
Rapport Final du projet

Contexte et Objectifs du projet:

Ce projet s'inscrit dans le cadre de la recherche sur le Self-Sovereign Identity (SSI) et le Digital Identity Wallet. L'objectif principal est de concevoir une taxonomie permettant de structurer et d'organiser les différentes facettes de ces concepts à partir des publications scientifiques disponibles. Ce projet vise à comparer et à combler les écarts entre les approches académiques et industrielles en matière de gestion des identités numériques

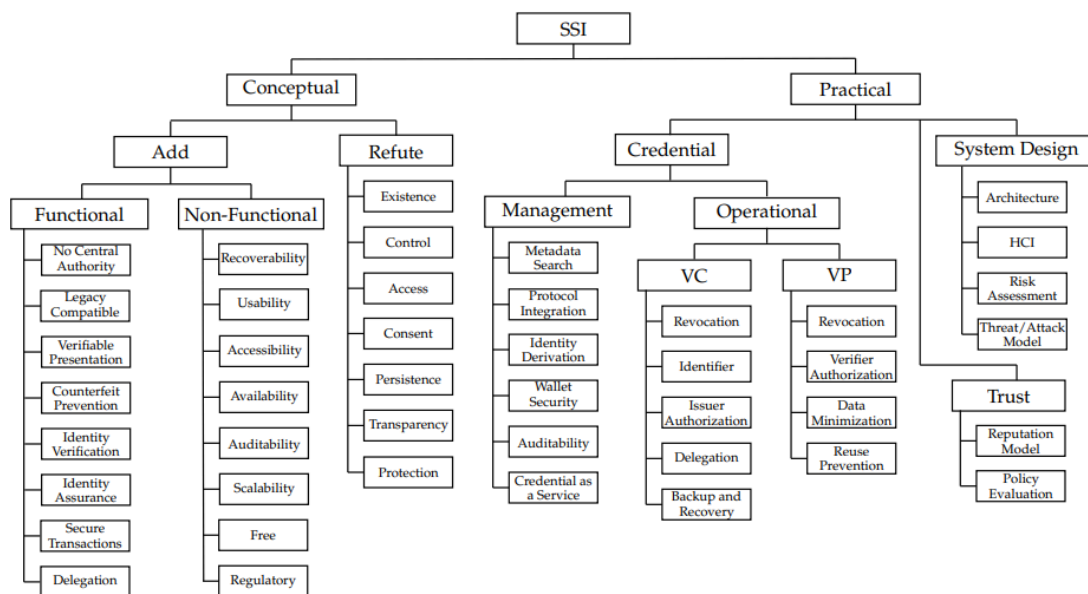


Figure 4: Taxonomy of SSI.

Méthodologie du projet:

Pour mener à bien ce projet, nous avons établi une nouvelle méthodologie finale (voir **Methodologie_Thales.drawio**). Pour obtenir plus d'informations, nous disposons d'un autre document détaillant plus précisément les différentes étapes de la méthodologie (voir **All_Methodologie_Thales.drawio**).

Le projet est divisé en quatre phases distinctes :

- Phase I : **Literature Review and Project Definition**
- Phase II : **Data Acquisition and Filtering**
- Phase III : **Taxonomy and Systematic Mapping**
- Phase IV : **Final Deliverables**

Papiers de recherches:

Pour comprendre les différents concepts et termes scientifiques autour du **SSI/Digital Identity Wallet**, nous disposons d'un ensemble de papiers de recherche (voir dossier **ResearchPaper**). Cela permettra également de mieux appréhender la structure d'un papier de recherche, ce qui nous sera très utile pour la suite du projet.

Analyse et Extraction des papiers de recherche à partir d'une base de données académique :

Pour élaborer cette taxonomie, nous avons besoin de données, plus précisément de papiers de recherche. Pour récupérer ces papiers, nous devons les extraire à partir d'une base de données académique. Nous avons donc choisi la base de données académique **Semantic Scholar**. Ce choix repose sur plusieurs critères, mais la principale raison est qu'elle propose une API gratuite qui permet de manipuler facilement les données.

Nous te fournissons le lien vers la documentation de l'API :

https://api.semanticscholar.org/api-docs/#tag/Paper-Data/operation/get_graph_get_paper_autocomplete

Nous allons maintenant nous concentrer sur le notebook

ExtractResearchPaperSemanticScholar.ipynb. Dans ce notebook, nous avons procédé par différentes petites étapes afin d'atteindre une étape finale où toutes ces petites étapes mises bout à bout (pagination des résultats, application des critères, etc.) forment un code complet (voir section **Final Step: Retrieving and Saving Papers in Excel, JSON, and CSV Formats via the Semantic Scholar API**).

Lorsque nous parlons de récupération de papiers de recherche, nous faisons plus précisément référence aux champs que nous extrayons (authors, title, abstract, id, etc.).

