الجمهورية الجزائرية الديمقراطية الشعبية
وزارة التعليم العالي والبحث العلمي
جامعة سعيدة د. مولاي الطاهر
كلية التكنولوجيا
قسم: الإعلام الآلي

MEMOIRE DE MASTER

Spécialité : Sécurité informatique et cryptographie

Thème

La vie privée du web sémantique : Anonymiser des données RDF avec SHACL

Présenté par :

Fatima HAMADOUCHE

Dirigé par :

Dr. Ahmed ZAHAF

Année universitaire: 2022- 2023

Remerciement

Au nom d'ALLAH le tous Miséricordieux et que le salut soit sur notre prophète Mohamed aalayh Elssalet wa Elssalem.

Je remercie **ALLAH**, Le Tout Puissant, pour son aide et sa protection, et de m'avoir donné la patience et le courage pour accomplir ce travail.

Ce mémoire est le résultat d'un effort. Cet effort n'aurait pas pu aboutir sans la contribution d'un nombre de personnes. Ainsi se présente l'occasion de les remercier.

Je tiens à remercier en premier lieu, Mr **Zahaf Ahmed** mon directeur de mémoire pour son orientation, encouragement et conseils qui m'ont permis de mener à bien ce mémoire.

J'exprime ma profonde reconnaissance à Mr **Chaibi Hassene** enseignant chercheur au département d'informatique à l'université de Saida, pour ses précieux conseils et d'avoir été tout le temps disponible et à l'écoute.

Je suis très honoré par la présence des membres du jury, qui ont accepté d'être les examinateurs de ce mémoire. Qu'ils trouvent ici, mes plus vifs remerciements pour l'effort qu'ils ont fait pour lire mon manuscrit et l'intérêt qu'ils ont porté à mon travail.

Je remercie ma famille de m'avoir donné le courage d'accomplir ce mémoire.

J'adresse mes sincères remerciements à mes collègues du promotion « SIC 2023 », Pour leurs conseils et soutien continu durant toute l'année.

Enfin, je tiens à remercier qui, de près ou de loin, ont collaboré à l'aboutissement de ce travail.

Dédicace

Je dédie ce modeste travail

Á mes chers parents, que dieu les protège Pour leur amour, leurs encouragements et leurs sacrifices.

Á mon cher frère et mes chères

sœurs

Á mon marie et mes enfants

Ainsi qu'à toute ma famille et amies.

Résumé

L'anonymisation des données du web est un processus crucial pour protéger la confidentialité des informations sensibles tout en permettant leur utilisation à des fins d'analyse et de traitement. SHACL est un langage de contraintes destiné à spécifier des règles et des contraintes sur les données RDF. Il permet de définir des modèles de données, des schémas et des contraintes pour structurer et valider les données RDF.
Nous venons de découvrir qu'Il offre un moyen puissant de valider les données RDF dans le contexte de la vie privée. Le processus de validation avec SHACL (ShapesConstraintLanguage) implique deux étapes clés : définition des contraintes SHACL et exécution de la validation. SHACL est un langage générale de contraintes qui nous a permis d'exprimer des contraintes imposées par des techniques de l'anonymisation des données. Les résultats de la validation permettent d'identifier les problèmes dans les données RDF. En fonction des violations détectées, des actions de sauvetage peuvent être entreprises pour cacher les valeurs divulguées. La validation de la vie privée des données avec SHACL garantit la divulgation des données RDF tout en assurant la confidentialité des données sensibles.

Mot clés : Web Sémantique, Vie privée, RDF, ontologies, Anonymisation, SHACL.

ملخص

يعد إخفاء هوية بيانات الويب عملية حاسمة لحماية سرية المعلومات الحساسة مع السماح باستخدامها للتحليل والمعالجة.
SHACL هي لغة قيد لتحديد القواعد على بيانات RDF . يسمح بتحديد نماذج البيانات و المخططات و القيود الخاصة بهيكل بيانات RDF و التحقق من صحتها.يوفر طريقة قوية للتحقق من صحة بيانات RDF في سياق الخصوصية .تتضمن عملية التحقق من الصحة باستخدام SHACL (Shapse Contraint language) خطوتين رئيسيتين:
تحديد قيود SHACL و إجراء التحقق من الصحة في تحديد المشكلات . SHACL هي لغة تقييد عامة سمحت بالتعبير عن القيود ة التي تفرضها تقنيات إخفاء هوية البيانات . تساعد نتائج التحقق من الصحة في تحديد المشكلات في بيانات RDF . إعتمادا على الإنتهاكات المكتشفة ، يمكن إتخاذ إجراءات الإنقاد لإخفاء القيم المسربة . يضمن التحقق من خصوصية البيانات باستخدام SHACL الكشف عن بيانات RDF مع ضمان سرية البيانات الحساسة.

الكلمات المفتاحية : الويب الدلالي ، الخصوصية ، RDF ، الأنطولوجيا ، إخفاء الهوية ، SHACL .

Abstract

Anonymizing web data is a crucial process to protect the privacy of sensitive information while enabling its use for analysis and processing. SHACL is a constraint language for specifying rules and constraints on RDF data. It allows to define data models, schemas and constraints to structure and validate RDF data.

We just discovered that it offers a powerful way to validate RDF data in the context of privacy. The validation process with SHACL (ShapesConstraintLanguage) involves two key steps: defining the SHACL constraints and performing the validation. SHACL is a general constraint language that allowed us to express constraints imposed by data anonymization techniques. Validation results help identify problems in RDF data. Depending on the violations detected, rescue actions can be taken to hide the leaked values. Validating data privacy with SHACL ensures disclosure of RDF data while ensuring confidentiality of sensitive data.

Key words : Semantic Web, Privacy, RDF, Anonymization, SHACL.

Table des matières

Chapitre I : Le web sémantique

Chapitre II : La vie privé état de l'art

Chapitre III : Anonymisation des données RDF en SHACL

Chapitre V : Implémentation

Liste des Figures

Liste des tableaux

Introduction générale

Introduction

Avec l'émergence du Web Sémantique et l'accroissement de la quantité de données disponibles en ligne, la question de la vie privée est devenue un sujet essentiel. Alors que le Web Sémantique vise à faciliter l'interconnexion des données et à créer des systèmes intelligents, il est important de prendre en compte les implications de cette évolution sur la vie privée des utilisateurs.

La vie privée dans le contexte du Web Sémantique concerne la protection des informations personnelles et la préservation de la confidentialité des utilisateurs lors de l'échange et de l'utilisation de données sémantiques. Le Web Sémantique peut potentiellement recueillir, combiner et analyser des informations provenant de différentes sources, ce qui soulève des préoccupations en matière de vie privée.

L'une des préoccupations majeures est la possibilité de ré-identification des individus à partir de données anonymisées. Même si les données sont anonymisées au départ, des liens et des corrélations peuvent être établis entre différentes sources de données, ce qui peut permettre de retrouver l'identité des individus. Il est donc crucial de mettre en place des mesures pour minimiser les risques de divulgation non autorisée et de ré-identification.

En outre, la collecte et le traitement des données sémantiques peuvent également soulever des questions sur le consentement des utilisateurs. Il est important de s'assurer que les utilisateurs comprennent comment leurs données sont utilisées et d'obtenir leur consentement éclairé pour leur utilisation. Des politiques de confidentialité claires et transparentes doivent être mises en place pour informer les utilisateurs sur la collecte, le stockage, l'utilisation et le partage de leurs données.

La sécurité des données est également un aspect essentiel de la vie privée dans le Web Sémantique. Les données sémantiques peuvent contenir des informations sensibles ou confidentielles, et il est essentiel de mettre en place des mesures de sécurité robustes pour prévenir les violations de données et les accès non autorisés.

Enfin, l'éducation et la sensibilisation des utilisateurs sont des éléments clés pour promouvoir la protection de la vie privée dans le Web Sémantique. Les utilisateurs doivent être informés des risques potentiels et des bonnes pratiques en matière de vie privée, afin de pouvoir prendre des décisions éclairées sur la divulgation de leurs données et de faire valoir leurs droits en matière de confidentialité.

la vie privée dans le Web Sémantique est un enjeu important qui nécessite une attention particulière. Il est essentiel de mettre en place des mesures de protection des données, de garantir le consentement des utilisateurs et d'assurer la sécurité des informations sensibles. L'éducation des utilisateurs et la transparence sont des facteurs clés pour favoriser une utilisation responsable et respectueuse de la vie privée dans le contexte du Web Sémantique.

Contexte

Avec la prolifération croissante des données et des connaissances sur le Web, le Web sémantique offre une approche innovante pour organiser et structurer ces informations de manière à les rendre plus significatives et interconnectées. Cependant, cette interconnexion des données soulève des questions légitimes en matière de vie privée et de sécurité.

Le Web sémantique repose sur des technologies telles que le langage RDF (Resource Description Framework), les ontologies et les langages de requête comme SPARQL pour représenter et manipuler les données de manière sémantique. Il vise à permettre aux machines de comprendre et d'interpréter le contenu des pages web, ce qui ouvre la voie à des applications plus intelligentes et personnalisées.

Cependant, cette puissance sémantique peut également introduire des risques pour la vie privée des utilisateurs. En connectant les données provenant de différentes sources et en reliant les informations sur les individus, le Web sémantique peut potentiellement révéler des détails sensibles et personnels.

L'une des principales préoccupations en matière de vie privée dans le Web sémantique est la collecte et l'utilisation des données personnelles. Lorsque vous interagissez avec des applications et des services basés sur le Web sémantique, vous pouvez fournir involontairement des informations personnelles, telles que vos préférences, votre localisation ou vos habitudes de navigation. Ces données peuvent être agrégées et utilisées pour créer des profils détaillés sur les utilisateurs, ce qui soulève des questions sur le contrôle et la sécurité de ces informations.

En outre, la nature interconnectée du Web sémantique peut rendre difficile l'anonymat des utilisateurs. Lorsque des données sont liées et que des corrélations sont établies entre différentes sources, il devient possible de relier des informations apparemment anonymes à des identités spécifiques. Cela peut avoir des conséquences sur la vie privée des individus, car des informations sensibles peuvent être déduites ou inférées à partir de ces données interconnectées.

Afin de préserver la vie privée dans le Web sémantique, des efforts sont déployés pour développer des mécanismes de protection et des bonnes pratiques. Des approches telles que l'anonymisation des données, la gestion des autorisations et des contrôles d'accès, ainsi que l'éducation des utilisateurs sur les risques et les choix en matière de confidentialité, sont essentielles pour assurer un équilibre entre la puissance du Web sémantique et le respect de la vie privée des individus.

Le Web sémantique offre des possibilités passionnantes en termes d'organisation et d'interconnexion des données. Cependant, il est important de prendre en compte les questions de vie privée et de sécurité qui découlent de cette interconnexion. En adoptant des mesures de protection appropriées et en sensibilisant les utilisateurs aux risques, nous pouvons exploiter pleinement les avantages du Web sémantique tout en respectant la vie privée de chacun.

Problématique et contribution

Le Web sémantique offre des possibilités d'interconnexion et de partage de données plus riches et plus contextuelles, mais cela soulève également des préoccupations majeures en matière de vie privée. Dans ce contexte, la problématique de la protection de la vie privée dans le Web sémantique peut être formulée de la manière suivante :

Comment concilier l'interconnexion sémantique des données et la protection de la vie privée des utilisateurs dans le Web sémantique ?

La contribution de notre étude vise à exprimer et implémenter en SHACL des techniques d’anonymisation des données RDF pour préserver la vie privée dans le contexte spécifique du Web sémantique.

Organisation du manuscrit

En plus de ce chapitre introductif, ce document est composé de quatre grands chapitres, suivis d'une conclusion générale où nous résumons nos principales contributions ainsi qu'un aperçu sur les perspectives de ce travail. Le premier chapitre présente une introduction approfondie sur le Web sémantique, en expliquant ses principes fondamentaux, ses technologies clés (comme le RDF, les ontologies, etc.) et son rôle dans l'échange et l'interconnexion des données sémantiques. Vous pouvez également aborder les avantages et les défis spécifiques liés à l'utilisation du Web sémantique.

Le deuxième chapitre est consacré à l'état de l'art en matière de vie privée, en mettant l'accent sur les concepts clés, les cadres réglementaires et les approches de protection de la vie privée dans divers contextes. Nous explorons également les enjeux spécifiques de la vie privée dans le domaine du Web sémantique, en mettant en évidence les défis et les menaces auxquels sont confrontés les utilisateurs. Le troisième chapitre est le cœur de notre contribution en montrant comment peut-on exprimer l'anonymisation des données RDF en langage de contraintes SHACL. Quant au quatrième chapitre est la mise en œuvre de cette contribution ou on discute les technologies utilisées pour implémenter notre système d'anonymisation des données RDF. Nous achèverons ce mémoire par une conclusion générale.

Chapitre 01 : le web sémantique

Chapitre 1

Le web sémantique

1. Introduction :

Parmi les problèmes majeurs qui influencent négativement sur l'utilisation du Web, est l'aspect statique, actuellement le Web utilisé est purement statique (environ 3 milliards de documents statiques, consultés par près de 300 millions d'utilisateurs à travers le monde), cela signifie que l'aspect dynamique est presque nul dans ce contexte grâce au aspect statique dominant qui repose sur la définition complète de la structure des documents, ou ressources au sens large. D'une façon ou autre, le contenu sémantique des données reste quasi inaccessible aux traitements machines.

Le besoin d'un Web plus puissant, plus efficace, et plus léger fait sans doute la naissance du Web sémantique, donc le futur du Web sera le Web sémantique comme Tim Berners-Lee a mentionné. L'objectif du Web sémantique est de donner une relation forte entre les données qui représente leur sémantique, et de lever les difficultés rencontrées sur le Web d'aujourd'hui (recherche d'information, services,...), en rendant la grande masse d'information disponible, accessible et interprétable par les machines, en plus de l'automatisation de certaines fonctionnalités grâce à la représentation sémantique du contenu des documents, des services, des ressources sur le Web au sens large.

Une représentation explicite de la sémantique des données, programmes, pages html, services Web et autres ressources du Web, permettra d'avoir un Web basé "connaissances", qui fournira à son tour un niveau de service qualitativement meilleur. Pour réaliser cette représentation, les ontologies ont été introduites. Les ontologies représentent le point d'amorçage et la technologie clé pour la réalisation du Web sémantique.

2. Définition du Web sémantique :

Le web sémantique ou toile sémantique est déclaré par son créateur « Tim Berners-Lee » comme la prochaine évolution du Web. Il s'agit d'arriver à un Web intelligent où les informations ne seraient pas seulement stockées mais aussi comprises par les ordinateurs, afin d'apporter à l'utilisateur. ce qu'il cherche vraiment. D'après sa définition, le Web sémantique permettra de rendre le contenu sémantique des ressources Web, compris non seulement par l'homme mais aussi par la machine contrairement au Web actuel qui est vu comme un Web syntaxique.[1]

L'utilisation du terme « sémantique » est trompeuse il s'agit non pas de sens mais de métadonnées traduisant des relations[2]. Le Web sémantique vise donc à produire des relations logiques de données à partir d'autres relations logiques de données, mais ne produit aucun sens, le sens restant le fait de l'interprétation humaine. L'inventeur du World Wide Web et directeur du « W3C » (World Wide Web Consortium) Tim Berners-Lee définit le web sémantique comme « un web de données qui peuvent être traitées directement et indirectement par des machines pour aider leurs utilisateurs à créer des nouvelles connaissances »[1].

3. Architecture du Web sémantique :

L'architecture du web sémantique s'appuie sur une pyramide des langages. Cette pyramide est proposée par Tim Berner Lee , en 2000, pour représenter des connaissances sur le web en satisfaisant les critères de standardisation, d'interopérabilité, et de flexibilité. Cette architecture en couches permet une approche graduelle dans les processus de standardisation et d'acceptation par les utilisateurs.

Ces standards propre au web sémantique sont ouverts et issus du W3C[3] .

Cette architecture est présentée par la figure 1 :

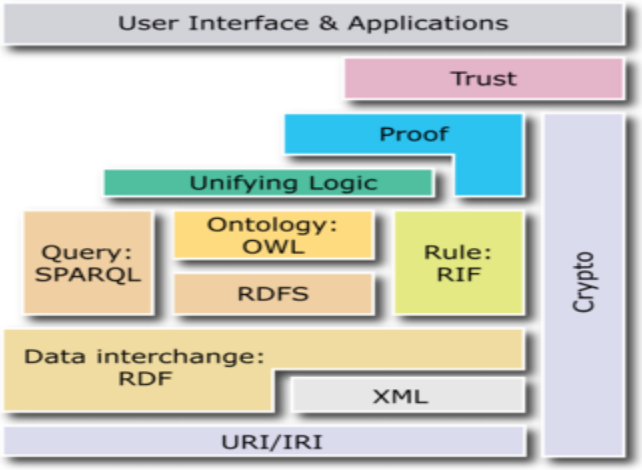


Figure 1. Architecture du web sémantique [4]

3.1. Niveau nommage et adressage

Contient la couche URI et Unicode.

La norme Unicode : est pour but de coder le texte dans toutes les écritures et avec tous les symboles nécessaires à tous les types de textes. Elle est une réponse unique aux problèmes résultants de la multiplicité de codage.

