PERLIC Vladana | LIU Zixuan

Master 2 Industrie de la Langue

Parcours Professionnel | Semestre 2

UE: Projet Professionnel

Rapport de projet

Data CNIL

Année universitaire 2023-2024

Table des matières

1. Présentation des parties

a. Maître d'œuvre

i. Coordonnées

Zixuan LIU
zixtrbl@gmail.com

Vladana PERLIC
vladana.perlic@gmail.com

ii. Description

Notre équipe est composée de deux étudiantes en 2ème année de Master professionnel en Industries de la langue. Nous avons réalisé ce travail dans le cadre d'un projet professionnel, encadré par Monsieur Thomas Lebarbé.

b. Maître d'ouvrage

i. Coordonnées

Emilie Masson
emilie.masson@inria.fr

ii. Description

- DPO Adjointe INRIA

2. Présentation du projet

a. Description et contexte de la demande

La commanditaire nous a sollicités pour améliorer la plateforme prototype en intégrant un moteur de recherche performant, une mise à jour automatique, une fonctionnalité d'exportation des résultats de recherche, une correction automatique des erreurs de balisage ainsi que des outils TAL et des représentations dynamiques pour explorer les délibérations de la CNIL.

Cette plateforme prototype a été développée pour répondre aux besoins des experts en protection des données en matière de recherche et de filtrage des délibérations de la CNIL. Ces recherches peuvent être effectuées en fonction de la nature de la délibération, du type de document, des dates de publication, des mots contenus dans les titres et des textes des délibérations. La plateforme exploite l'ensemble des documents XML du DATA CNIL, accessibles au public sur https://echanges.dila.gouv.fr/OPENDATA/CNIL.

En plus des fonctionnalités de recherche, la plateforme propose un concordancier et un graphique illustrant le nombre de délibérations trouvées par année. La plateforme a été mise à jour manuellement pour la dernière fois en date du 4 septembre 2023, et depuis lors, de nouvelles délibérations ont été publiées.

Afin de garantir que toutes les délibérations soient exploitables peu de temps après leur publication sur le site de la CNIL, et pour éviter de dépendre d'une mise à jour manuelle, le commanditaire souhaite intégrer une fonction de mise à jour automatique dans la plateforme.

Par ailleurs, des erreurs de balisage sont présentes dans les fichiers originaux, nécessitant ainsi la mise en place d'un correcteur automatique des erreurs de balisage.

Le moteur de recherche actuel de la plateforme n'est pas entièrement performant. Il ne prend pas en charge des fonctionnalités essentielles telles que la recherche d'expressions complètes, similaire à celle de Google en utilisant des caractères spéciaux (par exemple, en mettant l'expression entre guillemets). De plus, il présente des problèmes avec les acronymes, notamment lorsque ces derniers sont écrits avec des points entre les lettres (ex. C.N.R.S.). Cette particularité peut entraîner une non-correspondance lors de la recherche de "CNRS", qui ne retournerait pas les documents contenant "C.N.R.S.", bien que sémantiquement il s'agisse du même terme.

La commanditaire souhaite que nous trouvions une solution permettant de remédier à ces problèmes.

b. Public visé

La plateforme DATA CNIL s'adresse principalement aux experts en protection des données, tels que les responsables de la conformité, les avocats spécialisés, et les professionnels de la sécurité de l'information.

c. Langue

La plateforme DATA CNIL ainsi que l'ensemble des données de la DATA CNIL sont en français. Étant donné que l'utilisateur effectuant des recherches sur les données de la DATA CNIL doit être francophone, nous n'avons pas jugé nécessaire de créer une version en anglais.

d. Budget

Ce projet professionnel est mené dans le cadre d'un projet de fin d'année, et son budget ne peut être évalué en termes monétaires. Nous exprimons son coût en fonction des éléments suivants :

- Nombre de personnes : 2
- Durée : 2 mois
- Compétences requises : niveau BAC+5 en Traitement automatique des langues

