cibi da asporto, mentre la sera è dedicata allo sport ed alle informazioni sui trasporti notturni quando la metro sta per chiudere. Sarebbe certamente interessante osservare se in altre parti del mondo la navigazione in metro rispetta gli stessi canoni, oppure rifare l'analisi fra qualche anno e verificare se i comportamenti sono rimasti costanti o si sono modificati, oppure confrontare i dati riferiti alla metro rispetto ad altri luoghi pubblici o privati. Certo, un privato non raccoglie i dati per mezzo di fondi pubblici e pertanto può non essere equo imporre un adempimento di pubblicazione, tuttavia tra tante tasse che gravano sulle aziende, questa potrebbe essere una delle meno gravose; già in diversi casi sussistono obblighi di pubblicazione per le imprese, pensiamo ad esempio all'editoria, laddove chi riceve finanziamenti

[17] http://www.orthochristian.com/105682.html

statali deve pubblicare sulla testata una sintesi del bilancio[18], oppure alle Società che devono rendere pubblici lo statuto, i dati sulla composizione del CDA, il bilancio stesso[19]. Ma la pubblicazione di open data, più che un obbligo normativo, potrebbe soprattutto essere un fattore di qualità ed una dimostrazione di serietà, ovvero potrebbe essere una forma di pubblicità, perseguita dunque spontaneamente dalle imprese. Di nuovo, più che adempimenti normativi, si tratta di stabilire una mentalità orientata agli open data. D'altra parte - e questo vale per il pubblico come per il privato - è molto più facile imporre e far osservare una regola quando questa è culturalmente accettata dalla società piuttosto che imposta dall'alto non comprendendone il senso e l'importanza. Ovvero *"[una norma] richiede, oltre all'aspetto istituzionale fondativo, anche un'adesione da parte dei singoli, i quali devono effettivamente osservare quelle regole (principio di effettività), o almeno riconoscerle legittime"*[20].

Dove trovare gli open data

Il bello del WEB è che le barriere geografiche non esistono, per cui si possono utilizzare dati riferiti al Canada o ad un paesino perso nel wild australiano. Se i dati pubblicati su siti locali richiedono una ricerca specifica, non mancano comunque i tentativi di radunare gli open data, al fine di favorirne la ricerca e l'utilizzazione. Il primo e più famoso, che ha fatto da modello per tutto il mondo, è probabilmente data.gov (https://www.data.gov/), nato nel 2009 per raccogliere tutti gli open data degli Enti statunitensi (e si trova anche qualche informazione straniera). Tra gli assunti alla base dell'iniziativa segnaliamo il principio che i dati devono essere pubblicati in un formato aperto *"indipendente dalla piattaforma, machine-readable, e reso disponibile al pubblico senza restrizioni che ne impediscano il riuso".* La Gran Bretagna ha seguito poco dopo con data.gov.uk - sito peraltro fortemente sostenuto da Tim Baners Lee, inventore del WEB - quindi Australia con data.gov.au, Canada con data.gc.ca, Norvegia con data.norge.no, Francia, India, ...
In Italia il sito di riferimento è dati.gov.it, ma, in accordo con lo spirito un po' indipendente degli italiani, riflesso naturalmente nei suoi Enti locali, si avverte più che in altri Paesi il problema della dispersione dei dati in molti siti locali.
Anche l'Europa ha un suo sito dedicato agli open data:
http://data.europa.eu/euodp/it/home/.
E neppure può mancare un indice generale, infatti il sito:
https://www.data.gov/open-gov/ raccoglie tutti i portali governativi del mondo che pubblicano open data.

[18] Legge 5 agosto 1981, n. 416 Disciplina delle imprese editrici e provvidenze per l'editoria.
[19] Articolo 2478 bis Codice Civile.
[20] Compendio di diritto pubblico, Roberto Carlo DELCONTE, SEPEL Editrice, 2012.

Qualità e tipologie

Abbiamo detto che troppo spesso gli open data vengono pubblicati in formati non facilmente leggibili, quando non in formato immagine ottenuto per scansione. Non si tratta di un problema solo italiano, se Tim Berners-Lee, attualmente direttore del World Wide Web Consortium, ha sentito l'esigenza di studiare un sistema di punteggi che classifica la fruibilità degli open data[21]:

*Dato non strutturato e codificato in formato proprietario (esempi: un file pdf; un'immagine jpeg);

** Dato strutturato, ma codificato in formato proprietario (quindi abbastanza facile da poter essere elaborato da un'applicazione informatica);

*** Dato strutturato in un formato non proprietario (per esempio, il formato CSV, che può essere aperto da qualsiasi software);

****Dati strutturati e codificati in formato non proprietario e dotati di un identificativo unico di risorsa (URI). Un esempio è lo standard RDF: applica al dato un significato condiviso ("quel dato ha lo stesso significato in qualsiasi lingua, per qualsiasi Paese");

*****Dati aperti collegati ad altri insiemi di dati aperti (Linked data).

In questa scala, un dato può considerarsi aperto se ha almeno tre stellette.

