Le tecnologie ed i dati linguistici – quadro normativo

Rebecca Berto

FPPT.com

Dante: la Torre di Babele

Pieter Bruegel „La Torre di Babele“ (1563)

Dante:

- *Inferno* Canto XXXI, 58-81 „*Elli stessi s‘accusa; questi è Nembrotto per lo cui mal coto un linguaggio nel mondo non s‘usa*“ (75-78);
- *Purgatorio* Canto XII, 34-36 „*Vedea Nembrot a pie del gran lavoro quasi smarrito, e riguardar le genti che‘n Sennar con lui superbi fuoro*“

Risoluzioni del Parlamento Europeo

Risoluzione del Parlamento Europeo 2018/2028 – **uguaglianza linguistica nell'era digitale**:

- si sottolinea l'importanza delle tecniche di **estrazione di testo** e **di dati** per lo sviluppo delle **tecnologie del linguaggio**;
- si pone l'accento sull'esigenza di rafforzare la collaborazione tra l'industria e i proprietari di dati;
- si evidenzia la necessità di adeguare il quadro normativo e di garantire un utilizzo e una raccolta di **risorse linguistiche** più aperte e interoperabili;
- si riconosce che le informazioni sensibili non dovrebbero essere consegnate alle aziende commerciali e ai loro programmi gratuiti, in quanto non è chiaro in che modo esse potrebbero utilizzarle.

Proposta di risoluzione del Parlamento Europeo **2020/2217** – **strategia europea per dati**:

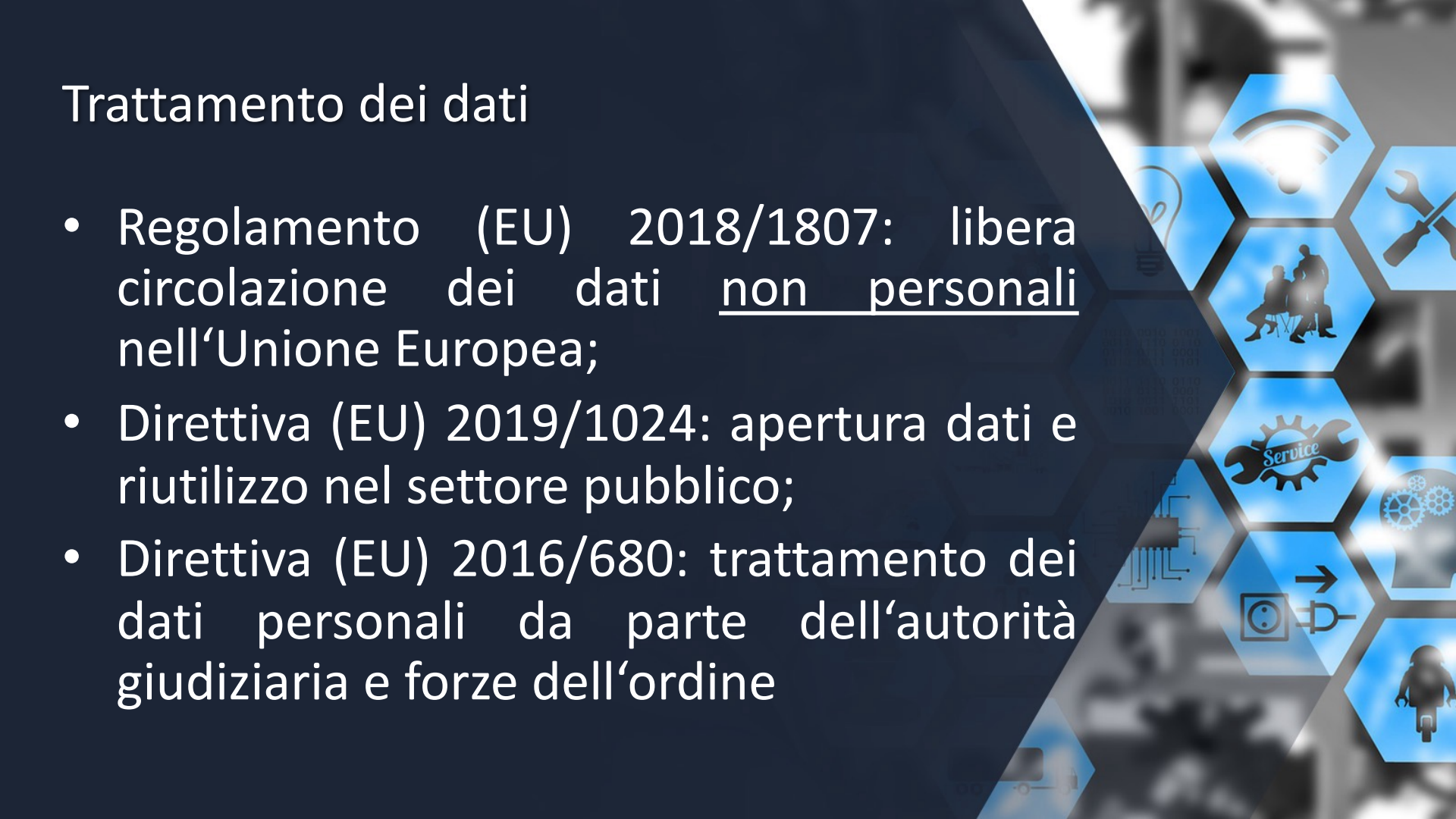
- la Commissione dovrebbe presentare una legge sui dati per incoraggiare e consentire in tutti i settori un flusso di dati più ampio ed equo fra imprese (B2B), fra impresa e pubblica amministrazione (B2G) e fra pubbliche amministrazioni (G2G);
- l'attuazione della strategia europea per i dati deve trovare un equilibrio tra la promozione di un uso e una condivisione più ampi dei dati e la protezione dei diritti di proprietà intellettuale (DPI) e dei segreti commerciali ed anche con i diritti fondamentali quali la privacy;
- si rileva che i dati utilizzati per l'addestramento dell'intelligenza artificiale (IA) si basano talvolta su dati strutturati quali banche dati, opere protette dal diritto d'autore e altre creazioni che beneficiano della protezione della proprietà intellettuale e che generalmente potrebbero non essere considerate come dati.