URI (Uniform Ressource Identifier) : L'URI est employé pour désigner le nom ou l'adresse ou les deux d'une ressource dans le Web sémantique [5]. Tout ce qui est disponible sur Internet doit être identifié par un URI. Un URI identifie de manière unique et non ambiguë chaque ressource du Web, comme une page, une adresse email, ou une image.

3.2. Niveau syntaxique :

Le niveau syntaxique est le niveau de la structuration des documents. La spécification de la structure logique des documents repose sur XML. Ce niveau contient la couche XML et espace de noms.

Espace de noms : est un conteneur abstrait contenant des noms, des termes, et des mots qui représentent des objets et des concepts dans le monde réel. Un nom défini dans un espace de noms correspond à un et seulement un objet, deux objets ou Le service web sémantique concepts différents sont référencés par deux noms différents dans un même espace de noms. Ils sont identifiés par une URI .

XML : XML**1** (Extensible Markup Language) fournit une syntaxe pour les documents structurés, mais

n'impose aucune contrainte sémantique à la signification de ces documents.

XML-S : Un XML-S2 (Schéma XML) est une description du type d'un document XML qui contient un ensemble de règles et contraintes sur la structure et le contenu du document aux quelles un document XML doit se conformer afin d'être considéré valide selon ce schéma. XML Schéma sont spécifiquement développés pour exprimer des schémas de XML .

3.3. Niveau sémantique :

Contient les couches restantes, RDF , RDF Schéma , ontologie, règles, cadre logique preuve, confiance, signature et chiffrement. *Couche RDF et RDF-Schéma* :(référencé souvent par RDF(s)) Elle fournit un moyen pour décrire les types des ressources et leurs caractéristiques.

RDF : (Resource Description Framework) est un modèle de métadonnées pour référencer des objets (ressources) et comment se sont reliés l'un à l'autre. Dans ce modèle, les ressources sont identifiées par les URI et nous pouvons faire des déclarations à propos de ces ressources en employant des expressions sous forme «sujet – prédicat – objet », appelées des triplets .

RDF-S : RDF Schéma est une extension sémantique de RDF. Il fournit des mécanismes pour décrire des groupes de ressources similaires et des relations entre ces ressources .

Couche Ontologie :

Est une modélisation du monde réel en concept et relation entre ces concepts. Les principales composantes qu'on peut distinguer sont donc les suivantes : Concept : Est la représentation abstraite des éléments (termes ou classes) du domaine.

Relation :Elles expriment les associations entre les différents concepts.

Fonctions : Il s'agit des relations particulières ou un élément est défini par les n-1 autres éléments [6].

La couche « logique »

S'appuie sur des règles d'inférence qui permettent le raisonnement intelligent exécuté par des agents logiciels.

4. Les Technologies de Base du Web Sémantique

4.1.RDF :

Comme il a été mentionné dans la section précédente, RDF est un élément fondamental du web sémantique puisqu'il permet de représenter des ressources sur le web de manière uniforme pour les agents logiciels là où ceux-ci ne voient dans un document texte qu'une succession de caractères inexploitables [7].

RDF (The Resource Description Framework) est un langage à base d'XML pour décrire des ressources sur le web. RDF est une création pour le traitement des métadonnées. Il fournit l'interopérabilité entre les applications qui échangent de l'information non compréhensible par les machines sur le Web. Il augmente la facilité de traitement automatique des ressources Web.

RDF peut être utilisé dans une variété de champs d'application, dans la découverte de ressources pour fournir une meilleure efficacité aux moteurs de recherche, dans le catalogage pour décrire le contenu et les relations entre les contenus disponibles sur un site Web particulier, sur une page, ou sur une bibliothèque numérique et dans l'évaluation du contenu, en décrivant des ensembles de pages qui représente un simple et unique "document", pour décrire les droits sur la propriété intellectuelle des pages Web, et pour indiquer les préférences de confidentialité (privacy policy) d'un utilisateur aussi bien que les politiques de confidentialité d'un site Web[8] .

Afin de décrire ces ressources, RDF se base sur la notion de triplets, qui permettent de définir des assertions au sujet de celles-ci . Chaque triplet se compose de :

- ✓ Un sujet, qui représente la ressource à laquelle on assigne une propriété, identifiée par une URI (Uniform Ressource Identifier);
- ✓ Un prédicat, désigne l'attribut ou la relation utilisée pour décrire une ressource. Le prédicat est également identifié par une URI;
- ✓ Un objet, c'est la valeur de la propriété. Celle-ci peut être de type primitif (chaîne de caractères, entier ...) ou elle peut être à nouveau une ressource. Elle peut ainsi être à son tour sujet d'un autre triplet conduisant à la formation d'un graphe, les nœuds tout comme les arcs étant représentés par des URIs.

  Par exemple : Soit la phrase simple suivante : la tour Eiffel est créée en1887 a paris .

  Cette phrase est composée des parties suivantes :

Tableau 1 : Le triplet RDF

Sujet (ressource)	Tour Eiffel
Prédicat (propriété)	Créée/située
Objet (littéral)	1887/ Paris

La phrase sera donc représentée par le diagramme suivant :

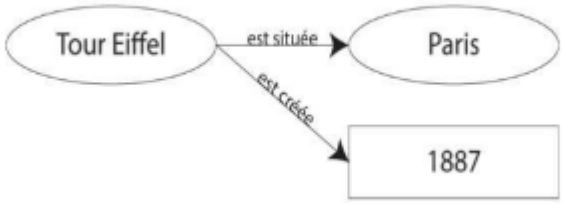


Figure 2- représentation graphique de triplets RDF

Pour représenter correctement ces triplets, il aurait fallu l'écrire de cette manière :

- ✓ (http://dbpedia.org/resource/Eiffel_Tower,http://dbpedia.org/ontology/location, http://dbpedia.org/resource/Paris) ;
- ✓ (http://dbpedia.org/resource/Eiffel_Tower,http://dbpedia.org/property/startDate, 1887)

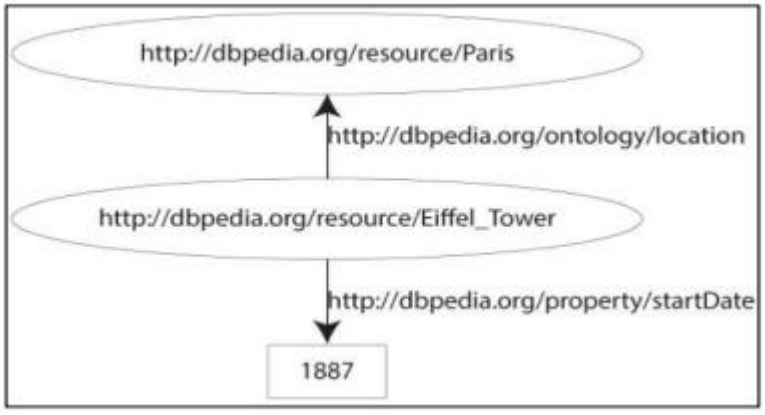


Figure 3- représentation graphique RDF

4.2. Rdfs

Il est nécessaire, de fournir un vocabulaire (classes et propretés) pour donner une sémantique ou un sens aux informations stockées sous forme de triplets RDF. Pour cela, W3C propose une extension de RDF appelée RDFS . Ce dernier, est un langage extensible de représentation des connaissances qui permet de décrire précisément par vocabulaires les ressources d'un domaine donné et les relations entre elles. RDFS fournit des éléments de base pour la définition d'ontologies ou des vocabulaires destinées à structurer les ressources RDF.

L'ensemble de ces ressources appartenant à un même groupe constitue ce que l'on appelle une classe ou une propriété.

- ✓ Une classe est le premier élément important de ce vocabulaire. rdfs (Class) représente le type de la ressource obtenue comme valeur lorsque la propriété rdf (type) est appliquée à une ressource et chaque ressource est une instance de cette classe.
- ✓ Une propriété rdf : type permet d'exprimer l'appartenance d'une ressource à une classe.

4.3.Ontologies

4.3.1.Définition d'ontologie

Le terme ontologie est défini selon deux points de vue :

- ✓ En philosophie : une ontologie signifie science ou théorie de l'être, c'est-à-dire, l'étude des propriétés générales de ce qui existe.
- ✓ En informatique le sens est différent, d'après Studer « Une ontologie est une spécification formelle et explicite d'une conceptualisation partagée. »[9]

En version originale: "An ontology is a formal, explicit specification of a shared conceptualization."

Dans ce contexte, plusieurs définitions des ontologies ont été proposées est la plus fiable, déclare

que « une ontologie définit les termes et les relations de base du vocabulaire d'un domaine ainsi que les règles qui permettent de combiner les termes et les relations afin de pouvoir étendre le vocabulaire» .

L'ontologie constitue en soi un modèle de données représentatif d'un ensemble de concepts dans un domaine, ainsi que des relations entre ces concepts. Elle est employée pour raisonner à propos des objets du domaine concerné. Plus simplement, on peut aussi dire que l' «ontologie est aux données ce que la grammaire est au langage».

4.3.2. Composants d'une ontologie :

Une ontologie est l'ensemble structuré des termes et des concepts représentant le sens d'un champ d'informations. Une ontologie constitue en soi un modèle de données représentatif d'un ensemble de concepts dans un domaine, ainsi que les relations entre ces concepts. Elle est employée pour raisonner à propos des

objets du domaine concerné. Ainsi une ontologie représente trois types de caractéristiques :

- ✓ Les concepts (les classes)
- ✓ Les propriétés
- ✓ Les relations

Les concepts :

Les classes représentent les concepts pris au sens large. Un concept peut être tout ce qui peut être évoqué : description d'une tâche, d'un objet matériel, d'une fonction, d'une action, d'une stratégie ou d'un processus de raisonnement, etc.

Les propriétés :

Une ontologie est, non seulement le repérage et la classification des concepts mais c'est aussi des caractéristiques qui leur sont attachées et qu'on appelle les propriétés (appelées, aussi, les attributs).

Par exemple : Dans une ontologie sur le monde animal, on pourra avoir les concepts de « mammifère » ou d' « oiseau » pour lesquels est précisé le type de tégument, respectivement à poil et à plume. En pratique, un attribut « tégument » pourra être attaché aux concepts et sa valeur variera suivant le concept auquel on fait référence.

Les relations :

Les relations représentent un type d'association entre les concepts du domaine.

4.3.3. Rôles des ontologies :

Les ontologies permettent d'améliorer la communication entre machines et entre humains et machines ou encore entre humains par le biais de logiciels. Les propriétés de ce type de structure de données ont permis de diversifier leur utilisation à différentes applications, en particulier la gestion des connaissances et le web sémantique. Les ontologies permettent de :

•Définir un vocabulaire commun et partagé par une communauté de pratiques.

•Avoir une compréhension commune.

•Donner un sens unique à des entités du monde réel.

•Construire et découvrir de nouvelles informations et/ou connaissances à partir des ontologies et des ressources existantes.

•Améliorer les processus de recherche d'informations.

• Faciliter la communication entre agents logiciels. Le langage d'ontologie le plus largement utilisé est le «Web Ontology Language», Qui a curieusement l'acronyme «OWL».

4.3.4. OWL

Le langage OWL a été conçu pour être le langage d'ontologie Web. OWL fait partie du paquet croissant des recommandations liées au Web sémantique. OWL est, tout comme RDF, un langage XML profitant de l'universalité syntaxique de XML. Il est Fondé sur la syntaxe de RDF/XML, OWL offre un moyen d'écrire des ontologies web. OWL intègre des outils de comparaison des propriétés et des classes : identité, équivalence, contraire, cardinalité, symétrie, transitivité, disjonction, etc. OWL offre aux machines une plus grande capacité d'interprétation du contenu web que RDF et RDFS, grâce à un vocabulaire plus large et à une vraie sémantique .

Le langage OWL offre trois sous-langages d'expression croissante destinés à des communautés de développeurs et d'utilisateurs spécifiques :

OWL Lite : est le sous langage de OWL le plus simple. Il est destiné aux utilisateurs qui ont besoin d'une hiérarchie de concepts simple. OWL Lite est adapté, par exemple, aux migrations rapides depuis d'anciens thésaurus .

OWL DL : est plus complexe que OWL Lite, permettant une expressivité bien plus importante. OWL DL est fondé sur la logique descriptive (d'où son nom, OWL Description Logics), un domaine de recherche étudiant la logique, et conférant donc à OWL DL son adaptation au raisonnement automatisé. Malgré sa complexité relative face à OWL Lite, OWL-DL garantit la complétude des raisonnements (toutes les inférences sont calculables) et leur décidabilité (leur calcul se fait en une durée finie) .

OWL Full : est la version la plus complexe d'OWL, mais également celle qui permet le plus haut niveau d'expressivité. OWL Full est destiné aux situations où il est plus important d'avoir un haut niveau de capacité de description. OWL Full offre cependant des mécanismes intéressants, comme par exemple la possibilité d'étende le vocabulaire par défaut de OWL .

4.4.Langage d'interrogation d'ontologie (SPARQL) :

A nos jours le stockage d'information est de quantité énorme et de manière structurée et il n'aurait aucune utilité s'il était impossible d'y accéder. Dans le cas du Web sémantique il existe des langages de requêtes connus tels que RDQL1, nRQL2, SPARQL. SPARQL est le langage standard d'interrogation du Web sémantique, plus particulièrement de graphe RDF.

SPARQL (acronyme obscur dérivé de SQL) est un langage de requêtes destiné à interroger les bases de données (fichiers) RDF, standardisé par le W3C et implémenté en divers langages, notamment en Java avec le framework JENA. La syntaxe du SPARQL ressemble à celle du SQL car elle a été inspirée de ce dernier.

• SPARQL permet d'exprimer deux types de requêtes:

Interrogative : extraction par SELECT des sous graphes d'un graphe RDF, satisfaisant les conditions spécifique dans la clause WHERE.

Constructive : génération d'un autre graphe à partir du graphe résultant de la requête d'interrogation.

• La syntaxe d'une requête SPARQL est définie de plusieurs parties suivantes :

PREFIX : définit une étiquette correspondant à l'URI d'un dépôt de données à interroger et permettent de définir des espaces de noms.

SELECT : identifie les valeurs à retourner parmi les variables de la clause WHERE.

ASK : interroge un dépôt ne retourne que true/false détermine la présence des résultats.

CONSTRUCT : crée un nouveau graphe RDF à partir des résultats de la requête.

DESCRIBE : retourne un graphe RDF décrivant les ressources trouvées lors d'une interrogation.

FROM : identifie les sources de données à interroger.

WHERE : associe les valeurs à retourner avec les objets du fichier RDF.

DISTINCT : garantit l'unicité des résultats.

REDUCED : supprime toutes les solutions présentent plus d'une fois dans les résultats.

FILTER : ajoute des contraintes sur les résultats.

OPTIONAL : permet de définir l'instruction suivante comme optionnelle.

UNION : combine les résultats de plusieurs instructions.

ORDER BY : trie les résultats.

LIMIT : limite le nombre de résultats.

OFFSET : fait commencer les résultats générés après le nombre de résultats indiqué.

Exemple :

```
SELECT ?personne ?age
WHERE { ?personne rdf :type
Personne.
?personne ns : Age ?age.
FILTER (?age >= 23) }
```

Résultat

Personne	Age
Sarra	23

La présence de FILTER filtre le résultat obtenu en introduisant une contrainte (? age >= 23) et le résultat obtenu de la requête donne toutes les personnes ayant l'âge supérieur ou égal à 23[10].

4.5. SHACL (Shapes Constraint Language) :

SHACL est une spécification du W3C (World Wide Web Consortium) permettant de valider des graphes RDF avec un ensemble de conditions. SHACL comprend, entre autres, des fonctionnalités permettant d'exprimer des conditions qui limitent le nombre de valeurs qu'une propriété peut avoir, le type de celles-ci, les plages numériques, les modèles de correspondance de chaîne et les combinaisons logiques de certaines contraintes. SHACL inclut également un mécanisme d'extension permettant d'exprimer des conditions plus complexes dans des langages tels que SPARQL .

SHACL permet d'exprimer différents types de contraintes de données pour valider des données RDF. Voici une liste non exhaustive de ces contraintes[11] :

- Contrainte de nœud : vérifie si un nœud existe dans le graphe de données.
- Contrainte de propriété : vérifie si une propriété spécifique existe sur un nœud.
- Contrainte de valeur : vérifie si une valeur spécifique existe sur un nœud.
- Contrainte de cardinalité : vérifie si un nombre spécifique de valeurs existe sur un nœud.
- Contrainte de type de données : vérifie si un type de données spécifique est utilisé pour une valeur.

