

ESME SUDRIA

**Étude / Audit Du Projet
&
Cahiers des Charges**

“Lis ça”

Florian Benamrane

Antoine Lorreyte

Thomas Gaudermen

4 Novembre 2022

Tuteurs de projet : M. Ben Romdhanne & Mme. Wafa Belhajali

Superviseur académique : M. Abdussalam Gemal

Etablissement / Formation : ESME SUDRIA / A3MSI

2022 - 2023

SOMMAIRE

I - Étude / Audit du projet	3
1 - Objectifs	3
2 - Analyse de l'existant	3
3 - Propriété intellectuelle	4
4 - Modèles économiques	4
5 - Cibles	5
6 - Stratégie du projet	5
7 - Répartition et réalisation des tâches.....	6
7.1 - Organisation du projet.....	7
7.2 - Planification et ressources	7
8 - Référencement	8
II - Cahier des charges.....	10
1 - Description du graphisme	10
2 - Logo.....	10
3 - Typographie	11
4 - Maquette du site.....	11
5 - Description des pages	14
6 - Nom de domaine	15
Glossaire	16
Annexe 1 : Estimation des coûts	17
Annexe 2 : Déploiement du site localement.....	18
Index des Figures et Tableaux	19

I - Étude / Audit du projet

1 - Objectifs

Dans le cadre du projet de fin d'études, nous avons décidé de réaliser un projet portant sur du "Machine learning", ou apprentissage automatique. L'objectif de celui-ci consiste à manier le "Deep learning" ou l'apprentissage profond, afin de détecter et de traiter du texte au sein d'images.

Cette démarche permettrait de simplifier la lecture d'informations contenues sur des supports physiques non numériques (IBAN, passeport, fichiers médicaux, etc.). Cette solution est pertinente du point de vue de la sécurité afin de lutter contre la fraude, mais également de simplifier le travail des opérateurs de vérification.

2 - Analyse de l'existant

La détection et le traitement des informations contenues au sein d'une image dépendent intrinsèquement du type de document analysé. En effet, il est nécessaire de choisir adéquatement les bibliothèques et dépendances nécessaires en fonction du document à analyser.

Pour les passeports, nous avons déterminé que les données principalement utilisées sont dans une zone appelée "machine-readable zone" (MRZ¹).

L'extraction des données est apparemment possible à l'aide du paquet "PassportEye"². Nous avons mis au point un petit script python qui permet de lire la zone MRZ et de récupérer les champs : "Nationalité", "Prénom", "Nom de famille", "Sexe", "Date de naissance", "Numéro d'identification", "Date d'expiration".

Pour les IBAN, l'extraction des données est apparemment possible à l'aide du paquet "schwifty"³.

Nous avons mis au point un petit script python qui permet de lire la zone MRZ et de récupérer les champs : "numéro IBAN", "code pays", "code banque", "code compte", "numéro BIC".

Pour les données médicales, il est apparemment possible d'extraire les données à l'aide du paquet "med7"⁴.

Nous avons mis au point un script python qui permet de lire une zone de texte et de récupérer les champs : "DRUG", "ROUTE", "DOSAGE", "FREQUENCY", "DURATION".

¹ Council of the European Union - PRADO (2022) Council of the European Union - PRADO - Search by Document Country in the Register of European ID Docs. General Secretariat. Available at: <https://www.consilium.europa.eu/prado/en/prado-documents/FRA/index.html> (Accessed: November 29, 2022).

² PassportEye documentation (no date) PassportEye documentation - PassportEye 1.0.1 documentation. Available at: <https://passporteye.readthedocs.io/en/latest/> (Accessed: November 29, 2022).

³ Schwifty (no date) PyPI. Available at: <https://pypi.org/project/schwifty/> (Accessed: November 29, 2022).

⁴ Kormilitzin (no date) Kormilitzin/MED7, GitHub. Available at: <https://github.com/kormilitzin/med7> (Accessed: November 29, 2022).

3 - Propriété intellectuelle

Le nom du projet et dans un sens plus large, notre marque, permet une exposition et une reconnaissance de nos services, mais également de les distinguer de potentiels concurrents.

Le dépôt d'une marque auprès d'un organisme tel que l'INPI⁵ (Institut national de la propriété industrielle) permet une protection juridique et un monopole d'exploitation sur le territoire français.