Petite précision concernant la **Query** : lorsque nous interrogeons la base de données, nous utilisons deux termes, **"Self-Sovereign Identity"** et **"Digital Identity Wallet"**. Nous exécutons donc le code une fois pour chaque terme. Une fois l'extraction des papiers de recherche réalisée, nous ajoutons une colonne

references et y indiquons les paperId des articles référencés dans chaque papier (voir section **Updating Paper References via the Semantic Scholar API**). Des explications plus détaillées sont directement disponibles dans le notebook.

IMPORTANT : L'ajout des références n'a pas été effectué sur l'ensemble des papiers de recherche, que ce soit pour les papiers sur le **Self-Sovereign Identity** ou ceux sur le **Digital Identity Wallet**. Il suffit d'exécuter les sections de code suivantes : **"Updating Paper References via the Semantic Scholar API"** et **"Updating Paper References with Pagination and Error Handling via the Semantic Scholar API"**, en chargeant les différents fichiers correspondants.

Algorithme et Clusterisation :

Nous allons maintenant nous concentrer sur le notebook **Cluster_Algorithm.ipynb**. Nous avons utilisé des données totalement différentes traitant d'un autre sujet pour plusieurs raisons :

- Nous connaissons déjà les résultats de ces données.
- Le but de ces différents algorithmes et méthodes est de tester leur efficacité et de vérifier leur bon fonctionnement sur des données dont nous maîtrisons déjà les résultats.

À court terme, l'objectif est d'appliquer ces algorithmes et méthodes aux différents papiers de recherche extraits précédemment (avec les références) sur les sujets du **Self-Sovereign Identity** et du **Digital Identity Wallet**.

Lien du site (expliquant les algorithmes et méthodes que j'ai d'ailleurs adaptés à mon cas d'utilisation) : <http://brandonrose.org/clustering>

Dans ce notebook, comme dans le précédent, nous avons procédé par différentes petites étapes pour atteindre une étape finale où toutes ces étapes mises bout à bout (tokenisation, stemming, suppression des stopwords, etc.) forment un code complet (voir section **"Final Step: Cosine Similarity Calculation for Combined Titles and Abstracts Using TF-IDF"**). L'objectif de ce notebook est d'évaluer la similarité entre les documents. Pour cela, nous avons décidé de vérifier cette similarité à la fois sémantiquement et par le nombre de références partagées en commun par paire de documents (voir section **"Calculating Document Similarity Based on References"**).

Pour évaluer la similarité des documents en fonction des références partagées, nous avons appliqué une première méthode (voir section **"Calculating and Visualizing Reference Similarity Between Documents"**) qui consiste à prendre le nombre de références communes entre les documents et à le diviser par le nombre total de

références des deux documents. Ensuite, nous multiplions le résultat par 100 pour obtenir un pourcentage de similarité. Le problème de cette méthode réside dans les cas où les documents ont un nombre de références très différent (par exemple, si le document 1 possède une dizaine de références et que le document 2 en possède une cinquantaine, cela peut fausser complètement le pourcentage de similarité).

Nous nous sommes donc orientés vers une deuxième méthode, en nous appuyant sur le papier de recherche **Van2007 (dans le dossier ResearchPaper)**, qui propose une formule logarithmique pour calculer la similarité des coréférences entre documents. Cependant, nous avons rencontré un problème avec cette méthode, car elle produit des résultats incohérents. Cela pourrait être dû à des erreurs dans le script ou à une mauvaise application de la formule (voir section "**Calculating Co-Reference Similarity Between Documents**" — cette section n'est donc pas opérationnelle pour le moment).

Enfin, il est important de mentionner que nous calculons la similarité des documents à la fois sémantiquement et par les références afin de voir s'il existe une corrélation entre ces deux types de similarité. Dans le cadre du schéma présenté dans la section "**Correlation Between Semantic and Reference-Based Similarity**", si les résultats tendent vers une diagonale, cela indiquerait une corrélation entre ces deux types de similarité.

Extraction de domaines d'application et de technologies :

Nous avons également un autre notebook (**Extract_Domain_Technologies.ipynb**) qui permet d'extraire deux informations clés à partir des abstracts des papiers de recherche : le domaine d'application et les technologies utilisées. Cela est réalisé en utilisant une API d'AI pour traiter les abstracts et extraire ces informations (voir section "**Automatic Extraction of Application Domains and Technologies from Abstracts**"). À court terme, il faudra être en mesure de trouver une corrélation entre les domaines d'application et les références.

Le dossier **DigitalIdentityWallet_ApplicationDomain_TechnologyUsed** contient les premiers résultats de ces scripts d'extraction. Il faudra réaliser cette extraction chaque année et pour les deux termes.