3. Solutions proposées et réalisées

Dans cette section, nous passons en revue les solutions techniques proposées dans le cahier des charges et ce qui a été effectivement réalisé. Nous abordons les principales fonctionnalités telles que la mise à jour automatique de la base de données, la correction des erreurs de balisage, l'amélioration du moteur de recherche, ainsi que la sauvegarde et l'export des résultats. Nous examinons comment ces solutions ont été mises en œuvre par rapport aux objectifs initiaux du projet.

a. Solution technique

Fonctionnalités principales:

1. **Mise à jour automatisée de la base de données:**

    - *Proposition :* La plateforme sera dotée d'un mécanisme de mise à jour automatique quotidienne de la base de données, assurant que les dernières délibérations de la CNIL soient instantanément intégrées à la plateforme dès leur publication.

    - *Réalisation:* La mise à jour automatique de la base de données a été implémentée de manière hebdomadaire, avec la possibilité pour l'utilisateur d'effectuer une mise à jour manuelle à tout moment en un seul clic sur le bouton dédié. Nous avons opté pour une mise à jour hebdomadaire de la base de données, car la CNIL ne publie pas de délibérations si souvent que cela nécessiterait une mise à jour quotidienne. De plus, nous évitons ainsi d'alourdir inutilement la plateforme. Dans des cas exceptionnels, il est toujours possible de mettre à jour la base de données manuellement en cliquant sur le bouton prévu à cet effet.

2. **Correction automatique des erreurs de balisage:**

    - *Proposition :* Un module de correction automatique des erreurs de balisage sera intégré, permettant de rectifier les anomalies présentes

dans les fichiers originaux. Cela garantira la cohérence et l'intégrité des données analysées.

- *Réalisation :* La mise en œuvre de la correction automatique des erreurs de balisage a été intégrée, offrant la capacité d'éliminer les doublons et d'unifier les catégories de certaines délibérations au sein de la table "Deliberation" de la base de données. Cette fonctionnalité a été réalisée au moyen de requêtes SQL visant à actualiser les valeurs de la colonne "NatureDeliberation", substituant des occurrences particulières par d'autres valeurs préalablement définies dans le code.
-

3. **Moteur de recherche performant:**

- *Description:* La plateforme sera équipée d'un moteur de recherche performant, incluant une correction automatique des fautes de frappe et une fonction de recherche d'expressions complètes. Par exemple, l'utilisation de l'algorithme de Levenshtein pour améliorer la recherche de termes proches.

- *Réalisation :*

  Opérateurs de recherche : Le moteur de recherche a été considérablement amélioré pour une recherche avancée. Il prend désormais en charge divers opérateurs de recherche avancés, tels que la virgule (ET logique), AND, OR, % (joker). Si plusieurs mots sont séparés par des espaces, ils sont considérés comme une expression complète. La recherche se fait en texte plein, sans distinction majuscules/minuscules.

  Champ de recherche "Titre ou texte contiennent" : A la demande de la commanditaire après la réalisation du cahier des charges, nous avons également ajouté un champ de recherche permettant de rechercher des termes présents soit dans le titre, soit dans le contenu des délibérations.

  La correction automatique des erreurs de frappe : Cette fonctionnalité a finalement été abandonnée afin de préserver la précision des résultats, en raison de son potentiel impact sur la pertinence des recherches. Pour que cette fonctionnalité soit opérationnelle, l'algorithme aurait dû comparer les mots de la requête avec ceux présents dans son dictionnaire, et rechercher des termes dans les délibérations où seulement une ou deux lettres diffèrent du terme présent dans la requête. En pratique, cela signifiait que si l'utilisateur saisissait

"cnrs", le moteur de recherche aurait pu retourner des résultats pour "cnrd".