Contrainte de nœud

Pour exprimer une contrainte de noeud en SHACL, on peut utiliser la contrainte de noeud (Nodeconstraint). Cette contrainte permet de restreindre les noeuds qui sont valides pour une forme donnée. On peut ainsi restreindre les noeuds qui sont des instances d'une classe donnée en utilisant la propriété sh:targetClass. Par exemple, pour restreindre les noeuds qui sont des instances de la classe "Person", on peut utiliser la contrainte de noeud avec la propriété sh:targetClass:Person. Voici un exemple de contrainte de noeud en SHACL :

```
:PersonShape
ash:NodeShape ;
sh:targetClass :Person ;
sh:property [
sh:path :name ;
sh:minCount 1 ;
] .
```

Contrainte de propriété

Pour exprimer une contrainte de propriété en SHACL, on peut utiliser la propriété sh:path. Cette propriété permet de spécifier le chemin de propriété qui doit être validé. On peut ainsi vérifier si une propriété existe sur un nœud donné en utilisant la contrainte de propriété. Par exemple, pour vérifier si un nœud a une propriété "name" avec au moins une valeur, on peut utiliser la contrainte de propriété avec la propriété sh:path:name et la propriété sh:minCount 1. Voici un exemple de contrainte de propriété en SHACL :

```
:PersonShape
ash:NodeShape ;
sh:targetClass :Person ;
sh:property [
sh:path :name ;
sh:minCount 1 ;
] .
```

Contrainte de valeur

Pour exprimer une contrainte de valeur en SHACL, on peut utiliser la propriété sh:hasValue. Cette propriété permet de spécifier la valeur attendue pour une propriété donnée. On peut ainsi vérifier si une propriété a une valeur spécifique en utilisant la contrainte de valeur. Par exemple, pour vérifier si une personne a un âge de 20 ans, on peut utiliser la contrainte de valeur avec la propriété sh:hasValue 20 et la propriété sh:path:age. Voici un exemple de contrainte de valeur en SHACL :

```
:PersonShape
ash:NodeShape ;
sh:targetClass :Person ;
sh:property [
sh:path :age ;
sh:hasValue 20 ;
] .
```

Contrainte de cardinalité.

Pour exprimer une contrainte de cardinalité en SHACL, on peut utiliser les propriétés sh:minCount et sh:maxCount. La propriété sh:minCount permet de spécifier le nombre minimum de valeurs attendues pour une propriété, tandis que la propriété sh:maxCount permet de spécifier le nombre maximum de valeurs attendues pour une propriété. Voici un exemple de contrainte de cardinalité en SHACL :

```
:PersonShape
ash:NodeShape ;
sh:targetClass :Person ;
sh:property [
sh:path :address ;
sh:minCount 1 ;
sh:maxCount 2 ;
] .
```

Cette contrainte s'applique à toutes les instances de la classe ":Person" et vérifie que la propriété ":address" a au moins une valeur et pas plus de deux valeurs. Cette contrainte permet de s'assurer que les données RDF respectent une structure ou un format spécifique.

Contrainte de type de données

Pour exprimer une contrainte de type de données en SHACL, on peut utiliser la propriété sh:datatype. Cette propriété permet de spécifier le type de données attendu pour une propriété donnée. On peut ainsi vérifier si une propriété a une valeur qui correspond à un type de données spécifique en utilisant la contrainte de type de données. Par exemple, pour vérifier si une personne a un âge qui est un entier, on peut utiliser la contrainte de type de données avec la propriété sh:datatypexsd:integer et la propriété sh:path:age. Voici un exemple de contrainte de type de données en SHACL :

```
:PersonShape
ash:NodeShape ;
sh:targetClass :Person ;
sh:property [
sh:path :age ;
sh:datatypexsd:integer ;
] .
```

SHACL permet également d'exprimer des contraintes plus complexes impliquant plusieurs nœuds ou propriétés, ainsi que de définir des contraintes personnalisées. Cependant, il ne peut pas exprimer toutes les contraintes qui peuvent être exprimées en utilisant OWL et un raisonneur [12].

5.Conclusion :

La manipulation des ressources du Web par des machines requiert l'expression ou la description de ces ressources. Plusieurs langages ont été définis à cet effet, ils permettent d'exprimer données et méta-données, de décrire les services Web et leur fonctionnement, et de disposer d'un modèle abstrait de ce qui est décrit grâce à l'expression d'ontologies (RDFS, OWL). Ces langages sont la base pour la réalisation du Web sémantique. Cependant, il reste beaucoup de choses à faire pour standardiser le tout, mais comme l'écrit, Tim Berners-Lee : « *le Web sémantique est ce que nous obtiendrons si nous réalisons le même processus de globalisation sur la représentation des connaissances que celui que le Web fit initialement sur l'hypertexte* ».

Chapitre 2 : La vie prive Etat de l’art

Chapitre 2

La vie privée : Etat de l'art

1. Introduction :

Les services web ont récemment émergés comme un moyen populaire pour la publication et le partage des données sur le web. Les entreprises modernes à travers tous les spectres s'orientent vers une architecture basée sur la mise des bases de données derrière des services web, fournissant ainsi un bien documenté, une plate-forme indépendante et une méthode interopérable d'interaction avec leurs données.

Par ailleurs, dans notre civilisation, la protection de la vie privée est considérée comme l'une des libertés individuelles fondamentales. Dans le monde réel, cette protection repose sur des lois, sur des difficultés matérielles et sur le coût de collecte d'informations qui porteraient atteinte à la vie privée des individus. En revanche, dans le monde virtuel de l'Internet, une telle collecte est à la fois facile et peu coûteuse.

D'autre part, cette vie privée des services est considérée comme un élément à trois dimensions pour les infrastructures des services web. Ces dimensions sont : la vie privée de l'utilisateur (telle que perçue par l'utilisateur d'un service web), la confidentialité des données (soutenue au niveau des données), et la vie privée des services (comme exposée aux utilisateurs à travers un service web). L'exécution des conditions de ces trois niveaux d'intimité lance un certain nombre de défis, par exemple: l'exécution de l'intimité au taux de disponibilité, peut exiger une manipulation complexe et dynamique sur les conditions de confidentialité des données utilisées par les individus.

2. Définition vie privée (privacy en anglais) :

«Le droit d'une personne de contrôler l'accès à sa personne et aux renseignements qui la concerne. Le droit à la vie privée signifie que la personne décide des renseignements qui sont divulgués, à qui et à quelles fins.»[13]

En d'autres termes, les questions de protection des renseignements personnels s'appuient sur l'utilisation abusive de données existaient et pas Analytics algorithmes. Cette section adressée au débat sur le contexte et les contours du mot « vie privée ».

3. Autre Définition :

Le concept de vie privée est un concept assez général et assez flou et difficile à cerner toutes ces facettes. Cela à répercuter sur les multiples définitions dans la littérature. L'une des premières définitions a été donnée par A.Westin. Cette définition a associé le terme de vie privée (ou protection de la vie privée) à la revendication d'un individu, d'un groupe ou d'une institution, à pouvoir contrôler les conditions d'acquisition et d'utilisation de leurs informations personnelles. Il s'agit également de savoir quand ces informations seront obtenues et quelles utilisations en seront faites par d'autres. Bien que cette définition a été introduite au début pour un monde hors-ligne, elle est aujourd'hui largement adoptée et étendue par plusieurs académiques pour définir la vie privée dans le monde numérique. D'autres chercheurssont classifié les le concept de vie privée en six catégories : (1) le droit d'être laissé seul ; (2) le secret; (3) l'intimité ; (4) Avoir le contrôle sur l'information ; (5) L'accès restreint au soi ; et (6) un regroupement de plusieurs catégories. La définition de la vie privée comme étant « le droit d'être laissé seul (ou tranquille) » a été donnée la première fois en 1890 par Warren & Brandeis comme réponse à l'utilisation intense de la photographie. Cette définition a été largement critiquée comme étant trop vague. En plus, ce n'est pas chaque violation du droit d'être laissé seul est une violation dela vie privée. L'exemple le plus simple c'est le frôlement de deux personnes sur un trottoir occupé. [14]

L'une des définitions de la vie privée qui tourne au tour du secret, a été donnée par Richard Posner, il a défini la vie privée comme : « le droit de dissimuler des faits discréditables sur soi-même ». Cette définition a été critiquée, du fait que les informations secrètes ne sont pas toujours privées (par exemple, des plans militaires secrets) et aussi, les affaires privées ne sont pas toujours secrètes (par exemple, les dettes d'une personne). Une opinion très répandue qui dit que la vie privée et l'intimité sont étroitement liées. Dans ce cadre, Julie Inness a défini la vie privée comme : « l'état de l'agent ayant le contrôle sur les décisions concernant des sujets qui tirent leur signification et leur valeur de l'amour, la compassion, ou les goûts,de cet agent ».

Cette définition a été critiquée, du fait qu'il est possible d'avoir des relations privées sans elles soient intimes et aussi d'accomplir des actes privés qui ne sont pas intimes.

Avoir le contrôle sur les informations personnelles a également été proposé comme définition de la vie privée. Dans ce contexte, Alan Westin, a donné la définition suivante : « La vie privée n'est pas simplement une absence d'information sur nous dans l'esprit des autres ; c'est plutôt le contrôle que nous avons sur cesinformations ». Les critiques sur cette définition

se basent principalement sur le fait que cette définition ne couvre pas l'aspect physique de la vie privée, comme le contrôle de l'accès aux lieux et aux corps. Parmi les définitions de la vie privée comme «l'accès limité au soi », celle donnée par S. Bok. Il a défini la vie privée comme : «la condition d'être protégé contre tout accès non désiré par d'autres personnes ». Malgré que cette définition couvre aussi bien l'aspect numérique que physique, elle a été critiquée d'être trop étroites. Enfin, beaucoup considèrent la vie privée en tant que concept de cluster qui comprend plusieurs des dimensions mentionnées ci-dessus. Par exemple J. DeCew a proposé que la vie privée est un concept qui s'étend sur l'information, l'accès et les expressions. Moore a argué que la vie privée est le droit de contrôler l'accès et l'utilisation des informations personnelles et les localisations spatiales.

4. Vie privée et Législation

L'importance de la législation relative à la protection de la vie privée réside dans le fait qu'elle constitue l'élément clé pour le développement du commerce et de la collaboration électronique à travers le monde. La figure 4 montre l'existence ou l'absence des lois de protection de données personnelles dans les pays du monde, ainsi que la compatibilité de ces lois avec celles de l'union Européenne. [15]

Cette section survole les principales lois relatives à la protection de la vie privée à travers le monde. Nous nous intéressons en particulier par : La loi islamique, la loi algérienne, les lois de l'union européenne, la loi internationale évoqué par les Principes directeurs de l'OCDE et enfin les Initiatives HIPAA des États-Unis.

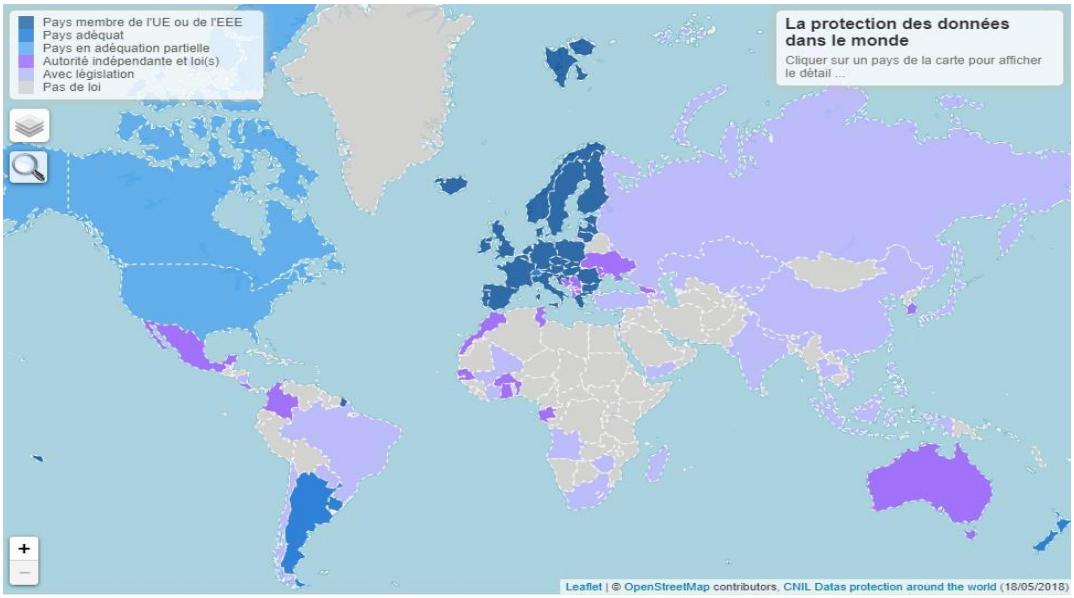


FIGURE 4– La protection des données dans le monde.

5. Vie privée Dans l'islam (Shariah)

L'Islam accorde une grande importance au droit humain à la vie privée. Cette importance, couvre différents aspects de la vie privée : l'aspect physique, l'accès et l'utilisation des informations personnelles, l'intimité et réputation. Cela est répercuté dans plusieurs versets du Coran et de le Hadith. Dans l'aspect physique de la vie privée, l'islam a interdit tout accès non autorisé à des propriétés privées, dans ce cadre, Allah a dit :"ô vous qui croyez ! N'entrez pas dans des maisons autres queles vôtres avant de demander la permission et de saluer leurs habitants. Cela est meilleur pour vous. Peut-être vous souvenez-vous"[1].

Le Prophète (paix et bénédiction de Dieu sur lui) est allé jusqu' à dire qu'un homme ne devrait pas entrer, même danssa propre maison soudainement ou discrètement[2].

L'interdiction d'acquérir ou de divulguer des informations personnelles sans autorisation est claire dans le verset coranique suivant :"ô vous qui avez cru ! évitez de trop conjecturer [sur autrui] car une partie des conjectures est péché. Et n'espionnez pas ; et ne médisez pas les uns des autres. L'un de vous aimerait-il manger la chair de son frère mort? (Non !) vous en aurez horreur. Et craignez Allah. Car Allah est Grand Accueillant au repentir, Très Miséricordieux"3. Il y a aussi des peines encas d'atteinte à la réputation des personnes (et spécialement les femmes), dans ce contexte, Allah a dit : "Et ceux qui lancent des accusations contre des femmes chastes sans produire par la suite quatre témoins, fouettez es de quatre-vingts coups de fouet, et n'acceptez plus jamais leur témoignage. Et ceux-là sont les pervers. [15]

1. Sourate Al-Nur, verset 27.
2. Sahih Muslim, Hadith N 3555.

6. La loi algérienne

Comme la figure 4 le montre, et jusqu'à l'écriture de ces lignes 5, l'Algérie ne dispose pas d'une loi en vigueur pour la protection de la vie privée et les données à caractère personnel. Le 27 décembre 2017, le conseil des ministres a adopté un projet de loi relatif à la protection des personnes physiques dans le traitement des données à caractère personnel. Selon le communiqué à l'issue de la réunion du conseil des ministres, ce nouveau texte "accompagnera le développement du traitement numérique des données administratives, juridiques et financières, dans des secteurs de plus en plus nombreux du service public" et "régulera la protection des personnes physiques lors du traitement de leurs données à caractère personnel". le projet de loi énonce notamment "l'exclusion des données à usage privé exclusif du traitement en l'objet, la nécessité de l'accord de la personne concernée lors du traitement de ses données personnelles, sauf dans des situations d'obligations légales, essentiellement judiciaires, l'institution d'une protection renforcée pour la protection des données personnelles de l'enfant et l'institution d'une Autorité nationale de protection des données à caractère personnel, placée auprès du Président de la République" .

7.Les risques relatifs à la vie privée

Au-delà de l'e-commerce, le simple fait de partager vos informations en ligne estune source de préoccupation pour de nombreux internautes. Une partie de l'échange est volontaire, comme sur les réseaux sociaux, et une partie est involontaire, comme lorsque votre information est vendue sur des réseaux de publicité en ligne. Sur les réseaux sociaux, par exemple, vous avez peut-être volontairement partagé votre domicile, votre âge, votre sexe, et vos intérêts personnels sur votre page Facebook, mais vous avez peut-être par inadvertance laissé vos paramètres de confidentialité publics. Si c'est le cas, les réseaux de publicité en ligne peuvent avoir déduit une grande partie de ces informations, en fonction des sites Web que vous visitez et des recherches que vous effectuez.

7.1.Le profilage :

Chaque fois qu'un client interroge un site Web, quelqu'un, quelque part suit sonactivité en temps réel. Ce profilage permet de collecter des informations sur l'utilisateur et se pratique sur plusieurs sites, toujours à l'inconnu du visiteur ou sans son consentement, et cela soumettre un risque d'atteinte à la vie privée du fait qu'ilpermet d'analyser avec précision le comportement des internautes.

Le profilage est une technique de surveillance ou d'exploitation des données qui permet d'établis différentes actions, mesures ou décisions touchant les personnes concernées dans le

cadre de finalités diverses. Les techniques de profilage représentent un intérêt important pour l'économie ou pour les administrations publiques ; elles peuvent aussi avoir des effets bénéfiques pour les personnes concernées, par exemple dans le domaine de la santé. Cependant, elles génèrent également des conséquences négatives sur le respect des droits et des libertés fondamentales, notamment le droit à la vie privée et à la protection des données.