Afin de déterminer la disponibilité de notre marque, nous avons utilisé la base de données WIPO⁶ (World Intellectual Property Organization).

4 - Modèles économiques

Notre modèle économique pour le site web reposera sur les points suivants :

- Système d'abonnements ;
- Publicité légère ;
- Vente de statistiques anonymisées (Données à porter commerciales permettant des études d'attractivité commerciale d'un produit.)

Notre modèle économique général reposera sur la disponibilité et la fiabilité par le recours à une innovation comportementale et aux technologies du "cloud" pour l'hébergement de la solution. C'est sur cet aspect que nous jouons afin de générer une marge de bénéfices qui permettront d'amortir les différentes pertes (coût de l'hébergement de la solution).

Nous proposerons trois types d'offres d'abonnements pour utiliser nos services. Une première offre sous forme de solution de cloud public avec une limitation des offres proposées.

Une seconde offre payante sous forme de solution de cloud hybride, où nous offrirons la possibilité à un organisme d'auto-héberger sa base de données à des fins de conformité et de gouvernance de données, le traitement des images se fera dès lors sur notre plateforme en ligne.

Une dernière offre payante sous forme de solution de cloud public, sans limitations, où nous nous chargeons d'héberger l'ensemble des services et délivrons une solution clés en main. L'ensemble des services seront instanciés et cloisonnés par clients pour s'accorder avec le respect de la confidentialité et de discrétion.

⁵ (no date) INPI.fr. Available at: <https://www.inpi.fr/> (Accessed: November 29, 2022).

⁶ WIPO (no date) Global Brand Database. Available at: https://branddb.wipo.int/en/quicksearch?by=brandName&v=&start=0&_id=1669738196256 (Accessed: November 29, 2022).

Une estimation des coûts de fonctionnement à l'aide de la calculette⁷ Azure nous permet d'estimer qu'une machine virtuelle qui répond à nos premières offres de service revient environ à 12.5€ par mois sur trois ans, tandis que pour notre offre inclusive, nous estimons un coût de 13.7€ par mois sur trois ans. Sur une ligne de fonctionnement, nous estimons un investissement minimum de 40€ mensuellement.

Nous pouvons donc justifier notre proposition de tarifs suivante :

- gratuite pour une utilisation limitée dans le temps et dans les services ;
- 25€ par mois pour l'offre sous forme de cloud hybride ;
- 28€ par mois pour l'offre sous forme de cloud public sans limitations.

Une documentation en annexe (Annexe 1) explicitera les estimations de coûts.

5 - Cibles

Notre public principal ciblé est les organismes gouvernementaux (la santé, l'administration, l'économie, la sécurité et la défense), mais aussi les entreprises (les ressources humaines et la sécurité). Un usage pour les particuliers est également envisageable bien que celui-ci reste moins pertinent et à usage variable.

6 - Stratégie du projet

L'ensemble de la réalisation du projet est divisé en trois phases distinctes et échelonnées dans le temps. Un travail de documentation est réalisé tout au long du projet.

Dans un premier temps, l'équipe projet doit réaliser un audit du projet et mettre au point le cahier des charges. Cette première phase se déroule sur une période de deux jours.

Puis l'équipe projet entre dans la seconde phase, celle de conception (générale et détaillée) et de réalisation pour une durée de deux semaines.

Ensuite, l'équipe projet passe dans la dernière phase, celle de déploiement, celle-ci se déroule sur une période d'une semaine. Des tests sont effectués en temps réel et dans des conditions en accord avec les besoins du commanditaire.

La stratégie de réalisation est synthétisée et illustrée ci-dessous.



FIGURE - PRINCIPE DE RÉALISATION DU PROJET

⁷ Pricing calculator: Microsoft Azure (no date) Pricing Calculator | Microsoft Azure. Available at: <https://azure.microsoft.com/en-us/pricing/calculator/> (Accessed: December 4, 2022).

7 - Répartition et réalisation des tâches

Pour partager les documents, chaque membre dispose de sa propre copie et peut la modifier à sa guise en utilisant ses outils habituels. Si plusieurs membres modifient le même objet en parallèle, alors les copies divergent.

Cette divergence n'est pas problématique en soi. Elle implique de programmer des temps de concertation pour échanger sur les différentes modifications réalisées en parallèle et se mettre d'accord sur leur intégration dans la version définitive.