4. **Sauvegarde et export des résultats:**

    - *Description:* La solution permettra la sauvegarde des résultats de recherche et offrira la possibilité d'exporter ces résultats au format HTML ou fichier texte.

    - *Réalisation :* Nous avons intégré un bouton permettant d'exporter et de télécharger les résultats filtrés de la recherche au format HTML. Il n'est pas intéressant d'exporter les résultats en format txt car ils ne seraient pas très lisibles.

5. **Outils TAL et représentations dynamiques:**
    - *Description:* La plateforme intégrera un concordancier pour le close-reading, permettant aux utilisateurs d'effectuer une analyse approfondie des délibérations. De plus, des représentations dynamiques, telles qu'un graphique statistique par année, seront fournies pour faciliter le distant reading.

    - Réalisation : Nous avons intégré deux outils de Traitement Automatique du Langage (TAL) sur la plateforme : un générateur de nuage de mots et un classificateur.

        Le générateur de nuage de mots est conçu pour produire une représentation visuelle des mots les plus fréquents à partir du contenu des délibérations. Les données extraites sont stockées dans un tableau, et nous avons mis en place une fonction permettant de calculer la fréquence des mots uniques dans ce tableau. Le code HTML et JavaScript qui suit vise à créer cette visualisation en utilisant la bibliothèque D3.js. Nous avons utilisé la fonction 'showNewWords()' pour actualiser le nuage de mots avec de nouveaux termes de manière cyclique.

        Le classificateur est conçu pour analyser et catégoriser les délibérations de la Commission Nationale de l'Informatique et des Libertés (CNIL) en tant que favorables, défavorables ou neutres. Nous avons créé une interface utilisateur avec des onglets pour différentes catégories : "Favorable", "Défavorable", "Neutre". Chaque onglet est associé à un contenu vide initial qui sera rempli dynamiquement lors du clic sur l'onglet correspondant. Le script JavaScript utilise jQuery pour

détecter les clics sur les onglets. Lorsqu'un onglet est cliqué, une requête AJAX est envoyée au serveur pour récupérer le contenu à afficher dans l'onglet à partir d'un fichier "classify_helper.php" en fonction de la catégorie sélectionnée. Le contenu récupéré est inséré dynamiquement dans l'onglet correspondant de la page HTML, remplaçant le contenu vide initial.

b. Solution esthétique

- Palette de Couleurs

Notre commanditaire a choisi un arrière-plan en rouge clair. Les fonds clairs sont choisis pour maintenir un contraste élevé entre le texte et l'arrière-plan, ce qui est crucial pour garantir l'accessibilité et la lisibilité du contenu. Nous avons ajusté la palette de couleurs en adaptant les règles CSS en conséquence.

- Icônes Intuitives
  Nous avons utilisé des icônes intuitives pour améliorer la compréhension de l'interface utilisateur, notamment des boutons pour visualiser les résultats et pour exporter ces derniers. En choisissant des icônes bien conçues, nous rendons les actions et les fonctionnalités plus évidentes pour les utilisateurs, ce qui améliore l'expérience globale de navigation.

4. Spécificités techniques

a. Langages et outils

Pour la réalisation de ce projet, nous utiliserons les langages et les outils suivants:

1. HTML version 5.0:
   Utilisé pour structurer le contenu des pages web afin d'afficher les résultats de recherche de manière organisée. Chaque résultat de recherche est encapsulé dans des balises HTML appropriées.

2. PHP version 8:
   Un langage de programmation côté serveur couramment utilisé pour développer des applications web. Dans le contexte de ce projet, PHP est utilisé pour créer la partie backend de l'interface utilisateur. Il gère les requêtes provenant du frontend (HTML, JavaScript) et interagit avec la base de données pour récupérer, mettre à jour ou supprimer des données

3. MySQL version 8.0:
   MySQL utilisé comme système de gestion de base de données relationnelle pour stocker de manière structurée les données relatives aux délibérations et effectuer des mises à jour.

