Page de garde -> Cassie

# Remerciements

# Historique du document :

|  |  |  |  |
| --- | --- | --- | --- |
| **Date** | **Numéro de version** | **Chapitres concernés** | **Cause de la modification** |
| 04/02 | v1.0 | Tous | Création du rapport |
| 06/02 | V1.1 | Parties I, II et III | Ajout du cahier des charges, des rôles et des documents |
| 23/02 | V1.2 | Parties IV et VI | Ajout du SADT et description de la gestion de configuration |
| 09/03 | V1.3 | Parties I, II, III, V et VI | Modifications et ajouts |
|  |  |  |  |

Table des matières

[Remerciements 3](#_Toc413767559)

[Historique du document : 4](#_Toc413767560)

[Tables des figures 7](#_Toc413767561)

[I. Objet et but du document 8](#_Toc413767562)

[1) Objectif du document 8](#_Toc413767563)

[2) But du document 8](#_Toc413767564)

[3) Domaine d’application 8](#_Toc413767565)

[4) Problématique 8](#_Toc413767566)

[II. Documents applicables, références et terminologie 10](#_Toc413767567)

[1) Documents applicables 10](#_Toc413767568)

[2) Documents de références 10](#_Toc413767569)

[3) Terminologie 11](#_Toc413767570)

[4) Présentation du document 11](#_Toc413767571)

[III. Organisation du projet 12](#_Toc413767572)

[1) Equipe de développement 12](#_Toc413767573)

[2) Planification 12](#_Toc413767574)

[IV. Démarche de développement 13](#_Toc413767575)

[1) Processus de développement 13](#_Toc413767576)

[2) Application du processus de développement 13](#_Toc413767577)

[a) Définir et valider le domaine 13](#_Toc413767578)

[b) Collecter les données 13](#_Toc413767579)

[c) Extraire les données 14](#_Toc413767580)

[d) Administrer les données 15](#_Toc413767581)

[e) Exploiter les données 15](#_Toc413767582)

[f) Valoriser les données 15](#_Toc413767583)

[V. Matériels et outils 16](#_Toc413767584)

[1) Plateforme de développement 16](#_Toc413767585)

[2) Outils utilisés 16](#_Toc413767586)

[3) Langages utilisés 16](#_Toc413767587)

[VI. Gestion de configuration 17](#_Toc413767588)

[VII. Assurances et contrôle qualité 20](#_Toc413767589)

[1) Revues 20](#_Toc413767590)

[2) Contrôle qualité 20](#_Toc413767591)

[Bilan 21](#_Toc413767592)

[Bilan par rapport au client 21](#_Toc413767593)

[Bilan pour l’équipe 21](#_Toc413767594)

[Sources 22](#_Toc413767595)

[Annexe 23](#_Toc413767596)

# Tables des figures

[Figure 1 Interface en local de GitHub 17](#_Toc413765928)

[Figure 2 Interface sur le Web - Répertoire 18](#_Toc413765929)

[Figure 3 Interface sur le Web - Détails des sychronisations 18](#_Toc413765930)

[Figure 4 Interface sur le Web - Gestion des collaborateurs 19](#_Toc413765931)

# Objet et but du document

## Objectif du document

Au cours du 2ème semestre de Master 1 Statistique et Informatique Décisionnelle, un projet de "Conception d'un système de recherche d'information décisionnelle" nous a été confié. Ce projet, dit Bureau d’études, a été réalisé en groupe de cinq étudiants.

Ce document est le résultat de notre travail effectué tout au long du bureau d’études. Il est principalement destiné à nos enseignants ayant le suivi notre projet tout au long de son développement, Mme Wahiba BAHSOUN, Mr Bernard DOUSSET et Mr Riad MOKADEM. Ils tiennent également le rôle de client, et notre groupe celui de fournisseur.