Afin de résoudre les problèmes de divergences entre les documents, nous avons recours à la plate-forme Gitlab. La collaboration au sein de l'équipe projet est ainsi facilitée et optimisée. En effet, il est possible de planifier et d'organiser les tâches de chacun, assurant ainsi une avancée concertée. Le contrôle de version permet également de retrouver des données effacées, mais aussi d'effectuer de nombreuses manipulations, comme la comparaison des sous-parties d'un logiciel travaillées parallèlement.

Lorsque de tels mécanismes sont mis en place, l'organisation, du point de vue opérationnel, devient optimale. Une administration centrale par le chef de projet permet alors de vérifier et de valider les différentes versions pour garantir la cohérence des informations mises à jour. Après validation, toutes les copies ont la même valeur.

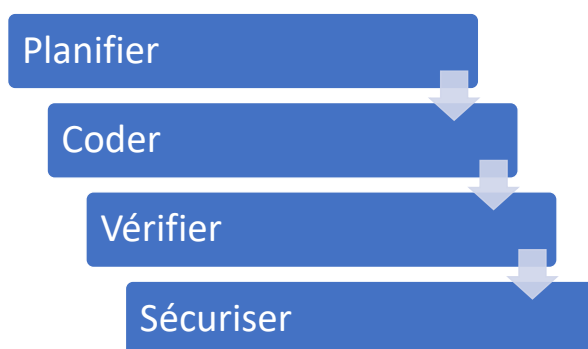


FIGURE - PRINCIPE D'UTILISATION DE GITLAB

7.1 - Organisation du projet

Dans son rôle de maîtrise d'oeuvre, le groupe a la responsabilité de réaliser les activités dont la validation conduira à son transfert en production.

Il est établi par l'équipe projet sous la responsabilité de Lorreyte Antoine (chef de projet), avec les contributions du service de maîtrise d'ouvrage en ce qui concerne la définition du service attendu, la planification et en collaboration avec tous les services de formation concernés (M. BEN ROMDHANNE) en ce qui concerne la validation des scénarios proposés, la validation de la solution réalisée, le suivi du projet et les modalités de communication en cours de projet.

Nous identifions quatre grands jalons clés liés à ce projet qui sont :

- Un travail de recherche pour déterminer le cadre et les objectifs du projet ;
- Un travail de recherche et d'apprentissage de l'utilisation d'un algorithme de Deep Learning qui correspond à notre besoin ;
- Une implémentation d'une solution d'ingénierie qui intégrera notre algorithme de Deep Learning ;
- Des tests de fonctionnement et de validation de la solution réalisée.

7.2 - Planification et ressources

Le projet débute le mardi 17 octobre 2022 et s'achève le 03 avril 2023, soit un peu plus de quatorze semaines. Afin de terminer ce projet dans les délais impartis, il est nécessaire de planifier et gérer les ressources mises à disposition par le biais d'outils de gestion adéquats.

Pour ce faire, nous avons recours à des outils de gestion tels que des diagrammes de GANTT, la plateforme GitLab et Microsoft Teams.

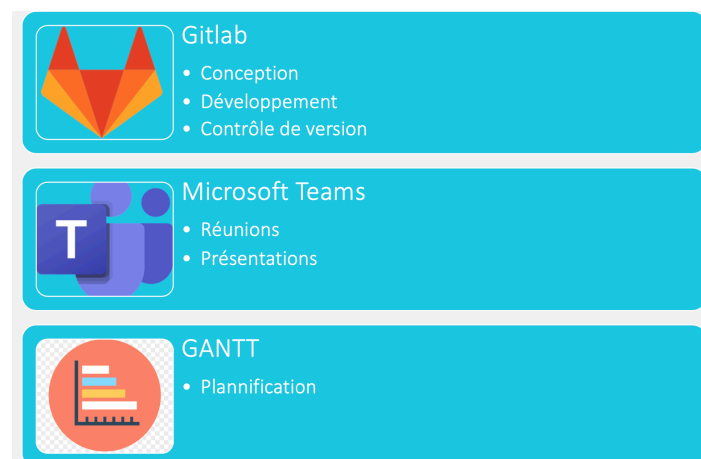


FIGURE - RESSOURCES

Le diagramme de GANTT ci-dessous est un outil efficace afin de donner une vue globale du projet



DIAGRAMME - GANTT PRÉVISIONNEL

8 - Référencement

- Promotion de notre site sur les plateformes des réseaux sociaux de ce dernier.
- Publicité sur Facebook, Instagram et YouTube.
- Possibilité de partager ses progrès à l'aide d'une fonctionnalité dédiée pour tous les utilisateurs.
- Recours à des mots-clés dans les textes de notre site pour améliorer le référencement (recours à la plateforme Google Trends).

