Elaborato Web Semantico
Un'ontologia sull'anonimizzazione dei dati
Anonym.me

Ismam Abu - 0001005104
Hamado Dene - 0001000974
Konrad Gomulka - 0000973128

November 2022

Contents

List of Figures

1 Introduzione

Nel contesto odierno il dato è una delle più grandi fonti di guadagno al mondo. Le organizzazioni sono sempre più esperte nella raccolta e monetizzazione di informazioni, ciò crea numerose opportunità di business ma altrettante di furto di dati, leak e violazione della privacy.

L'anonimizzazione dei dati è il processo mediante il quale i dati vengono "puliti", rimuovendo o criptando tutte le informazioni personali che potrebbero portare all'identificazione di un individuo. La pseudo-anonimizzazione, invece, è il processo mediante il quale si riduce la possibilità di identificare un individuo, senza però azzerarla. L'obiettivo di tali procedure è l'assicurazione di un livello di privacy adeguato mantenendo però la struttura dei dati, in modo che ne possano venire ricavate informazioni significative anche post-anonimizzazione.

L'anonimizzazione ha anche un ruolo fondamentale nello sviluppo e nell'impiego di Intelligenze Artificiali all'interno della società. In tale contesto vi è un ampio impiego di tecniche per l'anonimizzazione dei dataset mediante varie tecniche, quelle tradizionali si focalizzano sul "mascheramento dei dati" (i dati vengono offuscati es. criptandoli). Un buon livello di anonimato potrebbe convincere più soggetti a condividere i propri dati, migliorando significativamente la qualità dei dataset e quindi anche la performance di tali AI. Le AI si basano sui dati, pertanto l'accesso a moli più grandi di informazioni è il fulcro per sbloccarne il completo potenziale. Questo rende l'anonimizzazione un tema sempre più cruciale sia dal punto di vista delle organizzazioni che dello sviluppo e progresso della società.

Questo studio punta a realizzare un'ontologia completa sull'anonimizzazione dei dati. Tale ontologia dovrà rappresentare adeguatamente tutto il processo di anonimizzazione, a partire dalla differenziazione delle tipologie di documenti da cui estrarre i dati, alla scelta dei dati più significativi e tecniche di anonimizzazione adottabili. Si andrà inoltre ad implementare un applicativo che dimostri l'efficacia dell'ontologia.

In aggiunta a ciò il gruppo si è proposto di realizzare un'applicazione in grado di mostrare i contenuti dell'ontologia e di verificare in maniera molto semplice come funzionano processi di anonimizzazione dei dati.

2 Tecnologie e linguaggi utilizzati

- RDF: Linguaggio utilizzato per la definizione del modello e delle triple che descrivono il dominio di riferimento
- RDFS: Linguaggio utilizzato per estendere il vocabolario di RDF
- OWL: Linguaggio utilizzato per aumentare l'espressività dell'ontologia
- SPARQL: Linguaggio utilizzato per definire le query sulle istanze dell'ontologia e per effettuare l'inserimento dei dati
- Stardog Studio: Framework utilizzato per l'amministrazione dei dati e l'esecuzione delle query
- Protegè: Framework utilizzato per la modellazione e lo sviluppo dell'ontologia
- Node.js Framework utilizzato per la realizzazione dell'applicativo

3 Analisi del dominio e del contesto

3.1 Analisi del contesto

3.1.1 Anonimizzazione e pseudoanonimizzazione

Il GDPR (Regolamento generale sulla protezione dei dati) definisce le informazioni anonime come "informazioni che non si riferiscono a una persona fisica identificata o identificabile o a dati personali resi sufficientemente anonimi da impedire o da non consentire più l'identificazione dell'interessato".

Pertanto l'anonimizzazione è un processo che si occupa di rimuovere identificatori personali (diretti o indiretti che siano) che potrebbero portare un individuo ad essere identificato. Un identificatore è diretto quando un individuo può essere identificato attraverso un singolo dato, come il proprio nome, indirizzo, numero di telefono ecc. Invece un identificatore è detto indiretto se un individuo può invece essere identificato aggregando più dati, come luogo di lavoro, titolo, indirizzo postale o una diagnosi ospedaliera. Se un dato è stato correttamente anonimizzato, esso non può più essere associato ad un certo individuo, dunque il dato può rientrare nell'ambito della GDPR ed è più semplice da utilizzare

La pseudoanonimizzazione è un concetto leggermente diverso dall'anonimizzazione, viene definito dal GDPR come "il trattamento dei dati personali in modo tale che i dati personali non possano più essere attribuiti a un interessato specifico senza l'utilizzo di informazioni aggiuntive, a condizione che tali informazioni aggiuntive siano conservate separatamente e soggette a misure tecniche e organizzative intese a garantire che tali dati personali non siano attribuiti a una persona fisica identificata o identificabile". La pseudoanonimizzazione è quindi un processo reversibile, mentre con l'anonimizzazione l'identità degli individui è irrecuperabile.