L’objectif fondamental de ce projet est de mettre en pratique les connaissances théoriques acquises durant notre formation, plus spécifiquement en génie logiciel, fouille de données, entrepôts de données et bases de données. En pratique, il a pour objectif de collecter un grand nombre d'articles sur un domaine spécifique, pour par la suite effectuer des analyses statistiques au moyen du logiciel Tetralogie.

## But du document

Le but de ce rapport est de détailler et d’expliquer toutes les étapes de développement de notre projet. Pour cela, nous allons d’abord indiquer les documents de référence qui nous ont été utiles, la description de notre équipe de projet et le planning que l’on a suivi durant celui-ci. Ensuite, nous allons expliquer toutes les étapes suivies lors du développement du projet, afin d’aboutir au résultat attendu après plus de deux mois de travail.

Toute personne n’ayant aucune connaissance en Génie Logiciel, Fouille de données ou en Tetralogie doit pouvoir comprendre ce rapport.

## Domaine d’application

Ce bureau d’études a une visée informatique et statistique. En effet, le projet sur lequel nous travaillons requiert nos connaissances en bases de données et analyses statistiques. Nous utilisons également plusieurs outils et langages informatiques. Même si nous réalisons notre étude centrée sur le domaine du sport, notre projet pourrait être adapté à un autre secteur (ex : l’aéronautique).

## Problématique

Notre groupe a un intérêt particulier dans le domaine du sport. Nous voulions donc chercher un sujet pertinent et d'actualité. Au départ, nous avions choisi de travailler sur la psychologie sportive. Suite à de nombreux résultats lors d'une recherche dans les bases de données, nous avons décidé d'affiner notre sujet. Nous avons donc choisi l'étude de la préparation mentale des sportifs de haut niveau.

Y-a-t-il des différences entre la préparation mentale des sportifs par pays ? Les méthodes utilisées sont-elles identiques dans les pays dits émergents et développés ? Les managers préparent-ils mentalement les sportifs de la même façon ? S'il s'agit d'un sport collectif ou individuel ? Ou encore est-ce que la préparation mentale individuelle des joueurs d'une équipe est identique à celle des sportifs pratiquant un sport individuel ? Toutes ces questions ont suscité un vif intérêt pour notre groupe.

1. Documents applicables, références et **terminologie**

## Documents applicables

Le 19 janvier 2015, nous avons pris connaissance du cahier des charges. Ce cahier des charges étant le suivant :

Le cahier des charges contribue à la création d’un cadre formel de type client-fournisseur où les enseignants jouent le rôle du client, et les étudiants, celui du fournisseur.

Ce document précise l’organisation de l’équipe de développement et du fonctionnement du bureau d’études.

L’objectif de ce bureau d’études est de développer un système d’aide à la décision qui permet de faire des analyses détaillées sur des points d’intérêt à partir de données textuelles issues des bases de données en ligne.

Ce système va recouvrir plusieurs étapes du processus décisionnel telles que :

1. Recherche et collecte de données
2. Préparation de données
3. Stockage de données
4. Requêtage
5. Reporting (diffusion des connaissances)

Nous regroupons ces étapes en quatre grandes parties :

* Recherche d’informations
* DataMining
* Data Warehouse
* Reporting

Ces activités seront supportées par les méthodes de génie logiciel :

* Processus de développement
* Gestion de projet
* Gestion de configuration
* Assurance de qualité

## Documents de références

L’élaboration de ce rapport a nécessité les supports de cours de Licence 3 et de Master 1 Statistique et Informatique Décisionnelle. Les enseignements utilisés dans ce projet sont :

* Data-Mining, cours de M. Bernard DOUSSET
* Génie Logiciel, cours de Mme Wahiba BAHSOUN
* Langage PERL, cours de M. Jérôme FARINAS
* Concepts Fondamentaus des Bases de Données, cours de M. Franck MORVAN
* Data Warehouse, cours de M. Gilles HUBERT
* Gestion de Projet, cours de Mme Wahiba BAHSOUN
* Gestion des Risques, cours de Mme Véronique LAURENT

## Terminologie

Termes utilisés dans le contexte de ce projet :

**Préparation** : Action de préparer quelqu'un à, pour quelque chose.

**Mental** : Qui se passe exclusivement dans l'esprit, intérieurement, sans s'exprimer ou se manifester extérieurement.

**Sportif** : Qui pratique un ou plusieurs sports.

**SADT (Structured Analysis and Design Technic)** : méthode d'analyse par niveaux successifs d'approche descriptive d'un ensemble quel qu'il soit.