4. Javascript version ES6:
   Utilisé pour développer l'interface utilisateur pour la recherche et l'exploration des délibérations et pour créer des visualisations dynamiques.

5. Python version 3 :
   utilisé pour développer le script permettant la mise à jour automatique des délibérations de la CNIL.

b. Base de données

Nous avons utilisé PhpMyAdmin, application web open source écrite en PHP, pour gérer des bases de données MySQL à l'aide d'une interface graphique conviviale. Nous avons effectué diverses opérations sur les bases de données, telles que la création de tables, l'ajout, la suppression et la modification de données, l'exécution de requêtes SQL.

Notre base de données comprend actuellement 5 tables principales. Parmi celles-ci :

- la table 'Deliberation' qui comporte 16 colonnes, où sont stockées les informations essentielles relatives aux délibérations
- la table 'DTC_fichiers_open' stocke les informations relatives aux dates de mise à jour des délibérations
- la table 'token' calcule la fréquence des mots et le nombre d'occurrence des textes
- la table 'Token2Deliberation' contient l'identifiant du token associé à une délibération, l'identifiant de la délibération à laquelle le token est associé ainsi que le nombre d'occurrences du token dans la délibération
- la table 'MisesAJour' enregistre la date de la mise à jour chaque semaine

c. Hébergement

La plateforme en ligne est hébergée sur le serveur pédagogique du département Informatique intégrée en Langues, Lettres et Langage (I3L) (accessible à l'adresse http://i3l.univ-grenoble-alpes.fr). Doté de performances optimales, le serveur I3L assure la disponibilité continue de la plateforme, garantissant ainsi un accès fiable pour les utilisateurs. La sécurité des données est une priorité, avec des mesures de protection avancées mises en place pour préserver l'intégrité et la confidentialité des informations sensibles.

d. Espace de stockage

Au total, une allocation de 20MB est réservée pour stocker les fichiers xml de data CNIL. Nous anticipons une croissance de cet espace de stockage au fur et à mesure de la mise à jour des données.

e. Compatibilité

Le bon fonctionnement de la plateforme repose sur le navigateur utilisé pour accéder à la plateforme. En prenant en compte les navigateurs les plus répandus, à savoir Google Chrome, Mozilla Firefox, Internet Explorer, Safari et Opera, nous nous assurons que les utilisateurs peuvent accéder à toutes les fonctionnalités de la plateforme avec efficacité, quel que soit leur choix de navigateur.

6. Planning

a. Grandes dates

Date de début	Étape	Description
03/01/2024	Conception	Base de données, architecture du site…
08/01/2024	Réalisation	Implémentation de la mis à jour automatique de la base de données
08/01/2024		Correction automatique des erreurs de balisage
15/01/2024		Implémentation d'un moteur de recherche performant, concordancier et représentations dynamiques
29/01/2024		Sauvegarde et export des résultats
31/01/2024	Tests	Phase de tests sur la plateforme