3.2 Processo di anonimizzazione

Deloitte descrive il processo attraverso cui si anonimizzano i dati con il diagramma in figura 1.

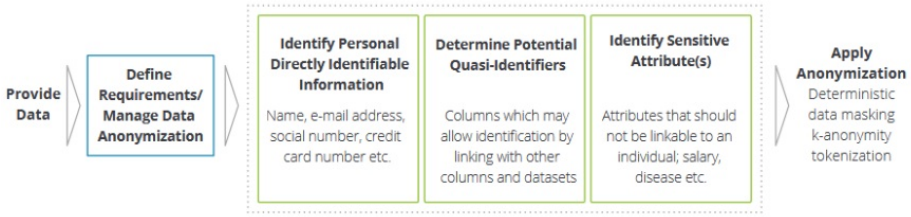

Figure 1: Rappresentazione di Deloitte del processo di anonimizzazione dei dati.

Alla base di tutto vi sono i dati, essi potrebbero essere rappresentati in una moltitudine di formati come es. tabelle di excel, file csv o liste di stringhe. Per anonimizzare tali dati il primo passo è una corretta analisi della loro struttura, il cui fine è la comprensione e il riconoscimento da parte dell'analista o del personale incaricato di quali dati sono i più significativi, la cui integrità dovrà pertanto essere mantenuta.

Una volta che si sono ottenuti i dati, si classificano gli attributi dei dati, questi possono essere:

1. 1. Identificatori diretti: come nome, indirizzo, numero di telefono, numero di targa, indirizzo e-mail ecc.

2. 2. Identificatori indiretti: sesso, data di nascita, età, codice postale, numero di figli ecc.

Successivamente viene definito quello che si dice attributo sensibile, partendo da questo si misura il rischio di identificazione e lo si collega ad una soglia di rischio che rappresenta il grado di rischio che si ritiene accettabile. Un valore di soglia pari a zero non consentirebbe la condivisione di dati utili, mentre un valore uno significa che nessun dato è anonimizzato.

Dopo aver indentificato quali dati dovrebbero essere anomizzati e quali no, diventa importante identifcare la tecnica più adeguata in base al settore di riferimento. Più precisamente, la tecnica utilizzata deve permettere il trattamento dei dati in maniera tale da non consentire l'identificazione della persona interessata mediante l'insieme dei mezzi che possono essere ragionevolmente utilizzati. Nel prossimo capitolo si andranno a descrivere alcune delle tecniche utilizzate (che sono tutt'ora oggetto di ricerca).

3.3 Tecniche di anonimizzazione

Di seguito sono descritte alcune delle tecniche di anonimizzazione che vengono generalmente utilizzate:

1. **Data Masking**: il data masking si riferisce alla divulgazione di dati con valori modificati. L'anonimizzazione dei dati viene eseguita creando un'immagine speculare di un database e implementando strategie di alterazione, come il mescolamento dei caratteri, la crittografia, i termini o la sostituzione dei caratteri. Ad esempio, un carattere valore può essere sostituito da un simbolo come "" o "x". Rende difficile l'identificazione o il reverse engineering.

2. **Pseudoanonimizzazione**: la pseudonimizzazione è uno strumento di anonimizzazione dei dati che sostituisce gli identificatori privati con falsi identificatori o pseudonimi, ad esempio scambiando l'identificatore "John Smith" con l'identificatore "Mark Spencer". Mantiene la precisione statistica e la riservatezza dei dati, consentendo di utilizzare i dati modificati per la creazione, la formazione, il test e l'analisi, mantenendo allo stesso tempo la privacy dei dati.

3. **Generalizzazione**: la generalizzazione implica l'esclusione di alcuni dati di proposito per renderli meno identificabili. I dati possono essere modificati in una serie di intervalli o in un'ampia regione con limiti ragionevoli. Ad esempio, il numero civico di un indirizzo può essere cancellato, ma assicurati che il nome della corsia non venga cancellato. L'obiettivo è rimuovere alcuni degli identificatori mantenendo l'accuratezza dei dati.

4. **Data swapping**: il data swapping, spesso noto come permutazione e mescolamento, riorganizza i valori degli attributi del set di dati in modo che non si adattino alle informazioni originali. Cambiare attributi (colonne) che includono valori riconoscibili, come la data di nascita, può avere un enorme impatto sull'anonimizzazione.

5. **Data perturbation**: la perturbazione dei dati modifica marginalmente il set di dati iniziale applicando metodi di numerazione arrotondata e aggiungendo rumore casuale. L'insieme dei valori deve essere proporzionale al disturbo. Una piccola base può contribuire a una scarsa anonimizzazione, mentre una base ampia può ridurre l'utilità di un set di dati. Ad esempio, una base di 5 dovrebbe essere utilizzata per arrotondare valori come l'età o il numero civico.