**SQL (Structured Query Language)** : langage informatique normalisé servant à exploiter des bases de données relationnelles.

**PERL (Practical Extraction and Report Language)**

## Présentation du document

Présentation du contenu, l’objet des chapitres constituant le document

NE PAS DIRE CE QUI A DEJA ETE FAIT AVANT mais ce qui va venir !!!

Décrire les chapitres suivant en réutilisant les noms.

Dans le chapitre Organisation du projet, nous verrons les ressources humaines et ~~matérielles~~ utilisées lors de ce projet. Grâce à des outils comme les diagrammes de GANTT et PERT, on montrera le planning prévisionnel et réalisé.

Dans le chapitre Démarche de développement, nous expliquerons les différentes étapes à suivre tout au long du Bureau d’études, de la définition du domaine à la valorisation des données.

# Organisation du projet

## Equipe de développement

Notre groupe est composé de 5 étudiants de Master 1 SID. Chaque membre s’est vu confier un rôle bien particulier dans la réalisation de ce bureau d’études :

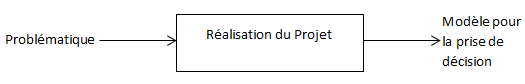
* Marine ISSERTES, chef de projet, chargée de superviser l’avancement du bureau d’études, d’attribuer les différentes tâches à réaliser à l’équipe de projet et gérer le planning.
* Cassie CHAUSSE, responsable configuration, chargée de la gestion des versions du travail produit (avec GitHub).
* Mélanie MOUCHARD, responsable rédaction, chargée de rédiger le rapport en parallèle du développement du projet.
* Thibault LEJAILLE, responsable qualité, chargé de tester le travail produit afin d’obtenir un résultat de qualité.
* Marie DUCHATEAU, responsable développement, chargée de concevoir et réaliser le code en langage Perl.

## Planification

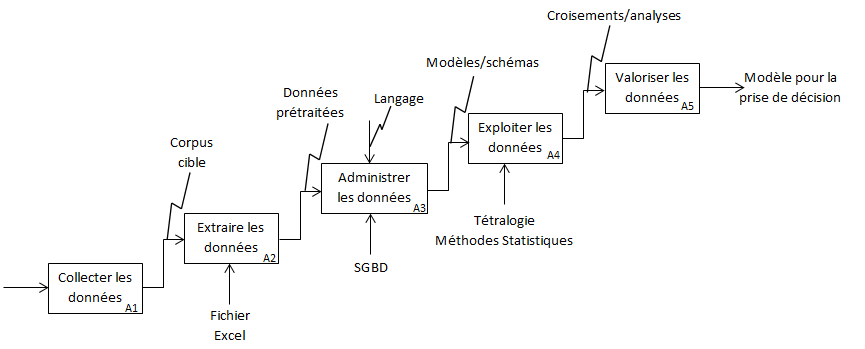
Gantt (en heures), Pert

# Démarche de développement

### Processus de développement



A-0 : Réalisation du projet



A0 : Réalisation du projet

### Application du processus de développement

### Définir et valider le d**omaine**

### Collecter les données

Cette étape permet de rechercher l’information et d’identifier celle qui est utile dans les sources sélectionnées pour alimenter l’analyse. Nous définirons un domaine d’analyse pour chaque groupe.