Une analyse des termes les plus recherchés en France métropolitaine à l'aide de Google Trends⁸ révèle que notre concept répond à une demande croissante au cours du dernier quinquennat.

⁸ (no date) Google trends. Google. Available at: <https://trends.google.com/trends> (Accessed: December 4, 2022).

Region	Deep learning	Optical character recognition	International Bank Account Number:	Passport	Electronic health record
Île-de-France	1 %	1 %	7 %	90 %	1 %
Provence-Alpes-Côte d'Azur	<1%	1 %	6 %	92 %	1 %
Rhone-Alpes	1 %	1 %	6 %	91 %	1 %
Pays de la Loire	<1%	1 %	5 %	92 %	2 %
Brittany	1 %	1 %	5 %	91 %	2 %
Midi-Pyrénées	1 %	2 %	5 %	90 %	2 %
Languedoc-Roussillon	<1%	1 %	6 %	91 %	2 %
Alsace	1 %	1 %	7 %	89 %	2 %
Nord-Pas-de-Calais	<1%	1 %	6 %	91 %	2 %
Aquitaine	<1%	1 %	7 %	90 %	2 %
Upper Normandy	<1%	1 %	7 %	90 %	2 %
Lorraine	1 %	1 %	7 %	89 %	2 %
Picardy	<1%	1 %	7 %	90 %	2 %
Champagne-Ardenne	1 %	1 %	7 %	87 %	4 %
Lower Normandy	<1%	1 %	7 %	90 %	2 %
Corsica		1 %	9 %	89 %	1 %
Centre-Val de Loire	<1%	2 %	8 %	88 %	2 %
Burgundy	1 %	1 %	8 %	88 %	2 %
Poitou-Charentes	<1%	2 %	7 %	88 %	3 %
Franche-Comté	1 %	1 %	8 %	88 %	2 %
Limousin		1 %	7 %	90 %	2 %
Auvergne	1 %	1 %	9 %	86 %	3 %

FIGURE - STATISTIQUES POUR LA FRANCE

Pour l'hébergement de notre site, nous avons pour objectif de créer notre propre serveur Apache⁹ sur une machine virtuelle basée sur la plateforme Microsoft Azure¹⁰. Cette décision s'appuie également sur la mise à l'échelle automatique ou l'autoscaling, une fonctionnalité offerte par Azure qui permet d'obtenir des performances fiables à moindre coût en augmentant et en diminuant de manière transparente les nouvelles instances en fonction des pics et des baisses de la demande. Cela permet d'assurer une cohérence entre la demande dynamique et imprévisible des applications.

Afin d'optimiser les performances, nous installerons une version légère du système d'exploitation Linux et utiliserons des programmes à code ouvert ne demandant pas beaucoup de ressources.

L'accès au dit système entre les collaborateurs se fera à l'aide du protocole SSH.

Une documentation en annexe (Annexe 2) explicitera la mise en service locale de notre site depuis notre dépôt GITHUB¹¹ : <https://github.com/ESME-MSI/lis-ca>

⁹ Group, D. (no date) Apache, Welcome! - The Apache HTTP Server Project. Available at: <https://httpd.apache.org/> (Accessed: December 4, 2022).

¹⁰ (no date) Google trends. Google. Available at: <https://trends.google.com/trends> (Accessed: December 4, 2022).

¹¹ (no date) GitHub. Available at: <https://github.com/> (Accessed: December 4, 2022).

II - Cahier des charges

1 - Description du graphisme

L'aspect sentimental des couleurs est un facteur important dans la prise de décision, tout comme la popularité d'une couleur ou encore sa capacité à se retrouver dans tous les secteurs d'activité.



FIGURE - PALETTE DE COULEURS

Les teintes choisies pour devenir la couleur de notre site sont celles qui, selon nous, communiquent le mieux ce qui se passe dans notre culture mondiale à ce moment précis, des couleurs qui traduisent l'état d'esprit de l'offre.