6. **Synthetic data**: I synthetic data sono informazioni generate algoritmicamente senza alcuna relazione con alcun caso reale. I dati vengono utilizzati per costruire set di dati artificiali invece di modificare o utilizzare il set di dati originale e compromettere la privacy e la protezione. Il metodo dei synthetic data include la costruzione di modelli matematici basati su modelli contenuti nel set di dati originale. Deviazioni standard, regressione lineare, mediane o altri metodi statistici possono essere utilizzati per produrre risultati sintetici.

Di seguito alcune tecniche che nello specifico utilizzano distribuzioni particolare per i propri algoritmi:

1. **Introduzione di rumore statistico**: Consiste nell'introduzione di una perdurbazione in alcuni tipi di attributi i cui valore diventa meno accurato pur mantenendo la distribuzione generale. Nonostante alcuni valori vengano approssimati, rimane comunque un tecnica efficace. Un esempio può essere l'altezza di una persone che invece di essere precisa è approssimata ai 5cm. Più individui si troveranno ad avere la stessa altezza come se fossere raggruppati per classi ad intervalli di 5 cm.

2. **k-anonymization**: ogni rilascio di dati deve essere tale per cui qualsiasi combinazione di quasi identificatori corrisponda ad almeno k identità; si ottiene generalizzando i valori dei identificatori indiretti, per esempio invece di fornire un'età precisa si raggruppano le persone per fasce di età e in ogni fascia sono contenuti almeno k individui. La tecnica si può applicare agli attributi numerici ma anche ad altri riducendo il livello di dettaglio come ad esempio la sostituzione della provincia al posto del paese. Una

tecnica simile è la generalizzazione che si ottiene per esempio modificando il valore di un CAP: eliminando l'ultima cifra si raggruppano le persone non più per paese ma per aree di 10 CAP diversi. La k-anonymity richiede che in una tabella ogni quasi identificatore sia presente almeno k volte.

3. **l-diversity**: in ogni classe di equivalenza di una tabella (insieme di tuple con uno stesso valore per un quasi attributo) un attributo sensibile deve avere almeno "l" valori diversi. Può essere usata per estendere la k-anonymity. Perché sia efficace la distribuzione in un insieme di record deve essere il più possibile vicina a quella della popolazione originale; per esempio se in una popolazione di pazienti il 30% ha una patologia, allora nell'insieme di record che soddisfa la l-diversity la percentuale di pazienti con la stessa patologia deve essere molto vicina al 30%, in caso contrario si possono verificare casi di asimmetria e similarità fra tuple tramite cui è possibile ridurre le classi di appartenenza e dedurre informazioni più dettagliate sui singoli pazienti ovvero stimare con una maggiore probabilità le informazioni riservate su un individuo.

4. **t-closeness (t-vicinanza)**: realizza un affinamento della l-diversity, ovvero tende a realizzarne la condizione migliore in cui la distribuzione iniziale degli attributi nella tabella viene rispettata anche nelle singole classi di appartenenza. Una classe di equivalenza rispetta la t-closeness se la distanza tra la distribuzione dei valori dei quasi identificatori nella classe e nell'intera popolazione è inferiore a "t"; in altre parole l'intero insieme di record e una sua parte hanno una distribuzione delle informazioni molto simile per cui non è possibile dedurre dettagli maggiori da quelli già noti.

5. **differential privacy**: lo scopo dell'algoritmo è fare in modo che la distribuzione di probabilità dei dati pubblicati sia la stessa indipendentemente dalla presenza nel dataset delle informazioni di un individuo oppure no. Lo scopo è evitare che si possa rilevare la presenza o l'assenza delle informazioni di un individuo in un dataset. Si realizza introducendo del rumore statistico casuale sul dataset ma a posteriori di una interrogazione (query), quindi senza modificare i dati originali. Maggiore è il rumore introdotto e maggiore è il livello di privacy ma minore è l'accuratezza. Come algoritmo fornisce una protezione migliore rispetto alla k-anonymization, tuttavia non garantisce una protezione completa.

Esiste un elemento importante che permette di attaccare i diversi algoritmi descritti. In ogni caso di pubblicazione dati è necessario considerare il cosiddetto "external knowledge" rappresentato dalla conoscenza di alcuni dettagli sugli individui di una popolazione che una terza parte possiede. Queste informazioni oltre che essere note a priori per i più disparati motivi, quali amicizia, popolarità, possono anche essere ricavate lavorando su diversi dataset rilasciati da diverse organizzazioni, da rilasci periodici dello stesso dataset con aggiornamenti e il livello di conoscenza che producono non è noto a priori.

4 Modellazione dell'Ontologia

Il grafo dell'ontologia è mostrato in figura 2.