7.2.Vol d'identité :

Le vol d'identité renvoie au processus initial consistant à acquérir les données personnelles d'une personne à des fins criminelles. Les voleurs d'ID s'approprient des éléments clés de renseignements personnels d'une personne de manière physique ou autrement, sans avertir et ils les utilisent pour usurper l'identité et commettre des crimes en nom d'une personne. En plus des noms, des adresses et des numéros de téléphone, les voleurs d'identité recherchent : les numéros d'assurance sociale, les numéros de permis de conduire, les renseignements sur les cartes de crédit et les renseignements bancaires, les cartes bancaires, les cartes d'appel, les certificats denaissance et les passeports. [16]

Les voleurs d'identité peuvent manipuler les renseignements d'une personne.Ils peuvent utiliser les identités volées pour faire des achats extravagants, ouvrir de nouveaux comptes bancaires, détourner le courrier, présenter des demandes d'emprunt, de cartes de crédit et de prestations sociales, louer des appartements et même commettre des crimes beaucoup plus graves.

8. Différents niveaux de protection de la vie privée

On peut définir quatre propriétés principales pour la protection de la vie privée : L'anonymat, pseudonymat, Non-"chaînabilité" et Non-observabilité. [17]

- **Anonymat** : Requiert que d'autres utilisateurs ou sujets soient incapables de déterminer le véritable nom de l'utilisateur associé à un sujet, une opération ou un objet.
- **Pseudonymat** : C'est l'utilisation d'un pseudonyme au lieu du vrai nom.
- **Non-chaînabilité** : C'est l'impossibilité pour d'autres utilisateurs d'établir un lien entre les différentes opérations faites par un même utilisateur.
- **Non-observabilité** : Consiste à ce que des utilisateurs ou des sujets nepuissent pas déterminer si une opération est en cours d'exécution.

9. Les principes fondamentaux de protection de la vie privée :

Il existe quelques principes universels liés au respect de la vie privée comme la minimisation des données, souveraineté des données, consentement explicite et la transparence.

9.1. Minimisation des données

La première mesure de minimisation consiste que la seule information nécessaire pour compléter une application particulière devrait être collectée ou utilisée et pas plus[18], par exemple le commerce électronique, impliquant un client, un marchand, un service de livraison, des banques. Le marchand n'a pas besoin en général de l'identité du client, mais doit être sûr de la validité du moyende paiement. La société de livraison n'a pas besoin de connaître l'identité de l'acheteur, ni ce qui a été acheté (sauf les caractéristiques physiques), mais doit connaître l'identité et l'adresse du destinataire. La banque du client ne doit pas connaître le marchand ni ce qui est acheté, seulement la référence du compte à créditer, le montant, etc [19].

9.2. Souveraineté des données

Lorsque des données personnelles se trouvent sur un site distant, c'est-à-dire une machine qui n'est pas sous le contrôle direct de la personne concernée (typiquement, un serveur d'une entreprise ou administration), soit pour un court moment (par exemple l'exécution d'une simple transaction), soit pour plus longtemps (par exemple des dossiers médicaux dans un hôpital), l'accès à ces données devrait être strictement limité à l'usage souhaité par leur propriétaire, c'est-à-dire la personne correspondant à ces données. Cela signifie que le propriétaire des données doit pouvoir imposer une politique de protection de la vie privée sur ses données et que le serveur qui conserve et traite ces données doitmettre en œuvre cette politique par des mécanismes de contrôle des accès à ces données. La politique en question peut définir des permissions et des interdictions précisant qui peut ou ne peut pas réaliser quelle opération sur ces données personnelles, mais aussi des obligations précisant, par exemple, que les données expirent (et donc doivent être effacées) après un délai donné suivant la terminaison de la transaction, ou que la divulgation de ces données à un tiers doit être notifiée au propriétaire par courriel, etc. Bien sûr, la politique de vie privée imposée par le propriétaire des données doit être compatible avec la politique de sécurité qui protège les biens de l'entreprise et gouverne l'exécution de l'application, et donc les accès effectifs aux données. La compatibilité entre ces deux politiques doit être vérifiée avant la divulgation par l'utilisateur de ses données personnelles [20].

9.3. Consentement explicite

Ça signifie qu'avant de collecter les données personnelles d'un individu, il faut luidemander son autorisation et lui expliquer comment elles seront utilisées.

9 .4.Transparence

Ça signifie que le système ne doit pas être considère comme une boite noire dans laquelle l'individu doit avoir une confiance aveugle.

10. Modèles de protection de la vie privée

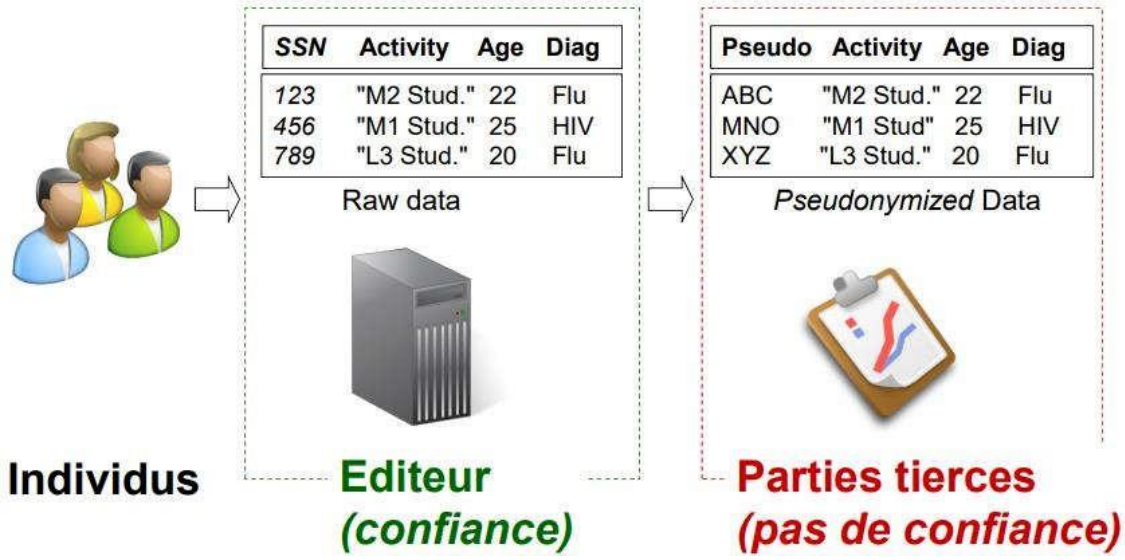
Les efforts de recherche consacrés à la protection de la vie privée ont donné naissance à plusieurs modèles et variantes de modèles. À titre d'exemple, (B. C.M. Fung et al. 2010) recense non moins de quinze modèles. Dans cet article, nous allons décrire cinq types de modèles d'anonymisation, qui cherchent à cacher ou briser le lien existant entre une personne du monde réel, et ses données sensibles : la pseudonymisation, le *k*-anonymat, la *l*-diversité, la *t*-proximité et la differential privacy.

10.1. La pseudonymisation

La pseudonymisation consiste à supprimer les identifiants explicites et leur remplacement éventuel par des pseudonymes c.à.d. à rajouter à chaque enregistrement un nouveau champ, appelé *pseudonyme*. Pour créer ce pseudonyme, on utilise souvent une *fonction de hachage* que l'on va appliquer à l'un des champs identifiants (par exemple le numéro de sécurité sociale), qui est un type de fonction particulier qui rend impossible (ou tout du moins extrêmement difficile) le fait de déduire la valeur initiale. On voit ainsi que deux entités possédant des informations sur une même personne, identifiée par son numéro desécurité sociale, pourraient partager ces données de manière anonyme en *hachant* cet identifiant. Il est également possible d'utiliser tout simplement une fonction aléatoire pour générer un identifiant unique pour chaque personne, mais nous verrons plus bas que cela ne résout pas tous les problèmes.

Le gros avantage de la pseudonymisation est qu'il n'y a aucune limite sur le traitement subséquent des données. Tant que l'on traite des champs qui ne sont pas directement identifiants, on pourra exécuter exactement les mêmes calculs qu'avec une base de données non-anonyme. Ainsi, on montre dans la **Figure 5** un exemple de calcul de la moyenne d'âge pour une pathologie donnée. L'utilisation de données pseudonymisées ne nuit pas à ce calcul.

Figure 5: Pseudonymisation et exemple de calcul



Toutefois, la pseudonymisation n'est pas reconnue comme un moyen d'anonymisation, car la pseudonymisation rend vulnérables les enregistrements dans lesquels une partie de la donnée permet de réidentifier l'individu concerné : la combinaison d'autres champs peut permettre de retrouver l'individu concerné. Sweeney[6] l'a mis en évidence aux Etats-Unis en 2001 en croisant deux bases de données, une base de données médicale pseudonymisée et une liste électorale avec des données nominatives. Le croisement a été effectué non pas sur des champs directement identifiants, mais sur un triplet de valeurs : code postal, date de naissance et sexe, qui est unique pour environ 80% de la population des Etats- Unis ! Elle a ainsi pu relier des données médicales à des individus (en l'occurrence le gouverneur de l'Etat). [21]

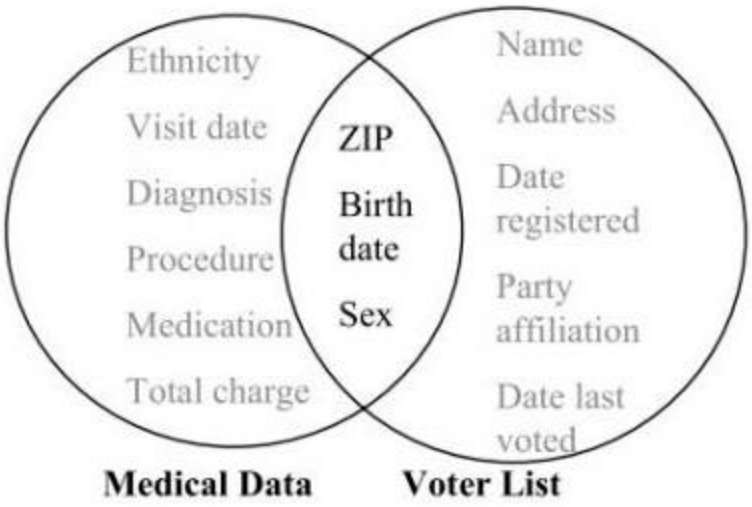


Figure 6: Un exemple de recoupement d'une base anonyme [21]

10.2.Le *k*-anonymat

Le modèle de k-anomymat est le premier modèle proposé dans la littérature. Lorsqu'il est mis en œuvre, ce modèle offre l'assurance que chaque n-uplet de valeurs de quasi-identifiants apparaît au moins k fois dans la table à publier. Ainsi, une table qui répond à ce type de modèle est dite «k-anonyme» , **Sweeney** a proposé la technique de *k*-anonymat. Celle-ci va flouter la possibilité de lier un *n*-uplet anonyme à un *n*-upletnon anonyme de la manière suivante :

- ✓ déterminer les ensembles d'attributs (appelés *quasi*identifiants) qui peuvent être utilisés pour croiser les données anonymes avec des données identifiantes ;
- ✓ réduire le niveau de détail des données de telle sorte qu'il y a au moins *k n*-uplets différents qui ont la même valeur de *quasi*-identifiant, une fois celui-ci généralisé (on dit alors que les individus font partie de la même classe *d'équivalence*). « Généraliser » signifie en fait « enlever un degré de précision » à certains champs. Ainsi, il est impossible d'êtresûr à plus d'une chance sur *k* qu'on a bien lié un individu donné avec son *n*-uplet anonyme. L'avantage du *k*-anonymat est que l'analyse des données continue de fournir des résultats exacts, à ceci près qu'on ne peut pas dissocier les individus d'un groupe.

Toutefois, Certains cas peuvent aussi apparaître : si tous les individus d'une classe d'équivalence possèdent les mêmes valeurs sur un champ intéressant l'attaquant, alors celui-ci sera capable d'identifier cette valeur.

10.3 La *l*-diversité

L'objectif de la l-diversité est de s'assurer que chaque groupe k-anonyme est également assez divers ou assez varié, c'est-à-dire avec suffisamment de valeurs sensibles distinctes à l'intérieur, ceci afin d'empêcher la déduction par homogénéité de la valeur sensible d'une personne. Chaque classe de k individus doit donc être associée à au moins l valeurs sensibles dites bien représentées. Celaveut dire qu'il faut avoir suffisamment de valeurs différentes, ou bien suffisamment de valeurs qui apparaissent souvent du point de vue statistique dans l'ensemble de la population.

10.4. La *t*-proximité

Bien qu'une table puisse être protégée par le principe de la l-diversité, il est possible pour un adversaire d'obtenir des informations au sujet d'un attribut sensible dès lors qu'il dispose d'informations sur la distribution globale de cet attribut.Sachant qu'il est impossible d'empêcherun adversaire d'avoir des informations sensibles globales sur une population, le principe de la t-proximité a pour objectif de limiter la capacité de cet adversaire à déduire des informations sensibles sur des individus ciblés. Pour ce faire, il fait en sorte que la distribution de l'attribut sensible au sein de n'importe quelle classed'équivalence soit proche de la distribution globale de l'attribut. En d'autres termes, il introduit le concept de distance entre ces deux distributions et propose que cette distance ne dépasse pas le seuil t. Ainsi, plus t est petit, plus la possibilité d'inférence de l'adversaire est réduite.

Age	Sexe	Département	Pathologie	Nombre d'individus
<45	M	75	Grippe	400
<45	M	75	Rhume	800
>45	M	75	Grippe	500
>45	M	75	Rhume	1000
<35	F	75	Grippe	300
<35	F	75	Rhume	600
>35	F	75	Grippe	600
>35	F	75	Rhume	1200
...				

Figure 7:t-proximité

La t-proximité souffre de plusieurs problèmes, le plus important étant sans douteson utilité ! En effet, il parait évident d'exploiter des données *k*-anonymes ou même *l*-diverses pour découvrir des corrélations entre des données appartenant auquasi-identifiant et des données sensibles. Toutefois, le but même de la t- proximité est de réduire au maximum ces corrélations, puisque toutes les données sensibles de chaque classe d'équivalence vont se ressembler ! Ainsi, comme on levoit dans la **Figure 7**, la *t*-proximité permet surtout de répondre à la question suivante : *comment partitionner mes données de telle sorte que toutes les partitions se ressemblent en termes de distribution ?* Par exemple, si on imagine une base de données nationale sur des pathologies, comment regrouper les départements, classes d'âge et sexes, de telle sorte qu'on ait la même distributiondes pathologies dans chaque sous-groupe. On peut s'interroger du jeu de donnéesqui résulte de cette opération lorsqu'on souhaite précisément réaliser une analysequi fait ressortir les facteurs qui différencient les individus.

10.5. La confidentialité différentielle (Différential Privacy)

Nous concluons ce survol des techniques d'anonymisation par la confidentialité différentielle, une méthode très en vogue dans les milieux de la recherche en informatique depuis quelques années, car contrairement aux méthodes précédentes, elle est la seule à donner des garanties formelles, c'est-à-dire des preuves mathématiques, sur la possibilité de borner les informations qu'on peut apprendre sur les individus. Cette méthode introduit un échantillonnage des données vraies (avec une probabilité de α) , et une génération de données fictives avec une probabilité $\beta << \alpha$ (mais ces données doivent naturellement rester réalistes...). Les garanties formelles sont cruciales, et permettent de quantifier le risque de ré-identification des n-uplets, d'où l'engouement pour cette méthode. En effet, en observant le jeu de données anonymes, l'information qu'on peut obtenir sur le fait qu'un n-uplet soit vrai ou faux est doublement bornée : on n'estjamais sûr qu' un n-uplet soit vrai avec une probabilité supérieure à , ni qu'il soitfaux avec une probabilité inférieure à β[21].

11. Conclusion :

La vie privée est un droit fondamental qui revêt une importance croissante à l'ère numérique. Les avancées technologiques ont facilité la collecte, le partage et l'utilisation des données personnelles, ce qui soulève des préoccupations légitimes quant à la protection de la vie privée des individus.

La vie privée concerne la capacité des individus à contrôler l'accès, l'utilisation et la divulgation de leurs informations personnelles. Elle englobe des concepts tels que la confidentialité, le consentement éclairé, l'anonymat, la protection des données, la minimisation des données et la transparence.

Les réglementations sur la protection des données, telles que le RGPD et la CCPA, ont joué un rôle important en établissant des droits et des obligations pour les organisations traitant des données personnelles. Elles ont accru la sensibilisation à la vie privée et ont renforcé les mesures de protection des données.

Cependant, la protection de la vie privée reste un défi complexe. Les avancées technologiques telles que l'intelligence artificielle et l'Internet des objets soulèvent de nouvelles questions en matière de vie privée. Il est essentiel de développer des technologies, des réglementations et des pratiques qui tiennent compte de ces évolutions et qui garantissent un équilibre entre la protection de la vie privée et d'autres intérêts légitimes, tels que la sécurité.

Les utilisateurs jouent également un rôle crucial dans la protection de leur vie privée. Ils doivent être conscients de leurs droits en matière de vie privée, être informés des pratiques de collecte et d'utilisation des données, et être en mesure de prendre des décisions éclairées sur la divulgation de leurs informations personnelles.

En conclusion, la protection de la vie privée est un enjeu majeur de notre société numérique. Il est crucial de promouvoir une approche équilibrée entre l'innovation technologique et le respect des droits fondamentaux des individus. En adoptant des réglementations appropriées, en développant des technologies respectueuses de la vie privée et en promouvant la sensibilisation et l'éducation des utilisateurs, nous pouvons créer un environnement numérique plus respectueux de la vie privée.