2 - Logo

Notre logo vient illustrer la mentalité derrière notre concept, pouvoir adopter une hygiène de vie saine. Le rapport symbolique de la fourchette et des feuilles vient accentuer le concept derrière notre site, tout en restant simple, dynamique et facilement identifiable.



FIGURE - LOGO

3 - Typographie

Pour le choix des polices, nous avons opté pour celles disponibles à usage personnel et commercial tout en conservant un aspect moderne.

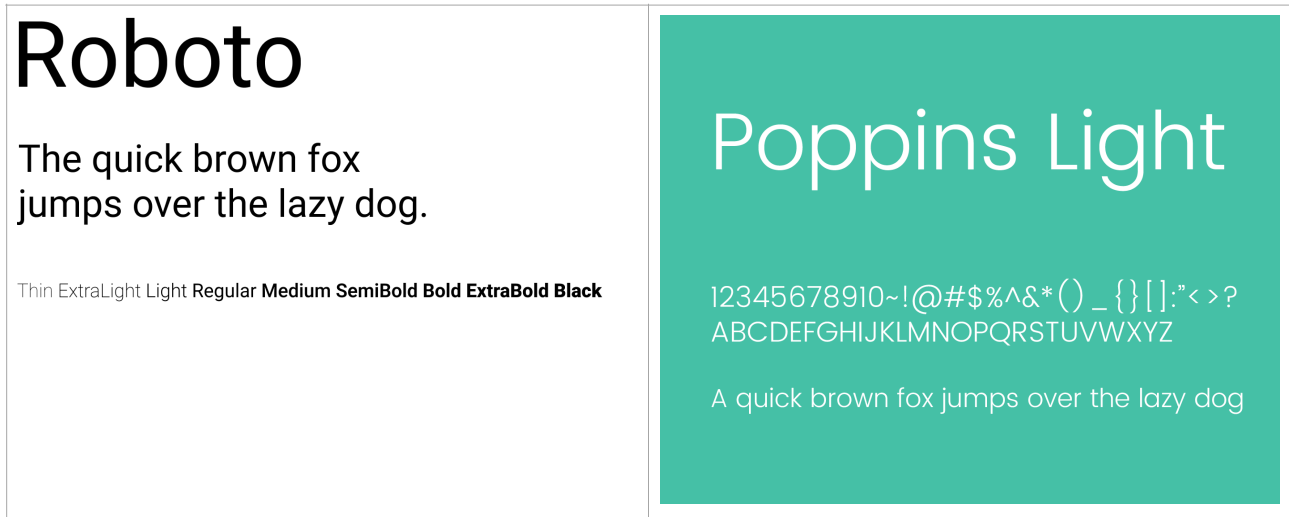


FIGURE - POLICES

4 - Maquette du site

Pages	Maquettes
Accueil	
Caractéristiques	

Pages	Maquettes
Tarifs	 A wireframe of a web page titled "Tarifs". It features a blue header bar with three white rectangular buttons. The page is enclosed in a blue border with a three-dot menu icon in the top-left corner.
Nous contacter	 A wireframe of a web page titled "Nous contacter". It features a blue header bar with a single white rectangular button. The page is enclosed in a blue border with a three-dot menu icon in the top-left corner.
Se connecter	 A wireframe of a web page titled "Se connecter". It features a blue header bar with a white "Header" label. Below the header is a blue box containing a white "Se connecter" label and a white rectangular input field. Below this is another blue box containing a white "S'inscrire" label and a white rectangular input field. At the bottom is a purple box with a white "Footer" label. The page is enclosed in a blue border with a three-dot menu icon in the top-left corner.

Pages	Maquettes
Compte	
Scanner	

5 - Description des pages

La page d'accueil se compose de la manière suivante :

- Header : Logo du site et son nom
- Body :
 - Une barre de navigation (accès aux différentes fonctionnalités du site et redirection vers les pages de création, connexion à un compte)
 - Une description du site.
- Footer : les liens utiles (redirections vers les pages du formulaire de contact et des mentions légales, conditions générales d'utilisation, la politique de confidentialité, et conditions générales de ventes).

La page Se connecter / S' Inscire se compose de la manière suivante :

- Header : Logo du site et son nom
- Body :
 - Un formulaire central d'authentification à l'aide d'un identifiant et d'un mot de passe.
- Footer : les liens utiles (redirections vers les pages du formulaire de contact et des mentions légales, conditions générales d'utilisation, la politique de confidentialité, et conditions générales de ventes).