Una classificazione in base alla tipologia di open data - per quanto le classificazioni possano essere imperfette o limitative - può essere quella proposta dall'Open Knowledge Foundation[22]:

- **Geodati**: dati utilizzati per realizzare mappe, per esempio la localizzazione di strade e palazzi, la topografia, la visualizzazione dei confini, la georeferenziazione di esercizi commerciali...
- **Cultura**: dati riferiti a opere e prodotti culturali (per esempio: titoli, autori ecc.) e generalmente conservati da biblioteche, gallerie, archivi, musei.
- **Scienze**: dati prodotti come parte della ricerca scientifica, dall'astronomia alla zoologia;
- **Economia & Finanza**: dati relativi ai conti pubblici (entrate e spese), informazioni sui mercati finanziari (titoli, azioni, obbligazioni ecc.)
- **Statistica**: dati prodotti da uffici e servizi statistici, indicatori sociali, economici, demografici...
- **Meteo**: i vari tipi di dati utilizzati per comprendere e prevedere il meteo e il clima;
- **Ambiente e Salute**: informazioni relative all'ambiente (presenza e livello di fattori inquinanti, qualità delle acque, rifiuti.), ai tassi e cause di mortalità, all'incidenza di malattie in determinate zone...
- **Trasporti**: orari, percorsi, statistiche sui tempi di percorrenza...

[21] http://5stardata.info/en/
[22] https://okfn.org/

Le opportunità

Volendo assumere che gli open data siano disponibili e correttamente gestiti, si ritiene che questo enorme patrimonio pubblico possa avere effetti positivi su molte ricerche, tanto nel fornire una base aperta e completa di dati di partenza, quanto agevolando la successiva verifica dei risultati da parte di chi deve controllare la validità del metodo e delle conclusioni, ma anche la congruità dei dati di partenza. Inoltre, poiché non sempre i dati in possesso della PA sono corretti e completi, potrebbe verificarsi anche il percorso inverso, ovvero la correzione di errori ed anomalie a seguito di rilevate discrepanze fra i risultati di una ricerca ed i dati da cui si è partiti, oppure fra diverse banche dati, il cui confronto ha prodotto risultati contraddittori. Pensiamo ad una qualunque ricerca medica, che difficilmente potrà prescindere da statistiche demografiche, magari corredate dai dati relativi all'etnia dei soggetti o alla loro professione. Pensiamo ad uno studio sui trasporti, che non può non considerare l'orografia del territorio, la distribuzione della popolazione, la tipologia e consistenza dell'apparato produttivo. Oppure consideriamo una ricerca storica, agevolata dalla disponibilità di mappe e di archivi on-line. Pensiamo soprattutto a cosa può significare poter mettere insieme tutti questi dati, e legare quindi una ricerca ad un'ampia rete di informazioni, che consenta correlazioni estese e magari originali, scoprendo interconnessioni a volte ovvie, altre volte inizialmente ritenute assurde o senza significato. Sarei ad esempio sorpreso se, analizzando dati di consumo e patologie, non si scoprissero correlazioni fra l'aumento o la diminuzione dei consumi di determinati prodotti alimentari e l'insorgere o il ridursi di determinate malattie. Però occorrono i dati dei supermercati e dei centri medici. Già analisi limitate ad un solo data-base hanno portato a scoperte imprevedibili, come lo studio dei dati riferiti alle auto usate negli USA, che ha riscontrato una correlazione statisticamente significativa fra i colori sgargianti ed il minor tasso di guasto riscontrato dai nuovi proprietari[23]. Poiché non appare ragionevole concludere che la vernice gialla o rossa preservi il motore più di quella grigia, è evidente che devono sussistere altri fattori, come una maggior visibilità che riduce la probabilità di incidente oppure il fatto che si tratta di colori scelti soprattutto dagli appassionati, che quindi usano meglio il mezzo. Al di là del caso curioso, comunque, si vuole evidenziare come i dati disponibili, soprattutto se incrociati fra loro, consentono analisi e conclusioni di ogni tipo. Certo, qualunque dato ricavato attraverso un servizio pubblico o privato dovrà essere verificato, analizzato correttamente, inserito in modo appropriato in un contesto più generale, ma trovarsi direttamente sul WEB un intero patrimonio di conoscenza agevola certamente l'attività del ricercatore, che non deve districarsi fra richieste inevase, uffici che non collaborano, dati in mille formati diversi, magari obsoleti o stampati su illeggibili moduli continui. E gli uffici pubblici, vedendo che la pubblicazione degli open data non è uno sterile adempimento normativo che comporta solamente una perdita di tempo, bensì uno strumento importante che viene utilizzato e verificato, saranno incentivati a pubblicare in modo costante

[23] Big Data: A Revolution That Will Transform How We Live, Work, and Think, Viktor Mayer-Schonberger - Kenneth Cukier, Mariner Book, 2013

dati di qualità. Se poi quest'incentivazione non fosse solo morale, ma fosse parte della valutazione sui progetti-obiettivi, che dovrebbero portare nella PA un minimo di meritocrazia, sarebbe certamente una spinta in più.

Insomma, un circolo virtuoso che non è facile innescare ma che, una volta messo in moto, non può che portare benefici a tutta la società.