Recherches effectuées, premiers résultats :

sport\* OR coach\* (mental\* OR coach\* OR psycholog\*)

PubPsych = 21723 résultats

PubMed = 3975 résultats

Web of Sciences = 12 857 résultats (à revoir)

//TODO revoir la requête -> affiner pour avoir moins de résultats

comportement sport\* OR behav\* sport\* OR coach\* (coach\* OR mental\* OR psycholog\*)

PubPsych = 8745 résultats

### Extraire les données

Chaque groupe doit définir des métadonnées pour la base étudiée. Une métadonnée est un ensemble structuré d’information servant à définir et à décrire l'article auquel elles font référence.

Aussi, un descripteur paramétrable de la structure des bases qui s’adapte fidèlement à chaque cas doit être conçu. Ce descripteur définit les différents champs de la base, en identifiant leur bannière, leurs séparateurs, leur utilité et les différents types d’information qu’ils contiennent. Il permet aussi de repérer le début de chaque notice ainsi que la structure physique de l’enregistrement (format et nombre d’occurrences des bannières).

Chaque groupe doit aussi définir pour chaque base les :

* Filtrages
* Synonymies
* Granularité

Et maitriser les fonctions de :

* Croisements
* Analyse de données
* Classification
* Dessin de graphe
* Géostratégie

### Administrer les données

Dans cette partie, un filtrage est appliqué sur la base de données afin d’obtenir un sous ensemble qui répond aux besoins des clients. Les sous ensembles constituent une vue restreinte de la base et sont utilisés comme point d’entrée pour chaque groupe afin que les étudiants puissent appliquer leurs requêtes.

### Exploiter les données

Cette partie consiste à générer des structures de données pour chaque groupe. Cela est possible en appliquant des requêtes sur les sous-ensembles de la base de données produits lors de l’étape précédente. Un programme en PERL fourni par l’enseignant produit des tables à partir de croisements obtenus par Tétralogie. Une étape intermédiaire permet de générer les instructions de création de la base de données et le peuplement de celle-ci. Ces étapes sont décrites dans la figure 1. Par la suite, les requêtes doivent être écrites. Ces requêtes concernent les données présentes dans les différentes tables obtenues. Suivant le type de requête, les sorties (résultats) sont classifiées en quatre (4) types (dimensions) :

* Une variable
* Un vecteur
* Une matrice
* Un cube

### Valoriser les données

Les fonctions de « reporting » sont essentielles pour réussir la présentation d’un travail de veille et pour convaincre les décideurs par un document lisible, pertinent et concis.

Pour chaque groupe, nous définissons différents types de sorties adaptés à chaque type de requête tels que :

* Histogrammes d’évolution 2D et 3D
* Histogramme comparatifs ou cumulatifs 2D et 3D
* Cartes géographiques
* Graphes relationnels.

Cet ensemble de possibilités doit permettre à chacun de trouver les bons réglages pour découvrir puis communiquer l’information stratégique ciblée à intégrer dans son rapport d’analyse personnalisé.

# Matériels et outils

## Plateforme de développement

Tétralogie

## Outils utilisés

Un outil est un logiciel qui supporte une méthode. Pour ce bureau d’études nous avons utilisé :

* GitHub
* Oracle
* Putty
* …
* Microsoft Word et Excel

WinHTTRACK :

## 

WinHTTrack est un aspirateur de sites web facile d'utilisation et [libre](http://www.gnu.org/philosophy/free-sw.fr.html) ([GPL](http://www.linux-france.org/article/these/gpl.html), logiciel libre).

Il permet de télécharger un site web d'Internet vers en construisant récursivement tous les répertoires, récupérant html, images et fichiers du serveur vers son ordinateur. HTTrack réorganise la structure des liens en relatif. HTTrack peut aussi mettre à jour un site existant, ou continuer un téléchargement interrompu. Le robot est entièrement configurable, avec un système d'aide intégré.