Risorse linguistiche e dati linguistici

- **Linguistica**: Scienza che studia sistematicamente il linguaggio umano nella totalità delle sue manifestazioni, e quindi le lingue come istituti storici e sociali, la loro ripartizione, i loro reciproci rapporti, nonché la funzionalità delle singole lingue sotto differenti aspetti (fonetico, sintattico, lessicale, semantico), sia nella struttura con cui si presentano in un determinato momento della loro storia sia nella loro evoluzione attraverso il tempo;
- **Risorse linguistiche**: insieme di dati e descrizione linguistiche in formato leggibile meccanicamente compresi corpora scritti e parlati, grammatica e banche dati terminolgiche (ELRC)

Trattamento dei dati

- Regolamento (EU) 2018/1807: libera circolazione dei dati non personali nell‘Unione Europea;
- Direttiva (EU) 2019/1024: apertura dati e riutilizzo nel settore pubblico;
- Direttiva (EU) 2016/680: trattamento dei dati personali da parte dell‘autorità giudiziaria e forze dell‘ordine



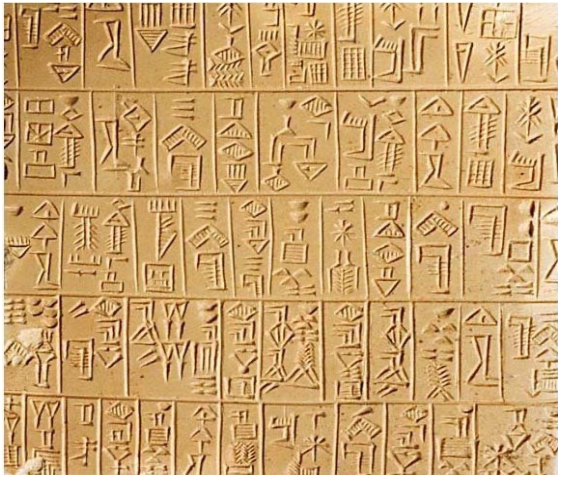
Patrimonio culturale: ricerca storico-archeologica