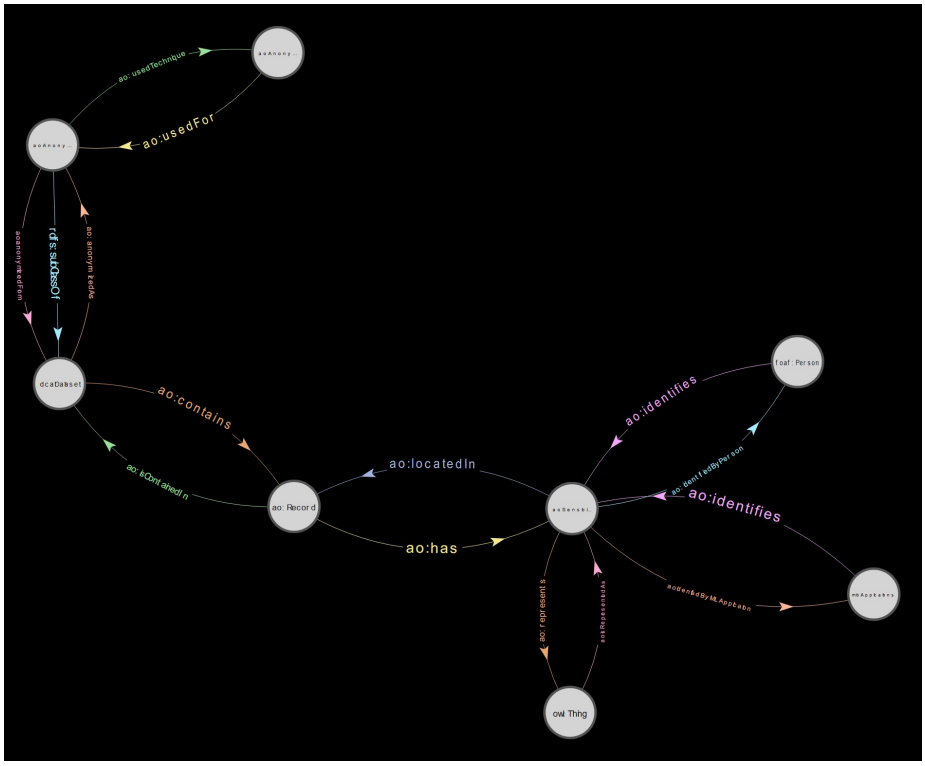

Figure 2: Grafo dell'ontologia

4.1 Classi modellate

4.1.1 Record

Entità che rappresenta un record appartenente ad un dcat:Dataset.
Attributi e associazioni

- text (String): attributo che rappresenta il testo del Record
- has (SensibleThing): relazione atta ad individuare gli elementi che compongono il record.

4.1.2 AnonymizedDataset

Entità che rappresenta un dataset ai cui record è stata applicata una ao:anonymizationTechnique. È implementata come sottoclasse di dcat:Dataset e si ottiene tramite la object

property anonymizedAs partendo dall'ao:Dataset.
Attributi e associazioni

- isPseudoAnonymized (Boolean): attributo che indica se il dataset è pseudoanonimizzato.
- usedTechnique (AnonymizationTechnique): relazione atta ad individuare le tecniche di anonimizzazione utilizzate per anonimizzare il dataset.

4.1.3 AnonymizationTechnique

Entità che rappresenta la tecnica di anonimizzazione che può essesre applicata ad un dataset.
Attributi e associazioni:

- name (string): attributo che rappresenta il nome della tecnica di anonimizzazione
- description (string): attributo che rappresenta una descrizione della tecnica di anonimizzazione

4.1.4 SensibleThing

Entità che rappresenta un elemento di un record che può essere un identificatore diretto o indiretto. SensibleThing permette di distinguere concetti potenzialmente sensibili, se presenti, e di metterli in relazione col soggetto che rappresentano. Si è scelto di modellare questa relazione col soggetto poichè le entità contenute nel dataset sono note soltanto a chi anonimizza il dataset, al quale potrebbe risultare utile ad es. data una foaf:Person collegarla con tutti i Dataset in cui è mezionata.
Attributi e associazioni:

- text (String): attributo che rappresenta il valore testuale della SensibleThing
- position (int): attributo che rappresenta la posizione all'interno del record della SensibleThing
- identifiedBy (foaf:Person, mlo:Applications): relazione atta ad individuare la persona o l'algoritmo di machine learning utilizzato per etichettare l'elemento come sensibile
- represents (owl:Thing): relazione atta ad individuare il dato sensibile a cui si fa riferimento

4.2 Object-Properties modellate

In seguito le Object-Properties modellate, alcune sono state modellate come subProperty di altre in modo da poter sfruttare il Property Chain.