Source : <https://httrack.com/page/1/fr/index.html>

## Langages utilisés

* **PERL (Practical Extraction and Report Language)**

PERL est un langage de programmation permettant de traiter facilement de l’information de type textuel. Au cours de notre projet, nous l’avons utilisé par exemple pour regrouper nos trois sources de données en ligne en une seule et même base de données. Plusieurs autres programmes, codés en PERL, ont été réalisés, et nous permettant de gagner beaucoup de temps.

* **SQL (Structured Query Language)**

SQL est un langage informatique normalisé qui servant à exploiter des bases de données. La partie langage de manipulation de données de SQL permet de rechercher, d'ajouter, de modifier ou de supprimer des données dans les bases de données. Le langage de définition de données permet quant à lui de créer, et de modifier l'organisation des données dans la base de données. Pour exemple, nous avons utilisé ce langage au cours de notre projet lors de la création de la base de données.

# Gestion de configuration

De nos jours, de nombreux outils de gestion de configuration ainsi que de synchronisation sont disponibles. Les principaux outils utilisés sont TurtoiseSVN et GitHub. Ces deux logiciels ont la particularité de gérer les conflits lors de modification d’un même fichier par plusieurs utilisateurs. Ils permettent de synchroniser sur leur poste de travail personnel des dossiers et fichiers.

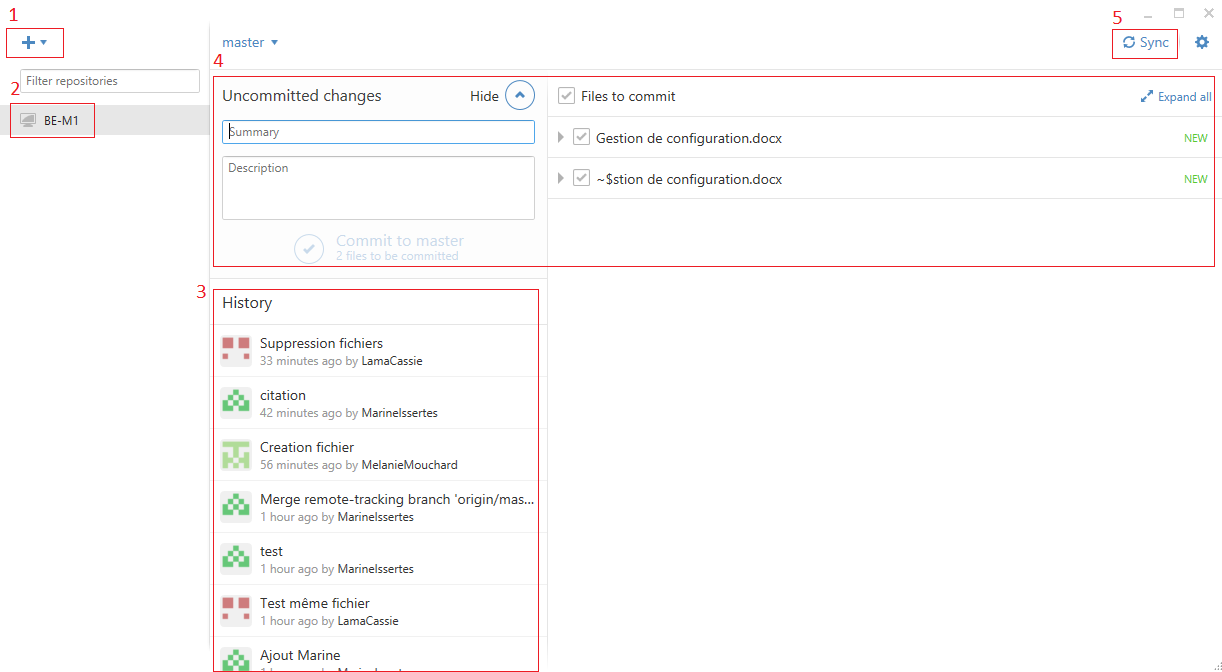
Lors du projet Vizurbi effectué début Janvier, l’ensemble des étudiants ont utilisé l’outil de synchronisation TurtoiseSVN qui récupérait des dossiers partagés sur la plateforme GitHub. Cependant cette plateforme propose un outil à télécharger sur le poste personnel qui gère les dossiers, tout comme TurtoiseSVN. L’avantage de GitHub est qu’il est plus simple d’utilisation grâce à son interface graphique fluide et très intuitive qui permet une adaptation rapide à cet outil. C’est pour cette raison que nous avons choisi d’utiliser GitHub pour ce projet d’étude car, d’une part, il est similaire à TurtoiseSVN et d’une autre part, nous voulions tester un autre outil de synchronisation et gestion de conflits pour nous en faire une idée personnelle.