Premio Sofia Kovalevskaya dalla Fondazione A. von Humboldt al fine di raccogliere un „Fragmentarium“ (= database dei frammenti presenti nei vari musei);

Enrique Jiménez impiega l'intelligenza artificiale per colmare le lacune presenti nei testi scritti;

Un singolo carattere può spesso fare la differenza tra una lettura chiara e un **enigma linguistico.**

Documento del XXVI a.C.: lista delle offerte fatte alle sacerdotesse di Adab in occasione della loro nomina



Lingue vive – interesse pubblico?

I dati linguistici o le risorse linguistiche sono dati personali? Costituisce un interesse pubblico allenare un'intelligenza artificiale con risorse linguistiche?

- Sono **dati personali** le informazioni che identificano o rendono identificabile, **direttamente o indirettamente**, una persona fisica e che possono fornire informazioni sulle sue caratteristiche, le sue abitudini, il suo stile di vita, le sue relazioni personali, il suo stato di salute, la sua situazione economica, ecc..

- Art.6 (f) GDPR: il trattamento è necessario per il perseguimento del legittimo interesse del titolare del trattamento o di terzi, a condizione che non prevalgano gli interessi o i diritti e le libertà fondamentali dell'interessato che richiedono la protezione dei dati personali, in particolare se l'interessato è un minore;

- Art2(2) Regolamento (UE) 2018/1807: Nel caso di un insieme di dati composto sia da dati personali che da dati non personali, il presente regolamento si applica alla parte dell'insieme contenente i dati non personali. Qualora i dati personali e non personali all'interno di un insieme di dati siano indissolubilmente legati, il presente regolamento lascia impregiudicata l'applicazione del regolamento (UE) 2016/679.

Decisione BGH I ZR 133/17

BGH I ZR 133/17: Caso che ha sollevato un dibattito giuridico, tutt'ora in corso:

- contratto d'opera fra un editore di materie giuridiche ed un Professore di diritto civile: contratto riguardava un commentario di diritto civile;
- collaborazione venuta meno: l'ultima edizione del commentario è del 2013;
- scaturì una controversia giudiziale, giunta fino al Bundesgerichtshof (=BGH).

Naturalmente le sentenze vengono pubblicate anonimizzando i dati personali.

Il Professore obiettò che la tecnica di anonimizzazione usata non era sufficiente. Inoltre, considerando la data dell'ultima edizione del commentario e le altre sue numerose pubblicazioni, egli risultava identificabile da un numero ampio di esperti civilisti. Sulla base di tale argomentazione la pubblicazione della sentenza venne temporaneamente sospesa per due anni.

Sul contrasto fra un interesse pubblico, come la pubblicazione di una sentenza, e la possibilità di ledere la sfera privata, la Corte d'Appello di Monaco di Baviera (OLG München Verfügung 24 agosto 2020 – 6 St 1/19) ritenne che eventuali lesioni alla sfera personale possano essere inevitabili nonostante le tecniche di anonimizzazione utilizzate.

Proposta - regolamento Intelligenza artificiale 2021/0106 (COD)

- Definizione di dati utilizzati per allenare un‘IA, adattando i parametri di allenamento (art.3 co.29);
- Allo scopo di individuare e/o correggere un sistema IA è possibile processare speciali categorie di dati personali (art.10 co.5);
- Nei limiti in cui sistemi IA innovativi impiegano dati personali o IA siano compresi nella supervisione di autorità competenti, quest‘ultime sono associate nell‘operazioni che coinvolgono l‘IA (art.53-54)

In conclusione

- Risorse linguistiche per allenare un IA non sono ancora propriamente disciplinate;
- La distinzione dati personali e dati non-personali non è d‘aiuto e nemmeno le tecniche di anonimizzazione sono risolutive;
- La proposta di regolamento sull‘IA prende in considerazione dati necessari per allenare un‘IA, ma non chiarisce il trattamento dei dati linguistici e/o delle risorse linguistiche.