Macerata, 30 marzo 2017

Open Data e protezione dei dati personali nel contesto dell'Agenda Digitale Italiana

La protezione dei dati personali nella pubblicazione degli open data

Francesco Ciclosi

Un nuovo documento di riferimento per le PA

- Il 14 dicembre 2016 l'Agenzia per l'Italia Digitale ha pubblicato le **Linee Guida Nazionali per la Valorizzazione del Patrimonio Informativo Pubblico 2016**
- Si tratta di un documento di riferimento per le PA che pubblicano i propri dati secondo il paradigma dei dati aperti

Dove reperire le linee guida

- Il documento è liberamente scaricabile dal sito http://www.dati.gov.it/content/linee-guida-open-data-2016
- La licenza di rilascio è la Creative Commons - Attribuzione 4.0 Internazionale (CC BY 4.0)
- Il formato di rilascio è duplice : PDF e ODT

La struttura delle linee guida 2016

- Il documento approfondisce:
  - da un lato un **modello** e un'**architettura di riferimento** per l'informazione del settore pubblico
  - dall'altro lato gli **aspetti organizzativi** necessari a individuare ruoli e figure professionali delle pubbliche amministrazioni coinvolte nella gestione dei dati aperti
- Il documento è carente per quanto riguarda la protezione dei dati personali oggetto di pubblicazione

Il contesto di riferimento

- Privacy dei dati in scenari di pubblicazione
  - Fa riferimento al rilascio a terze parti di dati che si hanno in gestione
  - Avviene in modo controllato

- **Senza dimenticare che lo scopo è quello di massimizzare il rilascio dell'informazione**

Il rilascio dei dati statistici

- Sono solitamente rilasciati in forma aggregata
- Due scenari classici in letteratura
  - Dati statistici
  - DBMS statistici
    - Presentano un maggiore grado di complessità
    - Presentano due grandi criticità: il problema della collusione e quello della finestra temporale

Scenari a confronto

Dati statistici	DBMS statistici
Vengono pubblicate delle statistiche precostituite	Si lascia all'utente la possibilità di crearsi delle statistiche «su misura»
Il controllo sull'eventuale rilascio indiretto di informazioni:	
È effettuato a monte prima della pubblicazione	È effettuato in tempo reale tenendo conto del problema della collusione e della finestra temporale
Viene effettuato un rilascio dei dati statistici	Non viene effettuato un rilascio dei dati statistici
Possono essere rilasciati senza esporre l'informazione confidenziali relativa ai rispondenti	La sorgente dati (microdati) ha delle informazioni sensibili che possono essere esposte se non si tiene conto della sequenza delle query statistiche

Le forme di rilascio

- I dati possono essere rilasciati in due forme:
  - Macrodati
    - Sono informazioni statistiche in forma aggregata
  - Microdati
    - Non sono informazioni in forma aggregata
    - Sono dati in forma specifica (le vere tuple)
- I microdati sono esposti a maggiori rischi di violazione della privacy

Attenzione

- A volte si stanno rilasciando delle informazioni anche quando si pensa di non rilasciare nulla

Infatti

- Il fatto che l'informazione venga rilasciata o meno non dipende dalla sua sensibilità bensì da ciò che già si conosce

Nel dubbio è meglio non rilasciare nulla

Esempio di tabella con microdati

Scuola di provenienza	Regione	Corso	Età	Nazionalità
Liceo Scientifico	Marche	Corso A	24	Italiana
Liceo Classico	Marche	Corso A	22	Italiana
Liceo Scientifico	Marche	Corso B	22	Italiana
Liceo Classico	Marche	Corso B	21	Italiana
ITC	Marche	Corso A	20	Italiana
ITC	Marche	Corso C	20	Francese
Liceo Scientifico	Marche	Corso B	21	Italiana
ITC	Marche	Corso A	20	Italiana
ITC	Marche	Corso D	24	Italiana

- **Solitamente l'identità è rimossa, ma questo non basta**

Esempio di tabella di conteggio

Risultati degli esami di profitto							
Votazione	**0-17**	**18-21**	**22-24**	**25-27**	**28-29**	**30+**	**Totale**
Insegnamento A	3	6	14	35	10	1	69
Insegnamento B			8	12			20
Insegnamento C	5		41	25	3		74
Insegnamento D				3			3

- Non vanno bene celle per cui:
  - il totale è uguale al numero di rispondenti della cella
  - il range di incertezza è al di sotto di un range ammesso
- Se in alcuni casi ho poche persone e queste cadono tutte all'interno di un range ⇒ queste sono esposte

Rivelazione delle informazioni - 1

- Si verifica quando una terza parte può identificare un rispondente a partire dai dati rilasciati
- Anche se, rivelare che un individuo è un rispondente in una raccolta dati non necessariamente determina una violazione dei requisiti di confidenzialità