4.2.1 contains

- descrizione: proprietà transitiva che mette in relazione un dataset con i record che lo compongono
- dominio: dcat:Dataset
- range: Record
- subProperty Of: hasPart
- proprietà inversa: isContainedIn

4.2.2 isContainedIn

- descrizione: proprietà transitiva che mette in relazione un record col dataset a cui appartiene
- dominio: Record
- range: dcat:Dataset
- subProperty Of: mlo:isPart
- proprietà inversa: contains

4.2.3 has

- descrizione: proprietà che mette in relazione un record con gli elementi sensibili che contiene
- dominio: Record
- range: SensibleThing
- subProperty Of: hasPart
- proprietà inversa: locatedIn

4.2.4 locatedIn

- descrizione: proprietà che mette un elemento sensibile col record a cui appartiene
- dominio: SensibleThing
- range: Record
- subProperty Of: mlo:isPart
- proprietà inversa: has

4.2.5 identifiedBy

- descrizione: proprietà che mette in relazione un elemento sensibile con l'elemento che lo ha classificato come tale. Nel nostro dominio non viene mai utilizzata direttamente ma solamente come "interfaccia" per le sub-Properties.
- dominio: owl:Thing
- range: owl:Thing
- proprietà inversa: identifies

4.2.6 identifiedByPerson

- descrizione: proprietà che mette in relazione un elemento sensibile con la persona che lo ha classificato come tale
- dominio: Record
- range: foaf:Person
- subProperty Of: identifiedBy
- proprietà inversa: identifies

4.2.7 identifiedByMLTechnique

- descrizione: proprietà che mette in relazione un elemento sensibile con l'algoritmo di Machine Learing che lo ha classificato come tale
- dominio: Record
- range: mlo:Applications
- subProperty Of: identifiedBy
- proprietà inversa: identifies

4.2.8 identifies

- descrizione: proprietà che mette in relazione una persona o un algoritmo di machine learning con gli elementi che ha classificato come sensibili
- dominio: foaf:Person, mlo:Applications
- range: SensibleThing
- proprietà inversa: identifiedBy

4.2.9 represents

- descrizione: proprietà funzionale che mette in relazione un elemento sensibile col soggetto che esso rappresenta
- dominio: SensibleThing
- range: owl:Thing
- proprietà inversa: isRepresentedAs

4.2.10 isRepresentedAs

- descrizione: proprietà che mette in relazione un'entità con l'elemento di un record che la rappresenta
- dominio: owl:Thing
- range: SensibleThing
- proprietà inversa: represents

4.2.11 anonymizedAs

- descrizione: proprietà inversamente funzionale che mette in relazione una dataset con la sua versione anonimizzata
- dominio: dcat:Dataset
- range: AnonymizedDataset
- proprietà inversa: anonimizedFrom

4.2.12 anonimizedFrom

- descrizione: proprietà funzionale che mette in relazione un dataset anonimizzato col dataset dal quale è stato ricavato
- dominio: AnonymizedDataset
- range: dcat:Dataset
- proprietà inversa: anonymizedAs

4.2.13 usedTechnique

- descrizione: proprietà che mette in relazione un dataset anonimizzato con le tecniche di anonimizzazione utilizzate
- dominio: AnonymizedDataset
- range: AnonymizationTechnique
- proprietà inversa: usedFor

4.2.14 usedFor

- descrizione: proprietà che mette in relazione una tecnica di anonimizzazione con i dataset che la utilizzano
- dominio: AnonymizationTechnique
- range: AnonymizedDataset
- proprietà inversa: usedTechnique

4.2.15 mlo:isPart

- descrizione: proprietà dell'ontologia mlo, è l'inversa di hasPart ed è stata modificata in modo da essere transitiva
- dominio: owl:Thing
- range: owl:Thing
- proprietà inversa: hasPart

4.3 Ontologie esterne

Sulla base delle entità individuate sono state importate alcune ontologie esterne:

1. **DCAT (Data Catalog Vocabulary)**: vocabolario designato per descrivere cataloghi, dataset e dataservice. Utilizzato per modellare le sorgenti dei dati da anonimizzare.
2. **MLO (Machine Learning Ontology)**: ontologia che descrive tutto il dominio di machine learning. Utilizzata per descrivere eventuali tecniche di machine learning mediante le quali identificare gli identificatori diretti/indiretti.
3. **FOAF (Friend Of A Friend)**: ontologia atta a descrivere persone, organizzazioni, le loro caratteristiche e le relazioni con altre persone. Utilizzata per referenziare i dati individuati riguardanti persone/organizzazioni.

5 Descrizione delle query

In questo capitolo verranno riportate alcune query per l'estrazione di dati, scritte in linguaggio SPARQL. Il codice delle query che susseguono è implementato e può essere eseguito dall'applicativo.