Chapitre 3 : Anonymisation des données RDF en SHACL

1. Introduction :

L'anonymisation des données est un processus crucial pour protéger la confidentialité des informations sensibles tout en permettant leur utilisation à des fins d'analyse et de traitement. L'émergence du Web sémantique et des données RDF a ouvert de nouvelles possibilités pour l'anonymisation des données, en utilisant des langages et des outils spécifiques tels que SHACL (ShapesConstraintLanguage).

SHACL est un langage de contraintes destiné à spécifier des règles et des contraintes sur les données RDF. Il permet de définir des modèles de données, des schémas et des contraintes pour structurer et valider les données RDF. En utilisant SHACL, il est possible de définir des règles d'anonymisation pour les données RDF, facilitant ainsi la protection des informations sensibles contenues dans ces données.

Enfin, l'anonymisation des données en SHACL permet d'utiliser les données sensibles à des fins de recherche, d'analyse ou de partage, en générant des connaissances, en extrayant des informations et en effectuant des analyses statistiques, tout en préservant la vie privée des individus.

2. l'anonymisation des données RDF

L'anonymisation des données RDF (Resource Description Framework) est une technique de préservation des données privées qui consiste à supprimer ou à modifier les informations sensibles dans le graphe RDF de manière à ce qu'elles ne puissent pas être associées à une entité spécifique. Cette technique permet de protéger la confidentialité des données personnelles contenues dans les graphes RDF en évitant que ces informations ne puissent être liées à des individus spécifiques.

L'anonymisation peut impliquer la suppression de noms, de numéros d'identification personnels ou d'autres informations sensibles dans le graphe RDF. Elle peut également impliquer la substitution de ces informations par des identifiants anonymes, appelés pseudonymes. L'objectif de l'anonymisation est de rendre les données RDF difficilement identifiables et de réduire les risques de divulgation d'informations sensibles.

3. Validation des données avec SHACL

SHACL est un langage de validation de graphes RDF par rapport à un ensemble de conditions. Il se compose d'un ensemble de contraintes qui nous permet de décrire un schéma (graphe de formes) pour nos données d'instance RDF (graphe de données). Le graphe des formes avec notre graphe de données constituent l'entrée d'un moteur SHACL, qui effectuera la validation du monde fermé (données RDF). Le moteur retourne un rapport de validation, sous forme de graphe RDF, contenant soit conforme (i.e. les données sont valides)soit les violations détectées. Le rapport de violations (i.e. le résultat de validation), peut prendre la forme de graphe contenant des informations sur le ou les triplets qui ont causé l'erreur [22].

3.1. Les principaux composants de SHACL :

1.Shapes : Les formes sont les composants fondamentaux de SHACL. Elles définissent les règles et les contraintes qui doivent être respectées pour une entité donnée dans le graphe RDF.

2.Propriétés SHACL : Les propriétés SHACL sont des propriétés spéciales utilisées pour définir des règles et des contraintes sur les formes et les données RDF. Par exemple, la propriété sh:property est utilisée pour définir une propriété dans une forme.

3.Validationengine : Le moteur de validation SHACL est responsable de la validation des données RDF par rapport aux formes spécifiées. Il compare les données RDF à la forme et génère des rapports d'erreurs et d'avertissements en fonction des règles et des contraintes définies dans la forme.

4.Fonctions SHACL : Les fonctions SHACL sont des fonctions prédéfinies ou personnalisées qui peuvent être utilisées pour définir des règles et des contraintes plus complexes sur les données RDF.

5.Shapes graph : Le graphe des formes est un graphe RDF qui contient toutes les formes définies pour un ensemble de données RDF. Il peut être utilisé pour naviguer et explorer les formes définies dans les données.
6. Extensions SHACL : Les extensions SHACL permettent d'ajouter des fonctionnalités supplémentaires à SHACL en définissant des propriétés et des fonctions personnalisées.

En utilisant ces différents composants, SHACL permet de définir des contraintes et des règles pour les données RDF et de les valider pour assurer la conformité des données [23].

3.2. Comment fonctionne SHACL

SHACL (ShapesConstraintLanguage) fonctionne en définissant des règles et des contraintes sur les graphes RDF (Resource Description Framework) pour assurer la conformité des données à un modèle ou une forme spécifiée.

Les contraintes SHACL sont exprimées sous forme de « shapes » (formes) qui définissent les propriétés et les contraintes qui doivent être respectées pour une entité donnée dans le graphe RDF. Par exemple, une forme peut être définie pour spécifier que toutes les entités de type « Person » doivent avoir un nom, un prénom et une date de naissance valides.

Les formes sont définies à l'aide de propriétés SHACL telles que sh:property, sh:datatype et sh:minCount, qui définissent des règles sur les propriétés, les types de données et les cardinalités des valeurs des propriétés.

La validation SHACL est effectuée en comparant les données RDF à la forme spécifiée. Si les données RDF respectent toutes les règles et les contraintes définies dans la forme, alors la validation est réussie. Si les données ne respectent pas les règles définies, une erreur ou un avertissement peut être généré, en fonction de la sévérité de la contrainte.

SHACL permet également de spécifier des formes imbriquées, des formes conditionnelles et des formes d'inclusion pour définir des contraintes plus complexes sur les graphes RDF.

SHACL fonctionne en définissant des formes qui spécifient les contraintes et les règles que les données RDF doivent respecter, puis en validant les données par rapport à ces formes pour assurer la conformité [24].

4. l’anonymisation en SHACL :

SHACL offre un cadre puissant pour définir des contraintes sur les données RDF, ce qui permet d’exprimerl'anonymisation comme étant un ensemble de contraintes de manière efficace.Nous aborderons les principales techniques d'anonymisation des données RDF et expliquerons comment peut-on les mettre en œuvre en utilisant SHACL.

4.1. Anonymisation par généralisation :

La technique de généralisation est utilisée pour anonymiser des données RDF en agrégeant ou en regroupant les valeurs des propriétés sensibles afin de créer des catégories plus générales. Cela réduit la spécificité des données sensibles tout en préservant l'utilité globale des données. Par exemple, nous pouvons définir une contrainte pour généraliser l'âge en remplaçant les valeurs selon une fonction d'escaliers comme suit : $\begin{cases} f(age) = 10 \; si \; 1 \leq age < 10 \\ f(age) = 20 \; si \; 10 \leq age < 20 \\ \ldots . \\ f(age) = 100 \; si \; 90 \leq age < 100 \end{cases}$

Dans les données RDF anonymes, les valeurs spécifiques de l'âge ont été remplacées par des catégories d'âge plus générales, préservant ainsi la confidentialité des individus. Une autre technique de généralisation qui servira de taxonomies pour remplacer les valeurs des propriétés par d'autres plus générales selon une taxonomie de concepts ou de classes. Par exemple, les valeurs de la propriété maladie qui relie une maladie (e.g. cancer) à la personne affectée par cette maladie peuvent être remplacées par le concept le plus générale dans la taxonomie des maladies qui est **disease**.

La validation des données anonymes revient à créer un document SHACL exprimant que la propriété age doit avoir sa valeur dans l'ensemble $\{10,20,\ldots,90,100\}$. Cette contrainte peut être exprimée en utilisant la primitive `sh:in`quispécifie la condition selon laquelle chaque nœud de valeur est membre d'une liste SHACL fournie.Le document SHACL doit aussi exprimer que la propriété maladie doit avoir la valeur « **disease** ». cette contrainte peut être exprimée en utilisant la primitive `sh:hasValue`. Ci-dessous un code SHACL qui montre l'utilisation de cette primitive pour anonymiser les valeurs des propriétésage et maladie.

```
@prefix rdf:   <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix sh:    <http://www.w3.org/ns/shacl#> .
@prefix xsd:   <http://www.w3.org/2001/XMLSchema#> .
@prefix rdfs:  <http://www.w3.org/2000/01/rdf-schema#> .
@prefix ex:    <http://www.semanticweb.org/ahmed/ontologies/2023/4/untitled-
ontology-3#> .
@prefix owl:   <http://www.w3.org/2002/07/owl#> .
ex:PersonShape
ash:NodeShape ;
sh:targetClassex:Person ; # Applies to all persons
sh:property [              # _:b0
sh:pathex:age ;       # constrains the values of ex:age
sh:maxCount 1 ;
sh:in (10 20 30 40 50 60 70 80 90 100) ;
] ;
sh:property [              # _:b1
sh:pathex:disease ;       # constrains the values of ex:disease
sh:datatypexsd:string ;
sh:hasValue "disease";
sh:severitysh:Warning ;
] ;
```

```
sh:closed true ;
sh:ignoredProperties ( rdf:typeowl:topDataPropertyowl:topObjectProperty ) ;
```

4.2. Anonymisation par suppression :

La technique de l'anonymisation des données RDF par suppression consiste à supprimer complètement ou partiellement les informations sensibles présentes dans les données RDF afin de préserver la confidentialité des individus. Voici les détails sur cette technique. Le premier pas consiste à identifier les propriétés ou les valeurs spécifiques dans les données RDF qui contiennent des informations sensibles. Cela peut inclure des propriétés telles que les noms, les adresses, les identifiants personnels, les numéros de téléphone, etc. Il est important de déterminer quelles informations doivent être supprimées pour garantir la confidentialité des individus. En utilisant SHACL, nous pouvons définir des contraintes spécifiant que le document RDF des données anonymes doit être dépourvu de ces données sensibles. Nous utilisons la composante de contrainte `sh:maxCount`qui représenteune restriction sur le nombre de nœuds de valeur pour le nœud de focus donné.. Ci-dessous un code SHACL qui montre l'utilisation de cette primitive pour anonymiser les valeurs de la propriété maladie.

```
@prefix rdf:   <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix sh:    <http://www.w3.org/ns/shacl#> .
@prefix xsd:   <http://www.w3.org/2001/XMLSchema#> .
@prefix rdfs:  <http://www.w3.org/2000/01/rdf-schema#> .
@prefix ex:    <http://www.example.org/#> .
@prefix owl:   <http://www.w3.org/2002/07/owl#> .
@prefix ex:   <http://www.semanticweb.org/ahmed/ontologies/2023/4/untitled-
ontology-3#> .
ex:PersonShape
ash:NodeShape ;
sh:targetClassex:Person ; # Applies to all persons
sh:property [              # _:b0
sh:pathex:disease ;       # constrains the values of ex:disease
sh:maxCount 0 ;
    ] ;

sh:closed true ;
sh:ignoredProperties ( rdf:typeowl:topDataPropertyowl:topObjectProperty ) ;
```

Dans les données RDF anonymes, les triplets contenant les informations sensibles ont été supprimés. Les noms et les adresses des individus ne sont plus présents, préservant ainsi leur confidentialité.Il est important de noter que lors de l'anonymisation par suppression, certaines informations peuvent être perdues, ce qui peut affecter l'utilité des données pour certaines analyses. Par conséquent, il est essentiel d'évaluer attentivement les besoins de confidentialité et d'utilité des données lors de l'application de cette technique.

5. Conclusion :

L’anonymisation des données RDF en utilisant SHACL offre une approche puissante et flexible pour préserver la confidentialité des informations sensibles tout en conservant l'intégrité et l'utilité des données. SHACL permet de vérifier la conformité des données RDF aux contraintes définies, assurant ainsi la qualité et la cohérence des données. Nous venons de découvrir qu’Il offre un moyen puissant de valider les données RDF dans le contexte de la vie privée. Le processus de validation avec SHACL (ShapesConstraintLanguage) implique deux étapes clés : définition des contraintes SHACL et exécution de la validation. SHACL est un langage générale de contraintes qui nous a permis d’exprimer des contraintes imposées par des techniques de l’anonymisation des données. Les résultats de la validation permettent d'identifier les problèmes dans les données RDF. En fonction des violations détectées, des actions de sauvetage peuvent être entreprises pour cacher les valeurs divulguées.La validation de la vie privée des données avec SHACL garantit la divulgation des données RDF tout en assurant la confidentialité des données sensibles.

Chapitre 4 : Implémentation

Chapitre 4

Implémentation

1. Introduction

L'anonymisation d'un graphe RDF est un processus nécessaire pour toute tentative de publication des données dans le web sémantique afin de préserver la vie privée des individus du maillage. Plusieurs travaux ont été réalisés pour démontrer le risque de violation de la vie privée en publiant un graphe RDF anonyme. Généralement, cette simple procédure n'est pas suffisante et il est possible de réidentifier certains individus dans le graphe si l'adversaire possède certaines connaissances préalables sur sa cible. Tout au long de ce projet, notre travail consistait à étudier et de comprendre les approches existantes d'anonymisation des données liées pour but d'évaluer ces techniques avec une approche d'intrusion informatique qui visent les violations des données personnelles, afin de proposer des perspectives qui concerne l'anonymisation des graphes RDF.

2.Préliminaire

L'objectif principale de cette partie est de présenter l'architecture générale de l'application ce chapitre démontre l'impact de notre approche sur les mesures d'anonymisation des données personnelleset indique notre choix du langage de programmation. Nous commençons par décrire le principe de cette approche et ses détails et nous terminons par un test et quelques résultats. Ensuite, nous décrivons courtement notre application et nous présentons les résultats obtenus. Enfin nous concluons par une conclusion.

3. Outils et langage

3.1. Le langage Python

Python est le langage de programmation le plus utilisé dans le domaine du Ma- chine Learning, du Big Data et de la Data Science. Découvrez tout ce que vous savoir à son sujet : définition, avantages, cas d'usage. . . Créé en 1991, le langage de pro- grammation Python apparu à l'époque comme une façon d'automatiser les éléments les plus ennuyeux de l'écriture de scripts ou de réaliser rapidement des prototypes d'applications. Depuis quelques années, toutefois, ce langage de programmation s'est hissé parmi les plus utilisés dans le domaine du développement de logiciels, de gestion d'infrastructure et d'analyse de données. Il s'agit d'un élément moteur de l'explosion du Big Data. Python est un langage de programmation open source créé par le programmeur Guido van Rossum en 1991. Il tire son nom de l'émission Monty Pythons Flying Circus.

Il s'agit d'un langage interprété, qui ne nécessite donc pas d'être compilé pour fonctionner. Un programme » interpréteur » permet d'exécuter le code Python sur n'importe quel ordinateur. Ceci permet de voir rapidement les résultats d'un changement dans le code. En revanche, ceci rend ce langage plus lent qu'un langage compilé comme le C. En tant que langage de programmation de haut niveau, Pythonpermet aux programmeurs de se focaliser sur ce qu'ils font plutôt que sur la façon dont ils le font. Ainsi, écrire des programmes prend moins de temps que dans un autre langage. Il s'agit d'un langage idéal pour les débutants. Langage Python : quels sont les principaux avantages ? Le langage Python doit sa popularité à plusieurs avantages qui profitent aussi bien aux débutants qu'aux experts. Tout d'abord, il est facile à apprendre et à utiliser. Ses caractéristiques sont peu nombreuses, ce qui permet de créer des programmes rapidement et avec peu d'efforts.De plus, sa syntaxe est conçue pour être lisible et directe. Un autre avantage du Python est sa popularité. Ce langage fonctionne sur

tous les principaux systèmes d'exploitation et plateformes informatiques. De plus, même s'il ne s'agit clairementpas du langage le plus rapide, il compense sa lenteur par sa versatilité. Enfin, mêmes'il est

principalement utilisé pour le Scripting et l'automatisation, ce langage est aussi utilisé pour créer des logiciels de qualité professionnelle. Qu'il s'agisse d'applications ou de services Web, le Python est utilisé par un grand nombre de développeurs pourcréer des logiciels. [25]

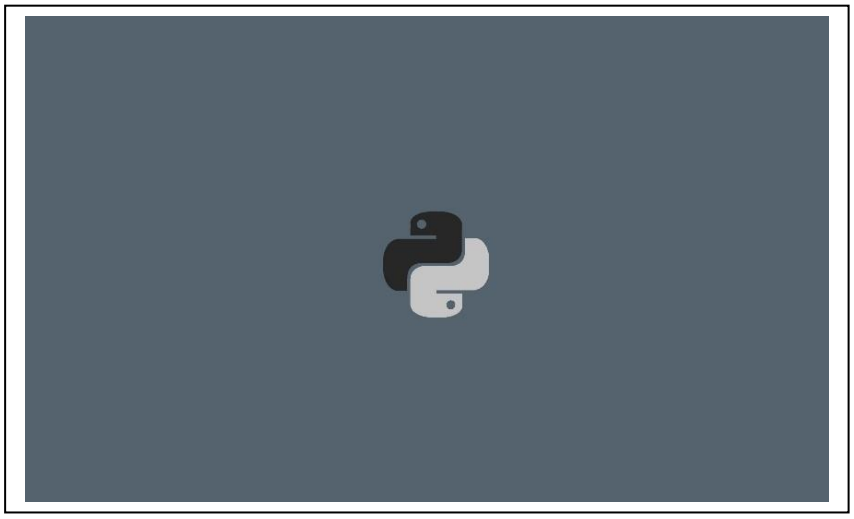


FIGURE 8 – Python Programming Language Logo

3.2. Le navigateur Anaconda

Le navigateur Anaconda est une interface utilisateur graphique (GUI) de bureau inclus dans la distribution Anaconda qui nous permet de lancer des applications et de gérer facilement les packages, les environnements et les canaux conda sans utiliser des commandes de ligne de commande [26] .