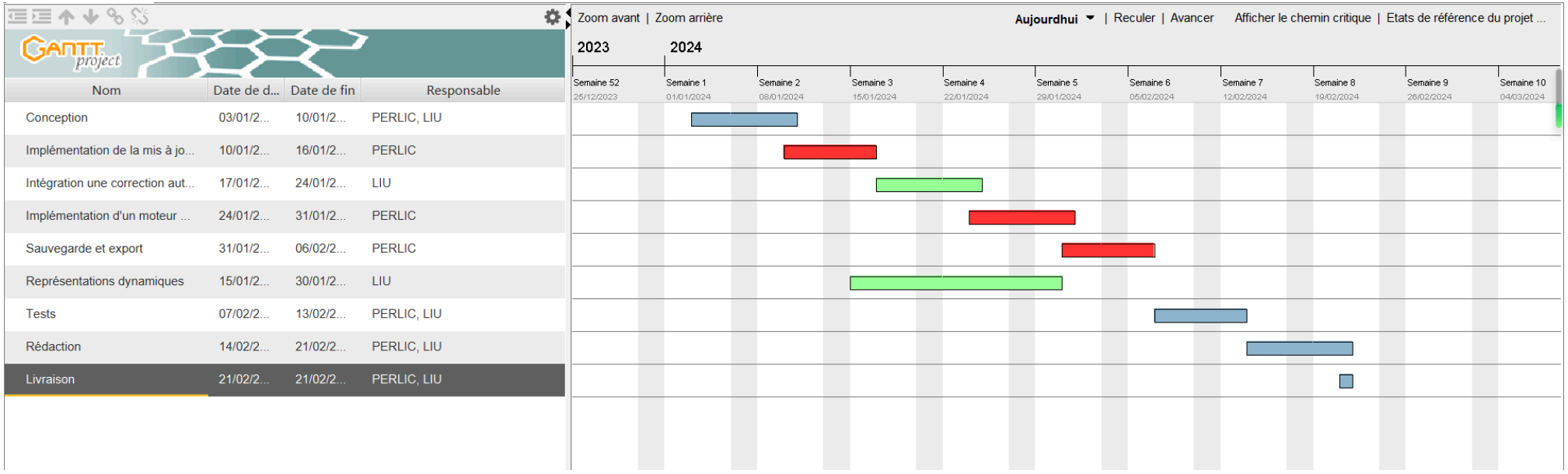
08/02/2024	Rédaction	Manuel d'utilisation, documentation, …
23/02/2024	**Livraison du produit**	

Nous avons eu une période de 7 semaines pour livrer le produit, et nous avons consciencieusement géré chaque étape pour garantir son bon déroulement. Malgré les défis rencontrés en cours de route, nous avons maintenu le cap sur les grandes dates prévues. Chaque fois que des obstacles se sont présentés, nous avons pris des mesures pour les surmonter et assurer l'achèvement des tâches dans les délais impartis.

b. Gantt

Définition de la tâche	Date de début	Date de fin	Jours	Personne en charge
Conception de la base de données	03/01/2024	05/01/2024	2	Vladana PERLIC
Conception de l'architecture du site	03/01/2024	05/01/2024	2	Zixuan LIU
Implémentation de la mis à jour automatique de la base de données	08/01/2024	12/01/2024	5	Vladana PERLIC
Correction automatique des erreurs de balisage	08/01/2024	12/01/2024	5	Zixuan LIU
Implémentation d'un moteur de recherche performant	15/01/2024	26/01/2024	10	Vladana PERLIC
Sauvegarde et export des résultats	29/01/2024	30/01/2024	2	Vladana PERLIC
Concordancier et représentations dynamiques	15/01/2024	30/01/2024	12	Zixuan LIU
Tests	31/01/2024	07/02/2024	6	PERLIC & LIU

Rédaction	08/02/2024	15/02/2024	6	PERLIC & LIU
Manuel d'utilisation	15/02/2024	19/02/2024	3	PERLIC & LIU
Documentation	19/02/2024	22/02/2024	4	PERLIC & LIU
Livraison	23/02/2024	23/02/2024	1	PERLIC & LIU



Nom	Date de d...	Date de fin	Responsable
Conception	03/01/2...	10/01/2...	PERLIC, LIU
Implémentation de la mis à jo...	10/01/2...	16/01/2...	PERLIC
Intégration une correction aut...	17/01/2...	24/01/2...	LIU
Implémentation d'un moteur ...	24/01/2...	31/01/2...	PERLIC
Sauvegarde et export	31/01/2...	06/02/2...	PERLIC
Représentations dynamiques	15/01/2...	30/01/2...	LIU
Tests	07/02/2...	13/02/2...	PERLIC, LIU
Rédaction	14/02/2...	21/02/2...	PERLIC, LIU
Livraison	21/02/2...	21/02/2...	PERLIC, LIU