Rivelazione delle informazioni - 2

- Esistono molti tipi di rivelazione delle informazioni
  - Rivelazione dell'identità
    - a partire da un dato rilasciato è possibile identificare un rispondente
  - Rivelazione degli attributi
    - a partire dai dati rilasciati vengono rivelate delle informazioni sensibili relative un rispondente
  - Rivelazione induttiva
    - a partire dai dati rilasciati è possibile determinare il valore di alcune caratteristiche di un rispondente senza che vi siano record che vi facciano esplicito riferimento

Strategie di protezione

- La strategia di protezione richiede più passaggi:
  1. Scelta della modalità di divulgazione, in base alla natura dei dati è possibile effettuare limitazioni in tal senso per proteggere la riservatezza
  2. Preparazione dei microdati per il rilascio, sono richieste varie azioni a partire da quella basilare di rimozione degli identificatori espliciti
  3. Controllo della quantità d'informazione rilasciata, eventualmente rivedendo i processi precedenti

Tecniche di protezione dei macrodati - 1

- Le tecniche di protezione includono:
  - Campionatura
    - pubblicare dati aggregati (su un campione statisticamente rappresentativo) e non l'intero insieme dei rispondenti
  - Regole speciali
    - imporre delle restrizioni sui dati che possono essere forniti in una tabella, per limitare il livello di dettaglio che potrebbe esporre l'informazione sensibile di una persona o la sua identità

Tecniche di protezione dei macrodati - 2

- Le tecniche di protezione includono:
  - Regole di soglia
    - imporre delle restrizioni sulle celle legate a un valore soglia, in modo da evitare il rilascio dei dati che fanno riferimento a un numero di rispondenti troppo piccolo

Tecniche di protezione dei microdati

- Servono a proteggere i microdati «sporcandoli» prima che vengano calcolate le statistiche
- Servono a evitare che la statistica esponga delle informazioni sui rispondenti
- Si classificano in:
  - Tecniche di mascheramento
  - Tecniche di generazione dei dati sintetici

Le tecniche di mascheramento

- Prevedono la trasformazione dell'insieme originale dei dati e possono essere di tipo:
  - Non perturbativo
    - Se i dati originali non vengono modificati
    - Anche se alcuni dati vengono soppressi e/o alcuni dettagli vengono rimossi *(campionamento, soppressione locale, generalizzazione)*
  - Perturbativo
    - Se i dati originali vengono modificati *(arrotondamento, scambio dei dati tra le tuple, introduzione controllata di errori)*

Le tecniche di generazione dei dati sintetici

- Prevedono il rilascio di dati plausibili (ma generati ad arte) al posto di quelli veri
- Le statistiche vengono calcolate su tali dati appositamente generati

Anonimato e re-identificazione - 1

- I cosiddetti dati «in forma anonima» sono de-identificati prima del rilascio, mediante rimozione di ogni identificatore esplicito (es. codice fiscale)

➡ La de-identificazione non è sufficiente

Anonimato e re-identificazione - 2

- Infatti, esistono altri insiemi di dati che includono le identità degli individui con le loro informazioni anagrafiche

Il **collegamento** tra tali **entità** e
i **dati de-identificati** può **comportare**
la **re-identificazione dei dati originali**

Microdati e classificazione degli attributi

- Identificatori
  - attributi che identificano in modo univoco il rispondente dei microdati (ad esempio il codice fiscale)
- Semi-identificatori
  - **attributi** che, in combinazione, possono essere **collegati** con informazioni esterne per **re-identificare** (o **ridurre l'incertezza sull'identità**) tutti o alcuni dei rispondenti a cui l'informazione si riferisce (es: data di nascita, CAP)

Il meccanismo di re-identificazione

Codice Fiscale	Nome	Cognome	Data di Nascita	Sesso	CAP	Stato civile	Sussidio
…	….	…	21/3/1980	M	62100	Celibe	Tipo 1
…	….	…	17/4/1977	F	62032	Coniugato	Tipo 3
…	….	…	19/5/1971	F	62029	Nubile	Tipo 1
…	….	…	6/8/1962	M	62100	Divorziato	Tipo 4
…	….	…	3/3/1945	F	00100	Coniugata	Tipo 7

Record anagrafico

Nome	Cognome	Indirizzo	Città	CAP	Data di nascita	Sesso	Stato civile
Maria	Rossi	V. Pino	Roma	00100	19/5/1971	F	Nubile
Carlo	Bianchi	V. Olmo	Roma	00100	1/4/1979	M	Celibe
Luisa	Verdi	V. Acero	Roma	00100	4/5/1945	F	Coniugata

Record social network

Una riflessione sui fattori di rischio

- Quello che è un semi identificatore cambia da persona a persona
  - Esempio: normalmente il lavoro non lo è ma esiste un solo Presidente della Repubblica
- Bisogna fare attenzione alle caratteristiche peculiari di un individuo che ne possono determinare l'identificazione
- Più attributi ho nella base dati più il collegamento è facilitato e i più rispondenti sono identificabili

I miei contatti

linkedin
http://it.linkedin.com/pub/francesco-ciclosi/62/680/a06/

facebook
https://www.facebook.com/francesco.ciclosi

twitter
@francyciclosi

www
http://www.francescociclosi.it
http://docenti.unimc.it/f.ciclosi