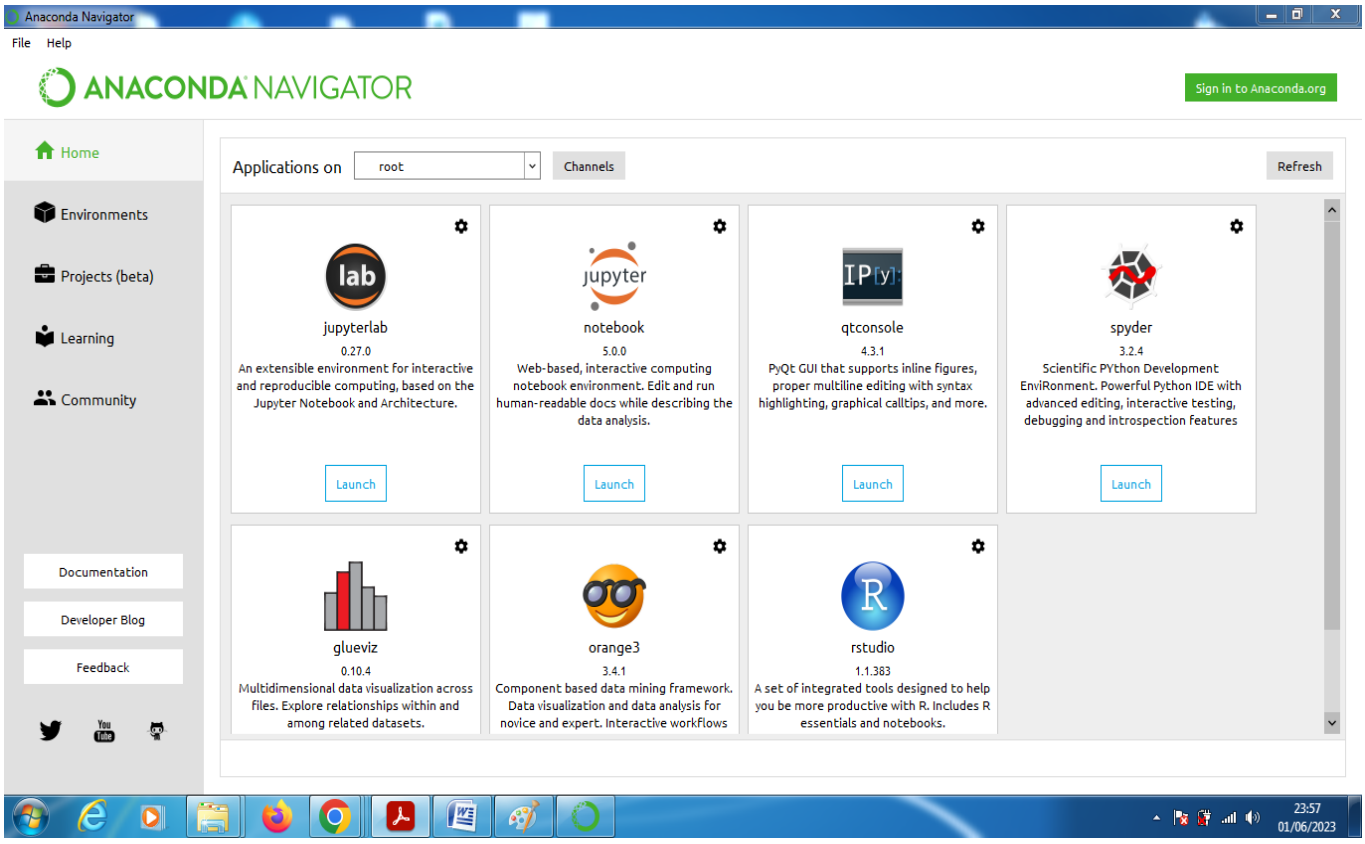


Figure 9: Le navigateur Anaconda

3.3. Jupyter NoteBook :

Jupyter Notebook (figure 4.2) et son interface flexible étendent le notebook au-delà du code à la visualisation, au multimédia, à la collaboration, etc. En plus d'exécuter le code, il stocke le code et la sortie, ainsi que les notes de démarque, dans un document modifiable appelé bloc-notes. Lorsqu'il est enregistré, celui-ci est envoyé du navigateur au serveur du bloc-notes, qui l'enregistre sur le disque en tant que fichier JSON avec une extension .ipynb [27] .

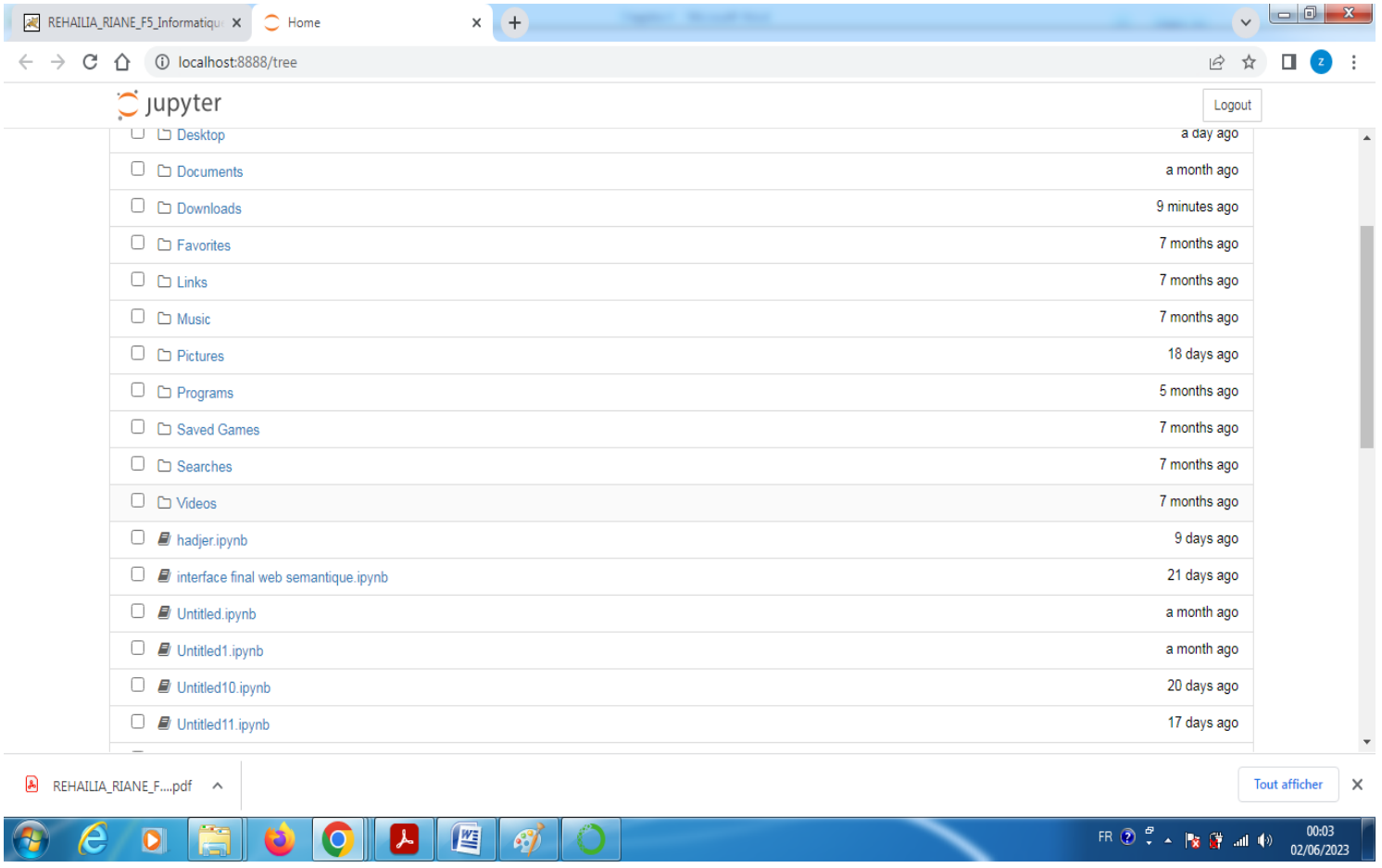


Figure 10 : Jupyter NoteBook

3.4. Implémentation et démonstration:

3.4.1 Description de l'application :

L'application permet de :

Phase 1 :l'identification basique réalisée en utilisant la bibliothèque Tkinter en Python. L'interface permet aux utilisateurs de saisir leur nom d'utilisateur et leur mot de passe, puis de cliquer sur le bouton "Se connecter" pour soumettre leurs informations. Si les informations de connexion sont correctes, la fenêtre de connexion se ferme et la fenêtre principale de l'application s'ouvre. Sinon, un message d'erreur est affiché.

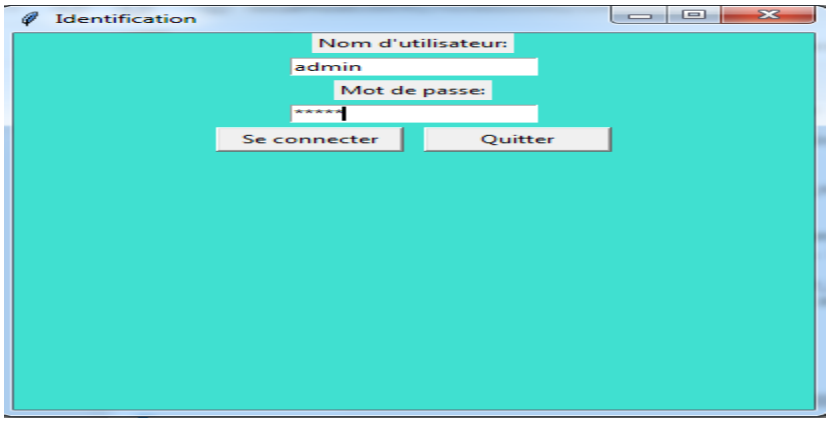


Figure 11 : fenêtre identification

Phase 2 : fenêtre principale :

La fenêtre principale affiche un menu avec différentes options pour créer , charger, manipuler et analyser des fichiers RDF. Les options du menu incluent le chargement de fichiers RDF, l'exécution de requêtes SPARQL, l'anonymisation des données, la validation des données par rapport à des règles SHACL, et la possibilité de quitter l'application. Lorsque vous sélectionnez une option du menu, une action correspondante est exécutée.

Figure 12 : fenêtre principal

Créer un fichiers RDF :

L'option "Créer RDF" du menu permet à l'utilisateur de créer un fichier RDF à partie des données textuel une fois créer, le fichier RDF est sauvegarder sur l'ordinateur.

Chargement de fichiers RDF :

L'option "Charger" du menu permet à l'utilisateur de sélectionner un fichier RDF à charger dans l'application.
Une fois chargé, le fichier RDF est parsé et stocké dans un graphe RDF.
En cas d'erreur lors du chargement ou du parsing du fichier, un message d'erreur est affiché.

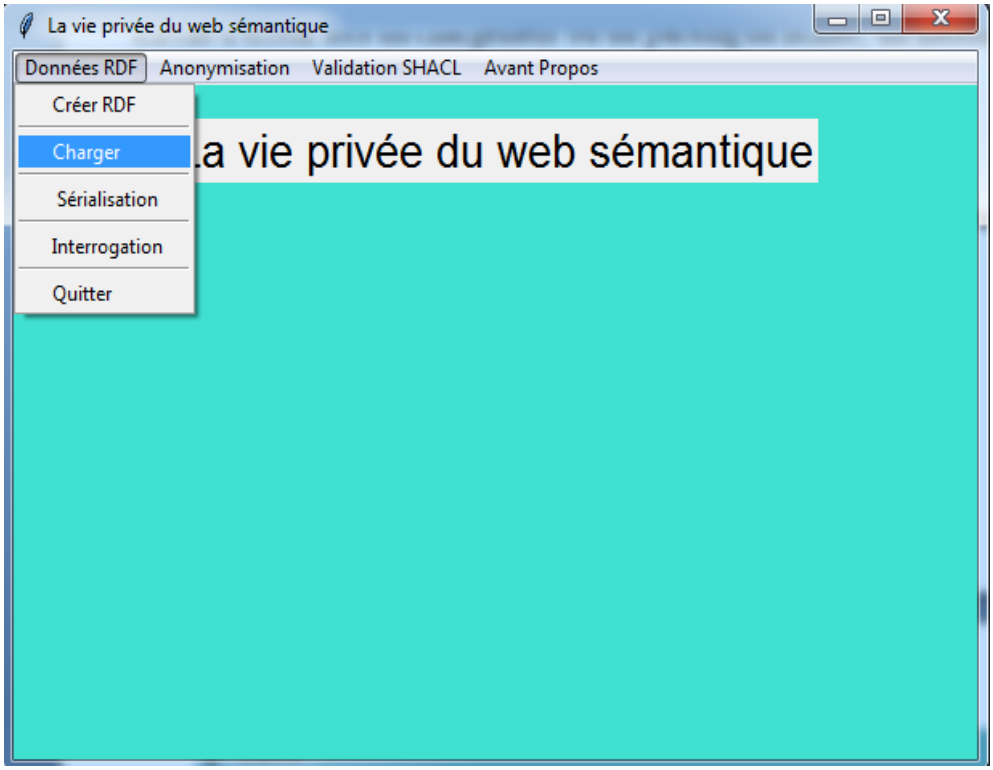


Figure 13 : fenêtre Chargement de fichiers RDF

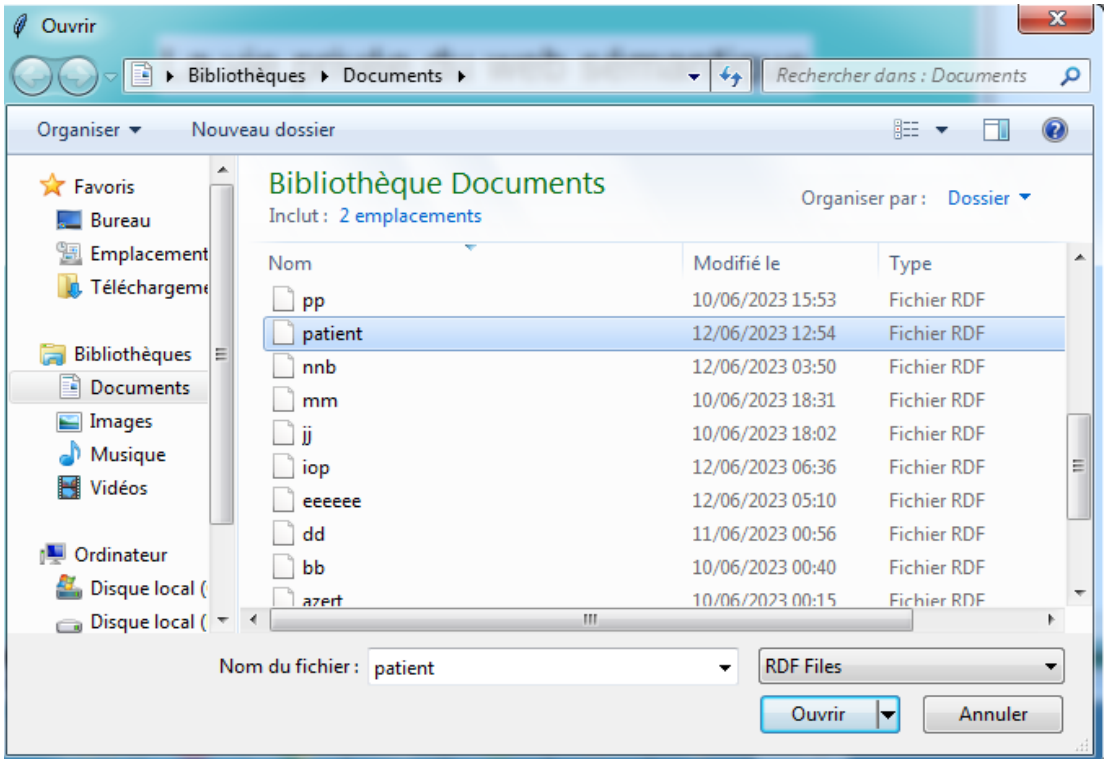


Figure 14 : sélectionnez le fichier RDF

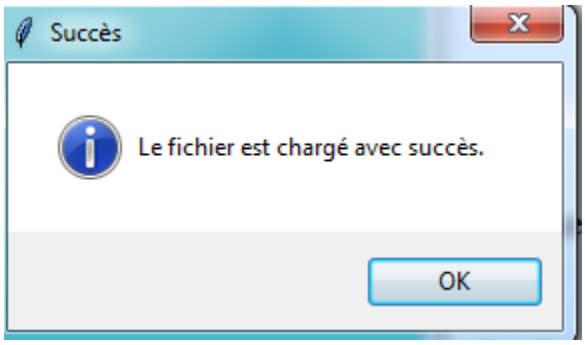


Figure 15: téléchargement termine du fichier RDF

Sérialisation :

L'option " Sérialisation" du menu permet à l'utilisateur de sélectionner un format de sortie (turtle, xml, nt, json-ld) et de spécifier un chemin d'enregistrement.