5.1 Dato un Dataset, lista Record inseriti

```
SELECT ?record ?text
FROM ${from}
WHERE {
    ?record
        a ao:Record ;
        ao:text ?text ;
        ao:isContainedIn ${dataset}.
}
```

Listing 1: Dato un Dataset ritornare la lista Record inseriti

5.2 Dato un Dataset, lista Record Anonimizzati

```
SELECT ?record ?text
FROM ${from}
WHERE {
    ?record
        a ao:Record ;
        ao:text ?text ;
        ao:isContainedIn ${anonymizedDS}.
}
```

Listing 2: Dato un Dataset ritornare la lista Record Anonimizzati

5.3 Dato un dataset, numero di record sensibili e non sensibili

```
SELECT (COUNT(?record) as ?NotSensibleRecords) (COUNT(?recordSens)
    as ?SensibleRecords)
FROM ${from}
WHERE {
    {
    ?record
        a ao:Record ;
        ao:isContainedIn ${dataset}.
        MINUS {?record ao:has ?sens}
    } UNION {
    ?recordSens
        a ao:Record ;
        ao:has ?sens;
        ao:isContainedIn ${dataset}.
    }
```

```
}
```

Listing 3: Dato un dataset ritornare il numero di record sensibili e non sensibili

5.4 Dato un dataset, tutte le organizzazioni menzionate

La query utilizza il reasoner per il Property chain della Object property *isPart*

```
SELECT ?name
FROM ${from}
WHERE {
    ?org
        a foaf:Organization ;
        foaf:name ?name ;
        ao:isRepresentedAs ?thing .
    ?thing
        mlo:isPart ${dataset} .
}
```

Listing 4: Dato un dataset ritornare tutte le organizzazioni menzionate

5.5 Dato un dataset anonimizzato, le tecniche usate per anonimizzarlo

```
SELECT ?techniqueName ?techniqueDescription
FROM ${from}
WHERE {
    ?technique
        a ao:AnonymizationTechnique ;
        ao:name ?techniqueName;
        ao:description ?techniqueDescription;
        ao:usedFor ${anonymizedDS}.
}
```

Listing 5: Dato un dataset anonimizzato ritornare le tecniche usate per anonimizzarlo

5.6 Creare un Dataset

```
PREFIX dcat: <https://www.w3.org/TR/vocab-dcat-2/>
        INSERT DATA {
            GRAPH <https://github.com/turbo-ismam/WS-AO/> {
                <#DS_${newID}>
                    a dcat:Dataset .
            }
        }
```

Listing 6: Creare un Dataset

5.7 Creare un Record

```
PREFIX dcat: <https://www.w3.org/TR/vocab-dcat-2/>
        INSERT DATA {
            GRAPH <https://github.com/turbo-ismam/WS-AO/> {
                <${record.id}>
                    a ao:Record ;
                    ao:text "${record.text}" ;
                    ao:isContainedIn <${StardogPrefix(record.
                        dataset)}> .
                <${StardogPrefix(record.dataset)}>
                    ao:contains <${StardogPrefix(record.id)}> .
            }
        }
```

Listing 7: Creare un Record

5.8 Aggiungere una Sensitive Thing

```
INSERT DATA {
            GRAPH <https://github.com/turbo-ismam/WS-AO/> {
                <${sensitiveThing.id}>
                    a ao:SensitiveThing ;
                    ao:text "${sensitiveThing.text}" ;
                    ao:position ${sensitiveThing.position} ;
                    ao:locatedIn <${StardogPrefix(sensitiveThing.
                        record)}> ;
                    ${sensitiveThing.represents ? "ao:represents <"
                         + sensitiveThing.represents + "> ;" : ""}
                    ao:identifiedByMLTechnique <http://www.a2rd.net
                        .br/mlo#Text_Classification> .
                <http://www.a2rd.net.br/mlo#Text_Classification>
                    ao:identifies <${StardogPrefix(sensitiveThing.
                        id)}> .
                <${StardogPrefix(sensitiveThing.record)}>
                    ao:has <${StardogPrefix(sensitiveThing.id)}> .
                ${sensitiveThing.represents ? "<" + sensitiveThing.
                    represents + "> ao:isRepresentedAs <" +
                    StardogPrefix(sensitiveThing.id) + "> .": ""}
            }
        }
```

Listing 8: Aggiungere una Sensitive Thing

5.9 Creare un Dataset Anonimizzato

```
PREFIX dcat: <https://www.w3.org/TR/vocab-dcat-2/>
        INSERT DATA {
            GRAPH <https://github.com/turbo-ismam/WS-AO/> {
                <${anonymizedDataset.id}>
                    a ao:AnonymizedDataset ;
```

```
6                       ao:usedTechnique <${StardogPrefix(
                            anonymizedDataset.technique)}> ;
7                       ao:anonymizedFrom <${StardogPrefix(
                            anonymizedDataset.dataset)}> .
8                   <${anonymizedDataset.dataset}>
9                       ao:anonymizedAs <${StardogPrefix(
                            anonymizedDataset.id)}> .
10                  <${anonymizedDataset.technique}>
11                      ao:usedFor <${StardogPrefix(anonymizedDataset.
                            id)}> .
12              }
13          }
```