La page Mon compte se compose de la manière suivante :

- Header : Logo du site et son nom
- Body :
 - Une vue globale des informations liées au client.
 - La possibilité de scanner des documents.
- Footer : les liens utiles (redirections vers les pages du formulaire de contact et des mentions légales, conditions générales d'utilisation, la politique de confidentialité, et conditions générales de ventes).

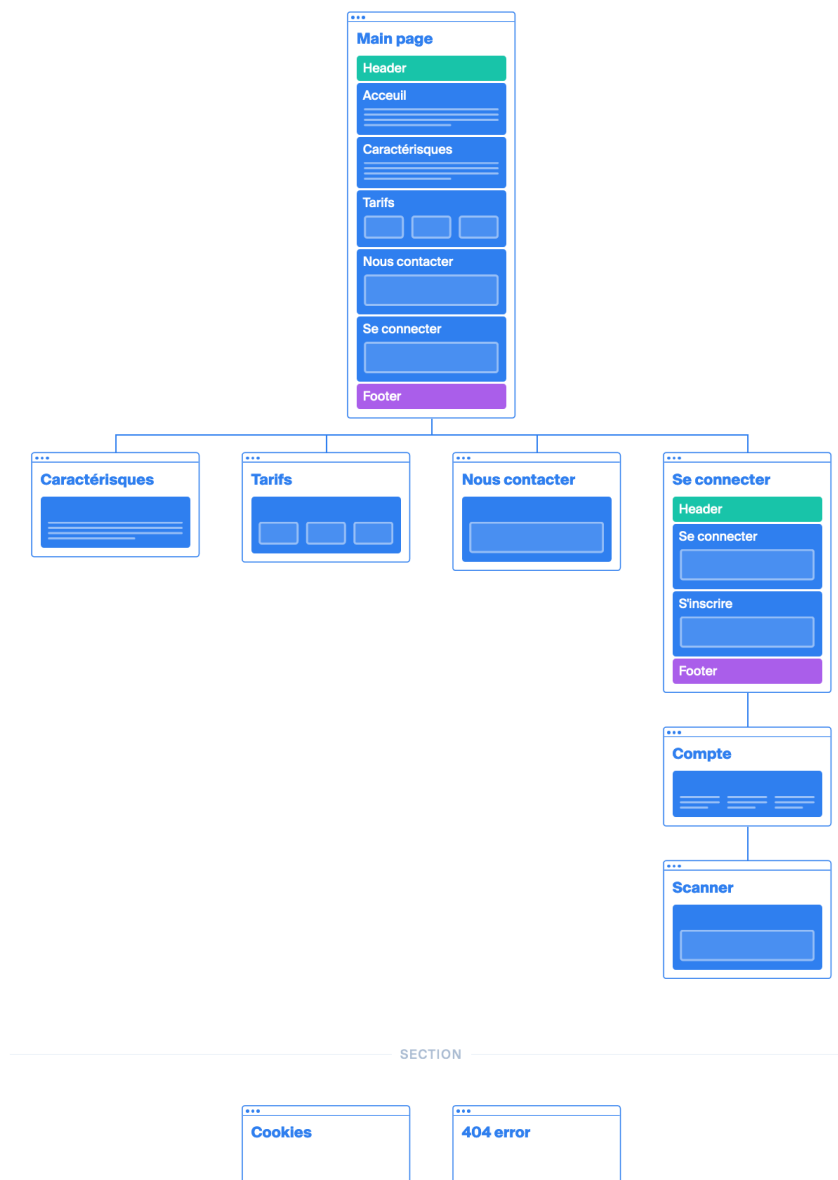


FIGURE - MAQUETTE DU SITE

6 - Nom de domaine

Le nom de domaine choisi est le suivant : lis-ca.

Glossaire

Contrôle de version : Suivi des différentes versions d'un programme, d'un projet, d'un document). C'est un processus permettant de conserver une trace des modifications successives apportées à un fichier numérique (documentation, code source, base de données), à travers un logiciel spécialisé.

Fonction : Sous-ensemble automatisable d'une activité qui compose le processus.

Maîtrise d'ouvrage : Personne morale ou organisme demandeur du projet. C'est le client qui sera l'attributaire final.