Le graphe RDF chargé est ensuite sérialisé dans le format choisi et enregistré dans le fichier spécifié.

Le contenu du fichier sérialisé est affiché dans une fenêtre de résultat.

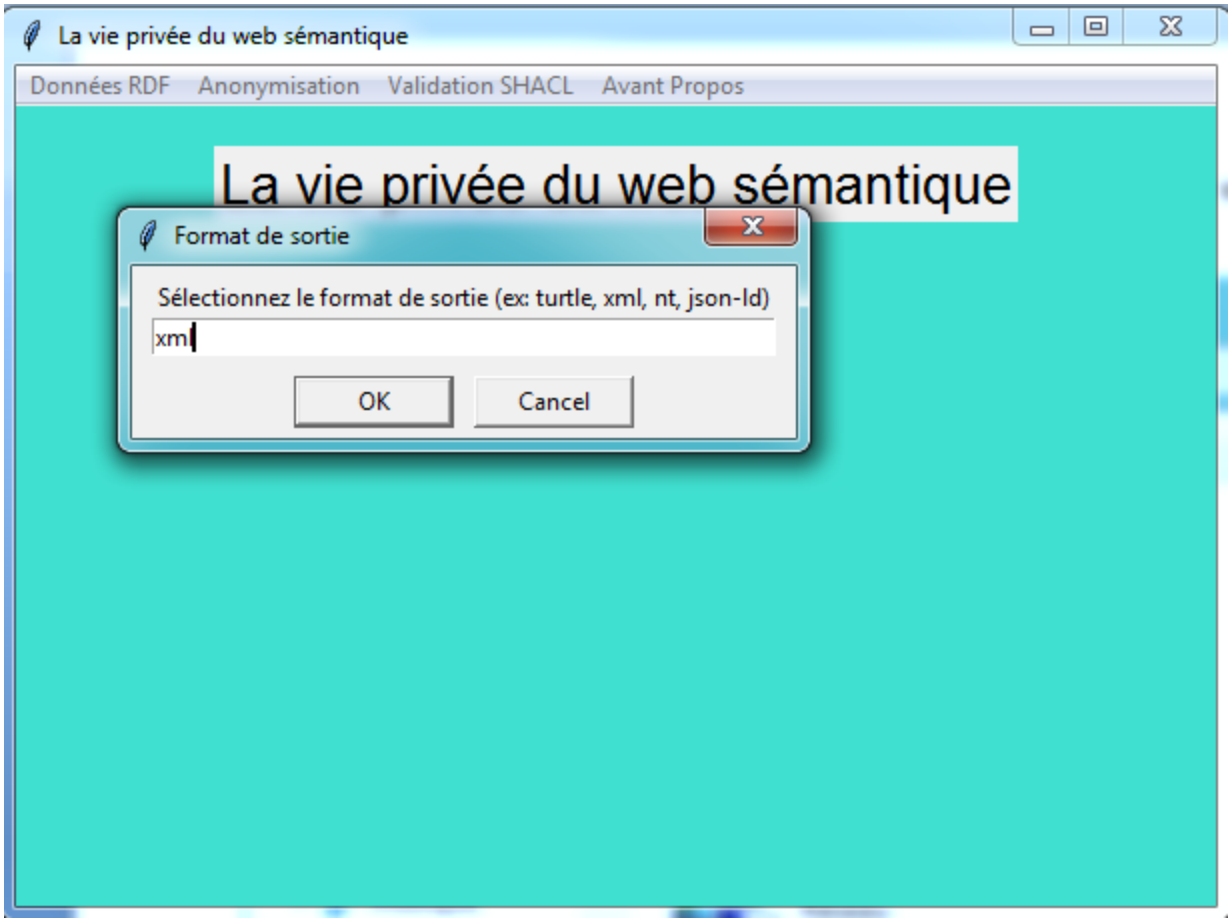


Figure 16: Choix du format de sérialisation

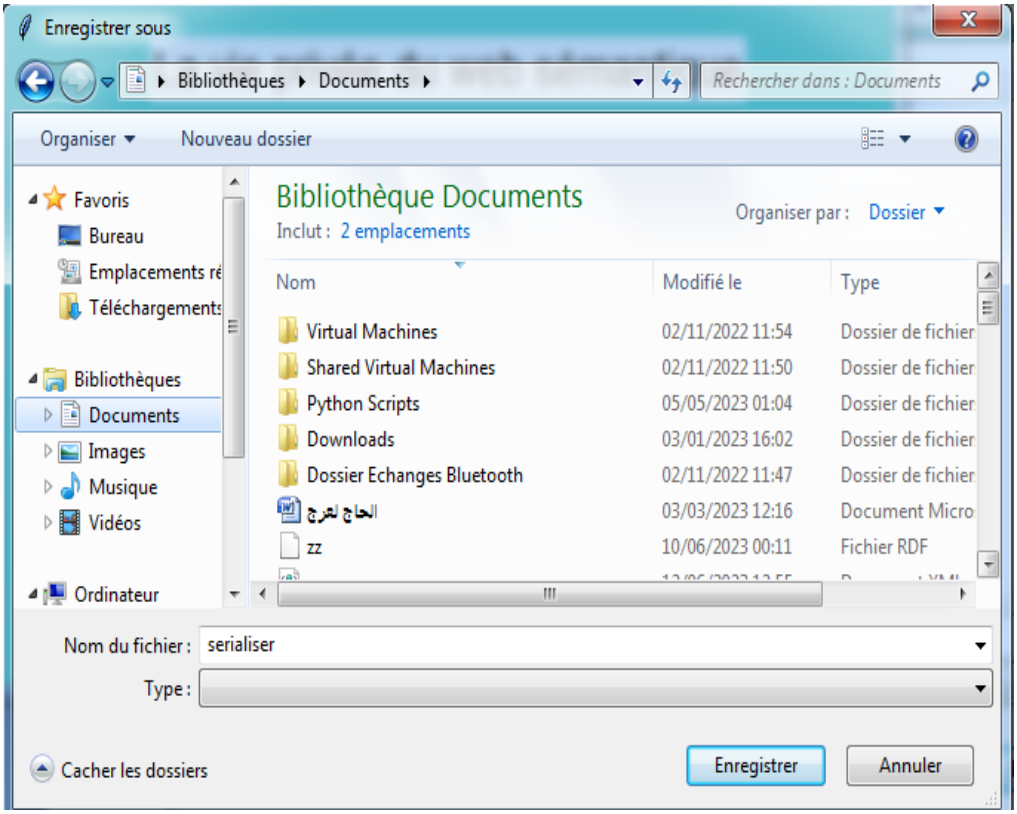


Figure 17 : enregistrement du fichier sérialisé

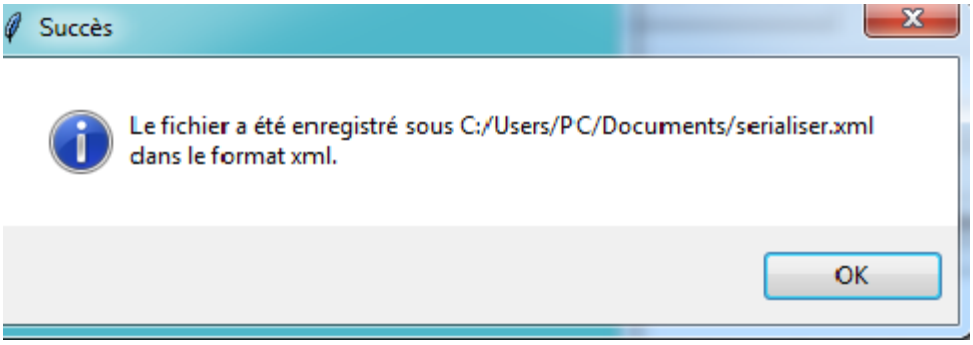


Figure 18: enregistrement termine du fichier sérialisé

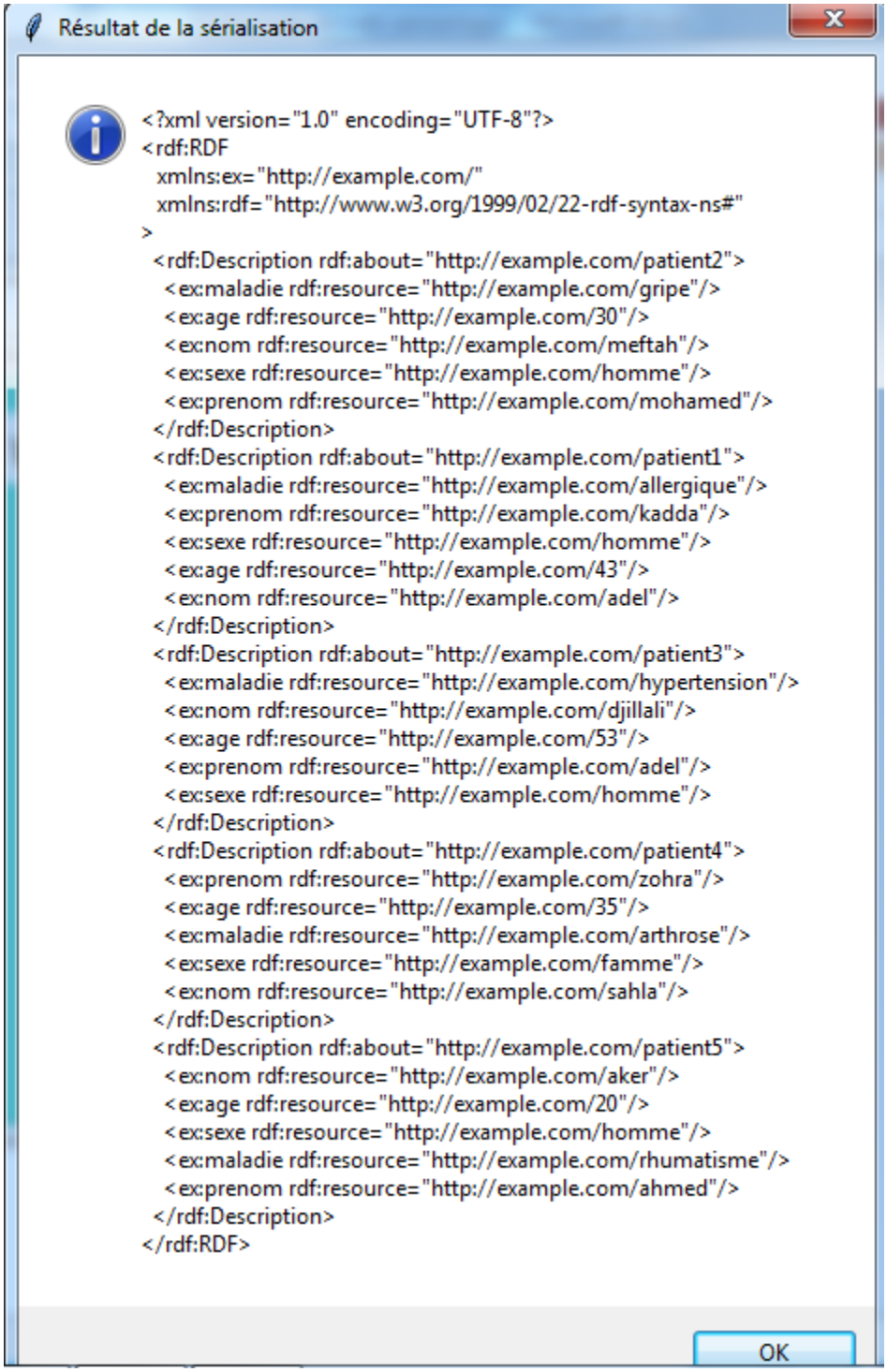


Figure 19: résultat de la sérialisation

Exécution de requêtes SPARQL :

L'option "Interrogation" du menu permet à l'utilisateur de sélectionner un fichier contenant une requête SPARQL à exécuter. La requête SPARQL est exécutée sur le graphe RDF chargé et les résultats sont affichés dans une fenêtre de résultat.

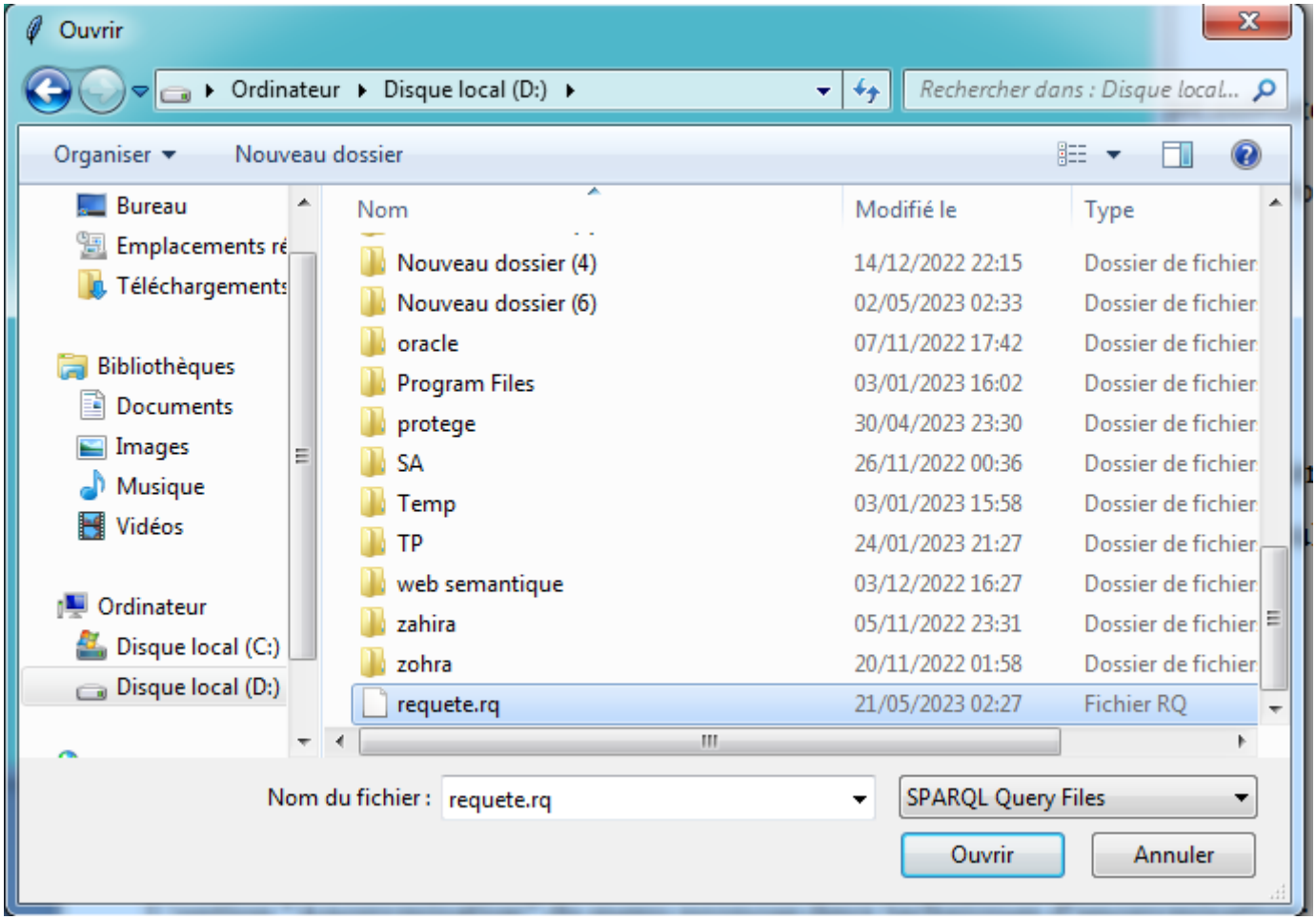


Figure 20 : sélectionnez le fichier une requête SPARQL

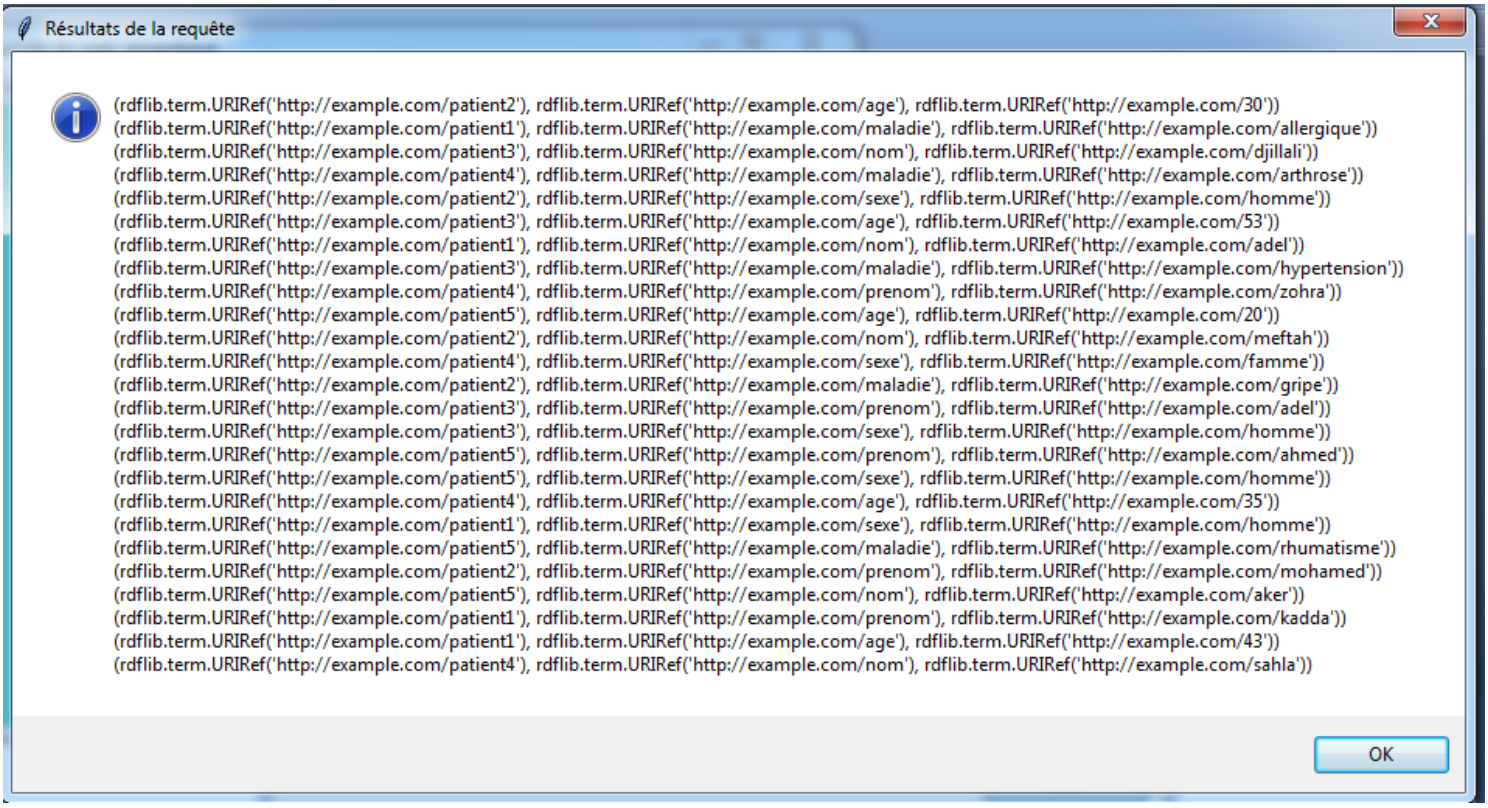


Figure 21 : résultats de la requête SPARQL

Anonymisation :

L'option "Anonymisation" du menu propose deux techniques d'anonymisation : généralisation et suppression.