Listing 9: Creare un Dataset Anonimizzato

6 Descrizione dell'applicativo

L'applicativo è stato realizzato con l'obiettivo di mostrare una possibile applicazione in domini reali dell'ontologia, fornendo un esempio per aiutare a comprenderne la struttura e mostrare come essa posso verire utilizzata nel processo di anonimizzazione dei dati. Per l'implementazione è stato scelto il linguaggio TypeScript attraverso il framework Node.js e le seguenti librerie: 1. Solidjs: libreria per la creazione di interfacce utente reattive 2. Stardog.js: pacchetto npm ufficiale per comunicare con un'istanza di Stardog 3. bert-large-NER: una libreria di Token Classification sviluppata dalla community di Hugging Face utilizzata per identificare all'interno di un testo i dati potenzialmente sensibli

Anonym.me è un applicativo Web che permette di anonimizzare una serie di record inseriti dall'utente. Dato che tale applicativo è soltato un esempio di come è possibile utilizzare l'ontologia, gli elementi considerati sensibili sono solamente i **nomi di Organizzazioni** riconosciuti dalla libreria best-large-NER con confidenza maggiore al 60

Prima che un utente possa utilizzare l'applicativo, è necessario inserire nell'ontologia tutte le entità che non dipendono dall'input dell'utente (per esempio le possibili tecniche di anonimizzazione applicabili ai dati sorgente).

Per un utente che utilizza l'applicativo è possibile anonimizzare delle informazioni passando in input del testo contentente, per ciascuna riga, un record da anonimizzare (ciascuna riga verrà contata come record). Una volta premuto il pulsante "Anonymize", dai record inseriti verranno estratte le variabili sensibili e l'utente sarà reindirizzato alla pagina delle query. Il processo effettuato dall'applicativo è il seguente: 1. Viene generata una richiesta mediante le API fornite da HuggingFace per identificare i dati sensibili all'interno del record (passaggio eseguito immediatamente data la possibilità che le API non siano disponibili) 2. Vengono create le entità Dataset e AnonymizedDataset in modo da avere un riferimeno per i Record 3. Vengono inseriti i Record del Dataset e record dell'AnonimizedDataset 4. Eventuali dati sensibili individuati vengono collegati al relativo Record nonchè alle relative entità che rappresentano

Una volta reindirizzato alla pagina delle Query è possibile eseguire (con o senza reasoner) le query descritte nel capitolo precedente o query create sul momento dall'utente.