Maîtrise d'œuvre : Personne morale ou organisme chargé par la maîtrise d'ouvrage de réaliser le projet.

Référentiel : Ensemble de bases de données contenant les "références" d'un système d'information. Il s'agit d'informations qui seront utilisées lorsque l'on doit faire évoluer une application : on parle alors « d'administration des données » ; ce sont des définitions, et aussi des indications sur le format de la donnée (« typage »), les conditions de sa mise à jour, son « propriétaire » (personne ou entité habilitées à la mettre à jour).

Socle interministériel des logiciels libres (SILL) : Catalogue de référence des logiciels libres pour les administrations en France. Celui-ci n'a pas de caractère obligatoire et ce sont des recommandations d'outils.

Annexe 1 : Estimation des coûts

Offre de cloud hybride :

Microsoft Azure Estimate						
Your Estimate						
Service category	Service type	Custom name	Region	Description	Estimated monthly cost	Estimated upfront cost
Compute	Virtual Machines		France Central	1 B2s (2 vCPUs, 4 GB RAM) (3 year reserved), Linux, (Pay as you go); 0 managed disks – E2; Inter Region transfer type, 5 GB outbound data transfer from France Central to East Asia	€12,45	€0,00
Support			Support		0 €0,00	
			Licensing Program	Microsoft Customer Agreement (MCA)		
			Billing Account			
			Billing Profile			
			Total		12,4480884828084 €0,00	
Disclaimer						
All prices shown are in Euro Zone – Euro (€) EUR. This is a summary estimate, not a quote. For up to date pricing information please visit https://azure.microsoft.com/pricing/calculator/						
This estimate was created at 12/4/2022 12:02:50 PM UTC.						

Offre de cloud public sous forme de Saas :

Microsoft Azure Estimate						
Your Estimate						
Service category	Service type	Custom name	Region	Description	Estimated monthly cost	Estimated upfront cost
Compute	Virtual Machines		France Central	1 B2s (2 vCPUs, 4 GB RAM) (3 year reserved), Linux, (Pay as you go); 1 managed disk – E3; Inter Region transfer type, 5 GB outbound data transfer from France Central to East Asia	€13,72	€0,00
Support			Support		0 €0,00	
			Licensing Program	Microsoft Customer Agreement (MCA)		
			Billing Account			
			Billing Profile			
			Total		13,7176244289493 €0,00	
Disclaimer						
All prices shown are in Euro Zone – Euro (€) EUR. This is a summary estimate, not a quote. For up to date pricing information please visit https://azure.microsoft.com/pricing/calculator/						
This estimate was created at 12/4/2022 12:04:24 PM UTC.						

Annexe 2 : Déploiement du site localement

Le dépôt principal de notre projet est sur la plateforme collaborative GITHUB, accessible par le biais du lien ci-après : <https://github.com/ESME-MSI/lis-ca>

Sous un système d'exploitation UNIX, les commandes suivantes permettront de récupérer une version du dépôt GITHUB localement :

```
cd ~ && git clone https://github.com/ESME-MSI/lis-ca.git && cd lis-ca
```

Nous utilisons des modules externes dans notre projet sous python, plus particulièrement :

- "schwifty" : pour la gestion des IBAN et des BIC;
- "PassportEye" : pour la gestion des passeports, plus spécifiquement de la zone MRZ;
- "Med7" : pour la gestion des dossiers médicaux.

Ces dépendances sont installables à l'aide de "pip"¹², il s'agit d'un utilitaire en ligne de commande permettant d'interagir avec notre dépôt et facilitant l'installation de paquets, la gestion des versions et la gestion des dépendances.

Maintenant, pour exécuter notre instance docker, il suffit de taper :

```
cd Docker-Projet && sh Make.sh
```

Il suffit dès lors de pointer votre navigateur sur l'adresse ci-après:

```
http://localhost:8080
```

¹² <https://pypi.org/project/pip/>

Index des Figures et Tableaux

Figure - Principe de réalisation du projet	5
Figure - Principe d'utilisation de GitLab	6
Figure - Ressources.....	7
Diagramme - Gantt prévisionnel	8
Figure - Statistiques pour la France	9
Figure - Palette de couleurs.....	10
Figure - Logo	10
Figure - Polices	11
Figure - Maquette du site.....	15