Lorsque l'utilisateur sélectionne l'une des techniques, les données sensibles du graphe RDF sont anonymisées en remplaçant les valeurs par des valeurs génériques ou en supprimant les triplets correspondants.

Le graphe RDF anonymisé est affiché dans une fenêtre de résultat.

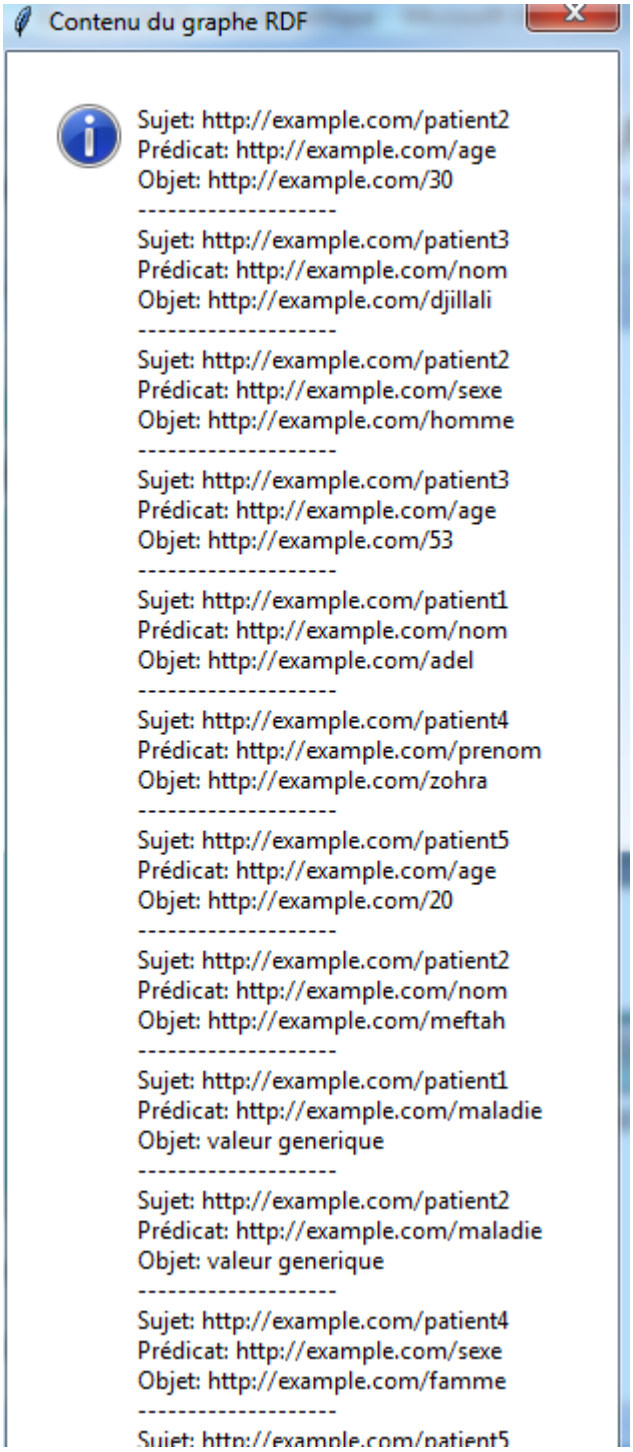


Figure 22 : résultat de l'anonymisation par généralisation

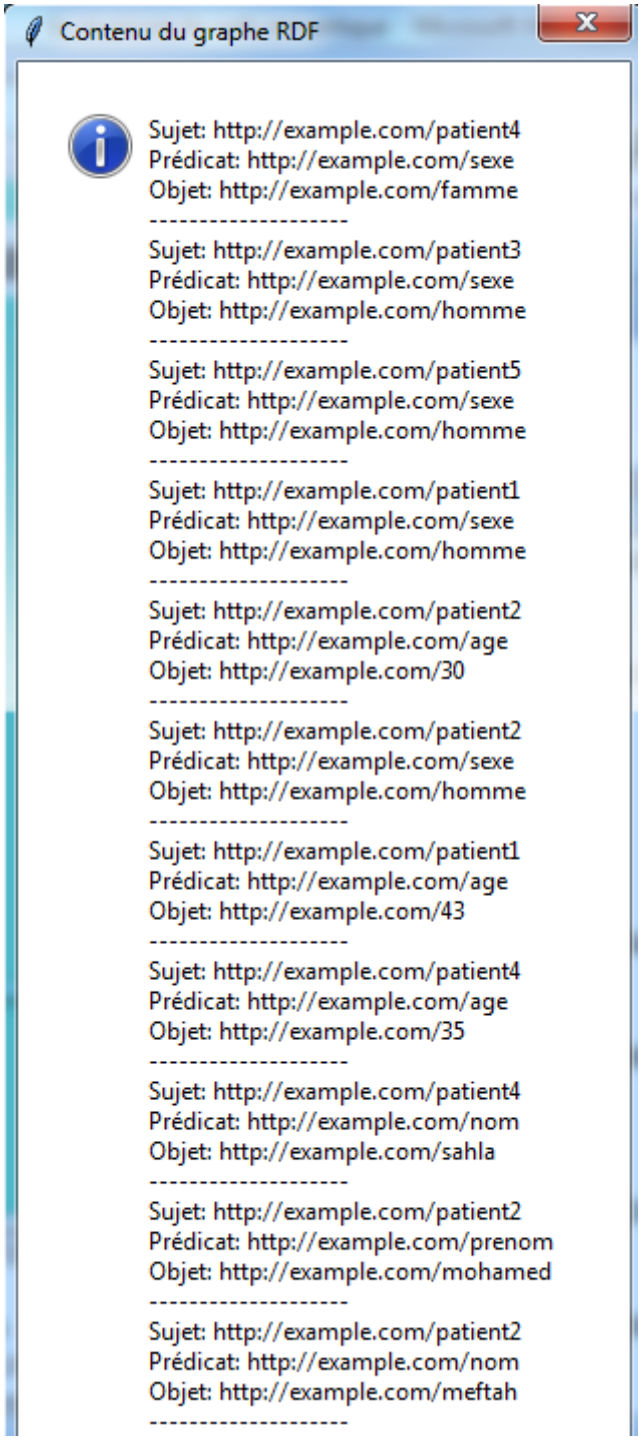


Figure 23 : résultat de l'anonymisation par suppression

Validation des données :

L'option "Validation" du menu propose différentes options de validation, notamment la validation par techniques de généralisation et de suppression, ainsi que la validation SHACL.

Lorsque l'utilisateur sélectionne l'une des options de validation, le graphe RDF chargé est validé par rapport à un document SHACL contenant des règles de validation.

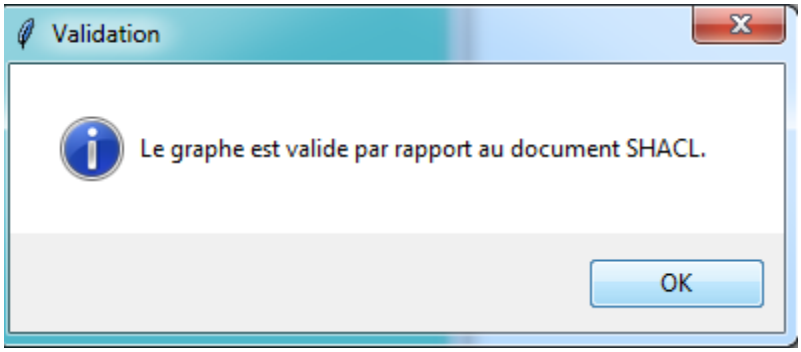


Figure 24: validation par rapport au document SHACL

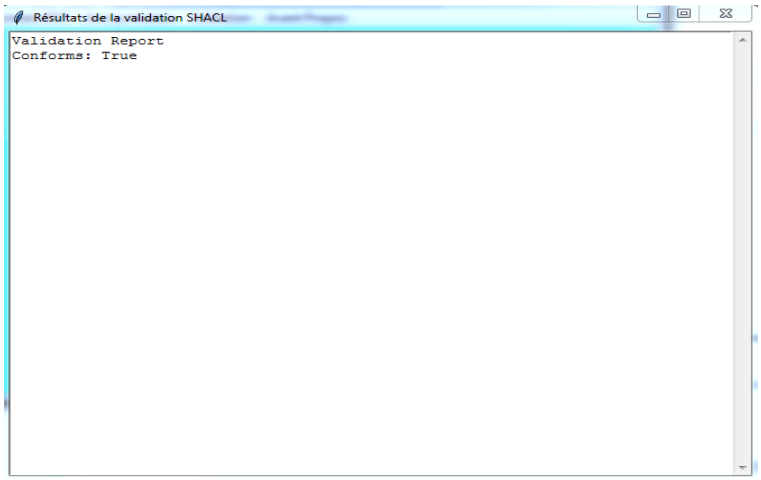


Figure 25: Résultats de la validation SHACL

4.Conclusion :

l'utilisation de SHACL (Shapes Constraint Language) pour l'anonymisation des données RDF offre des fonctionnalités puissantes pour protéger la confidentialité des informations sensibles. SHACL permet de définir des contraintes et des validations qui peuvent être appliquées aux données RDF afin d'assurer la conformité aux politiques de confidentialité et de protection des données

En combinant les fonctionnalités de SHACL avec d'autres techniques d'anonymisation des données, telles que la suppression, la généralisation ou la perturbation, il est possible de créer des processus d'anonymisation robustes pour préserver la confidentialité des données RDF. Cependant, il est important de mener une évaluation approfondie des risques et de suivre les meilleures pratiques en matière de protection des données lors de l'anonymisation des données RDF en SHACL.

Conclusion générale

La vie privée dans le web sémantique est un sujet complexe qui nécessite une attention particulière. Bien que le web sémantique offre de nombreux avantages en termes de représentation et de gestion des données, il présente également des défis en matière de protection de la vie privée. Les technologies et les protocoles utilisés dans le web sémantique permettent une interconnexion et un partage de données à grande échelle, ce qui peut potentiellement compromettre la confidentialité des utilisateurs.

La nature même du web sémantique, axée sur l'interconnexion des données, peut rendre difficile le contrôle précis de la diffusion et de l'accès aux informations personnelles. Les données sémantiques peuvent être agrégées, combinées et inférées pour obtenir des informations plus détaillées sur les utilisateurs, ce qui soulève des préoccupations en matière de vie privée.

Pour préserver la vie privée dans le web sémantique, il est essentiel de mettre en place des mesures de protection adéquates. Cela peut inclure l'utilisation de techniques d'anonymisation et de pseudonymisation pour masquer les informations sensibles, ainsi que l'application de politiques de contrôle d'accès pour limiter l'accès aux données personnelles uniquement aux entités autorisées.

De plus, les utilisateurs doivent être conscients des informations qu'ils partagent et de la manière dont elles sont utilisées dans le contexte du web sémantique. Il est important de lire attentivement les politiques de confidentialité et de comprendre comment les données personnelles sont collectées, stockées et partagées par les applications et les services utilisant le web sémantique.

Enfin, une réglementation et une gouvernance appropriées sont nécessaires pour encadrer l'utilisation des données dans le web sémantique et protéger la vie privée des utilisateurs. Les autorités réglementaires doivent travailler en étroite collaboration avec les acteurs du web sémantique pour développer des normes et des pratiques de protection de la vie privée qui répondent aux besoins et aux préoccupations des utilisateurs.

En adoptant une approche équilibrée entre l'exploitation des avantages offerts par le web sémantique et la préservation de la vie privée, il est possible de créer un environnement numérique sécurisé et respectueux de la confidentialité des utilisateurs. Notons que cette problématique est vaste et très complexe, et loin d'être traitée d'une façon complète dans un mémoire.

Bibliographie

[1] Xavier Lacot, "Introduction à OWL, un langage XML d'ontologies Web ", juin 2005.

[2] http://www.techno-science.net/?onglet=glossaire&definition=1450

[3] T. Berners-Lee, J. Hendler, and O. Lassila, (2001). "The Semantic Web". Scientific American, May 2001.

[4] https://fr.wikipedia.org/wiki/Web_s%C3%A9mantique.

[5] Bach Thành Lê, "Construction d'un Web sémantique multi-points de vue", thèse de Doctorat soutenue le 23 octobre 2006 à L'École des Mines de Paris à Sophia Antipolis.

[6] Ould Brahim Mohamed, découvert et sélection de service web sémantique mémoire de master soutenue publiquement le 05/04/2014, à Université Larbi Ben M'hidi – Oum El Bouaghi.

[7] Magkanaraki Aimilia Et Al, Benchmarking RDF Schemas For The Semantic Web. P.132-146

[8] Broekstra Jean, Kampman Arjohn And Van Harmelen Frank, Sesame: A Generic Architecture For Storying And Querying RDF And RDF Schema.

[9] A. Bouramoul, Recherche D'information Contextuelle Et Sémantique Sur Le Web, Thèse Pour Obtenir Le Grade De Docteur En Sciences, Université Mentouri De Constantine, 2011.

[10]Abdeloued sarah, Traitement des données sémantique basé sur l'utilisation de l'outil metis mémoire de master ,Université Abdelhamid Ibn Badis -Mostaganem .

[11] https://fr.wikipedia.org/wiki/SHACL

[12] https://www.w3.org

[13] Aimeur, E. (2016). Les Enjeux de la Vie privée sur Internet.

[14] Bastien,L. (2019).Python :tout savoir sur le principal langage big data et machine learning. https://www.lebigdata.fr/python-langage-definition.

[15] Belabed, A. (2018). La protection de la vie privée sur Internet. PhDthesis, Université Abou-Bekr Belkaid, Tlemcen.

[16]Shwan.K (2019). La préservation de la vie privée dans le web de données.memoire de master Universite Dr. Tahar Moulay Saida.

[17] Yves Deswarte. Intelligence ambiante et protection dela vie privée,cours, 2004.

[18] Sébastien Gambs. Introduction à la protection de la vie privée, cours, 2012.

[19] Yves Deswarte. Intelligence ambiante et protection dela vie privée,cours, 2004.

[20] Yves Deswarte , Sébastien Gambs. Protection dela vie privée : Principes et technologies, CNRS ; France Université de Toulouse,2016.

[21] Bendida .A ,Guendouze.B.La vie privee dans les graphes de connaissances.memoire de master Universite Dr. Tahar Moulay Saida.

[22] https://www.w3.org/TR/shacl/

[23] https://www.w3.org/TR/shacl/#constraints

[24] https://www.w3.org/TR/2016/WD-shacl-20160128/

[25] Bastien,L. (2019). Python : tout savoir sur le principal langage big data et machine learning. https://www.lebigdata.fr/python-langage-definition.

[26] https://docs.anaconda.com/free/navigator/index.html

[27] https://jupyter.org/