6.1 Immagini dell'applicativo

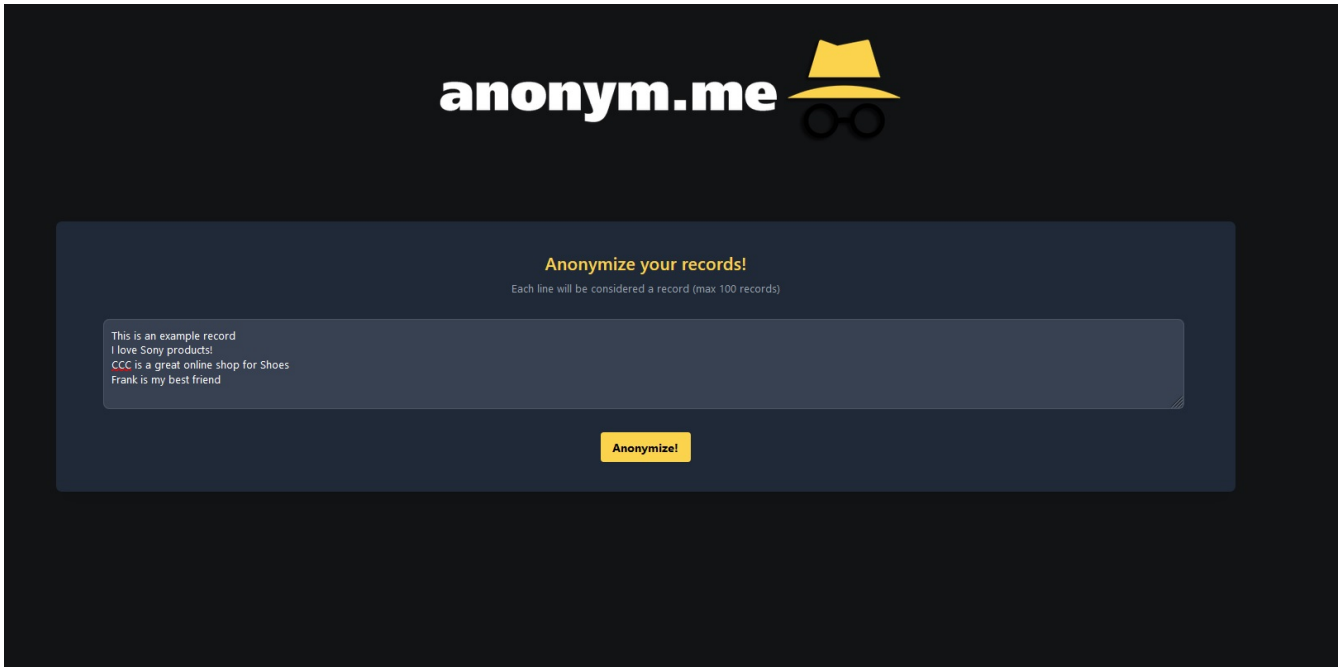

Figure 3: Home page

Figure 4: Pagine delle query con esempio di Dataset anonimizzato

Figure 5: Esempio di query con reasoner

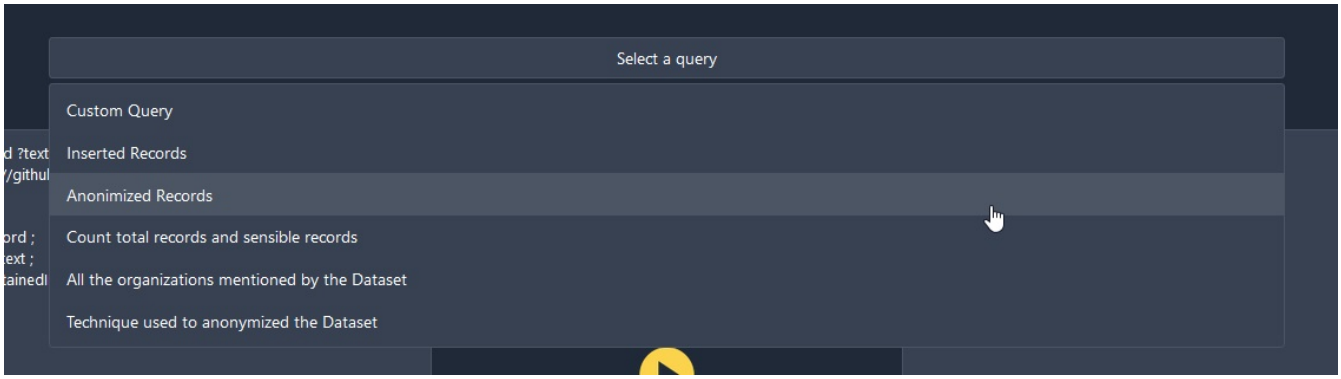

Figure 6: DropDown delle query

7 Conclusioni e sviluppi futuri

7.1 Considerazioni

Il Web Semantico è l'idea secondo la quale l'informazione dovrebbe essere espressa associandovi delle informazioni aggiuntive, in maniera da stabilirne il contesto semantico, le proprietà, relazioni e regole.

7.1.1 Vantaggi

I vantaggi non hanno solamente a che vedere con la velocità con cui i motori di ricerca possono elaborare quantità sempre maggiori di informazioni, ma soprattutto sull'usabilità delle risorse in rete. Se il Web Semantico fosse al massimo delle sue capacità, avremmo delle informazioni prive di ambiguità semantica, sarebbe possibile identificare tempetivamente solamente le informazioni rilevanti, colmando anche l'*information overload* che si verifica ai giorni nostri.

7.1.2 Svantaggi

Lo svantaggio più significativo è che per gli umani non è immediato inserire informazioni nel Web fornendo al contesto i metadati necessari a RDF. Tale difficoltà è uno dei punti critici del Web Semantico dato il rallentamento del processo di inserimento è alquanto limitante per la grande parte degli utenti nel web, che non sono disposti ad accettare tale rallentamento. Per tale motivo il passo principale da compiere per passare dal Web tradizionale al più ricco Web Semantico non dipende solamente dalle applicazioni intelligenti, ma anche dallo sviluppo di interfacce abbastanza semplici e di facile comprensione.

7.2 Conclusioni

L'obiettivo di questo progetto era quello di produrre un software applicando le conoscenze acquisite durante il corso, il gruppo si ritiene soddisfatto dal lavoro svolto e di come questo sia stato realizzato. Nonostante le iniziali perplessità riguardo il dominio propostoci, dopo un'attenta analisi del dominio siamo riusciti ad estrapolare le informazioni principali per poter realizzare un'applicazione che mostri il potenziale dell'ontologia. I dubbi principali riguardano l'aver compreso appieno il dominio dell'ontologia, essendo esso una cosa nuova e sconosciuta per tutti i membri del gruppo.

7.3 Sviluppi futuri

Il modello implementato è una versione basilare sull'ontologia pensata, un naturale ampliamento potrebbe derivare dalla correlate di altre onotologie esistenti, comprendendo ulteriori argomenti intersecanti come ad esempio la tipologia e il formato dei dati inseriti. Un aspetto un po' trascurato nell'applicativo è il formato di inserimento dei dati che in questo momento corrisponde unicamente ad una stringa di testo ma che potrebbe venire ampliato a vari altri come ad

es. .csv. Inoltre, vengono citate numerose tecniche ma nell'applicativo viene utilizzata sempre la stessa, si potrebbe pensare di far anonimizzare all'utente attraverso le tecniche che più gli aggradano. Sempre riguardo all'applicazione, si potrebbe far decidere all'utente la libreria di Machine Learning da utilizzare per il riconoscimento dei dati sensibili, o addirittura permettere all'utente di etichettare manualmente i dati che ritiene sensibili.