Figure 1 Interface en local de GitHub

**1 :** Ajout d'un dossier de partage ou clonage d'un dossier existant sur GitHub.

**2 :** Répertoire de travail actif.

**3 :** Historique des révisions effectuées sur le répertoire de travail. Indication du titre, de la date ainsi que le modificateur du fichier (identifiant affiché).

**4 :** GitHub affiche tous les changements effectués sur le répertoire de travail par l'utilisateur seul. Nous pouvons indiquer un titre correspondant à la modification ainsi qu'une description qui seront visibles par les autres collaborateurs du répertoire. Les détails des changements sur les fichiers sont disponibles sur la partie droite de l'affichage.

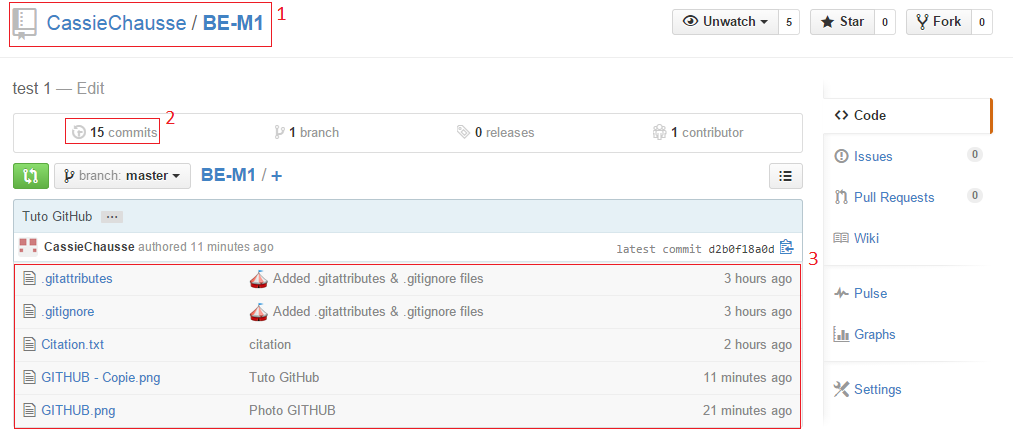
**5 :** Bouton de synchronisation des fichiers du répertoire. Dans ce cas, GitHub envoie aux collaborateurs les modifications effectuées en local et récupère les modifications des autres utilisateurs. De plus, cette partie affiche le nombre de fichiers modifiés par l'utilisateur (fichiers à synchroniser) ainsi que le nombre de fichiers que les autres utilisateurs ont modifié (fichiers à récupérer).

Figure 2 Interface sur le Web - Répertoire

**1 :** Nom du répertoire de travail

**2 :** Nombre de synchronisations effectuées par l'ensemble des collaborateurs

**3 :** Fichiers présents dans le répertoire

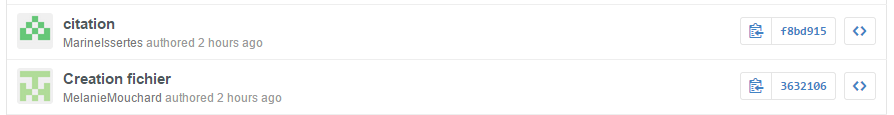


Figure 3 Interface sur le Web - Détails des sychronisations

La gestion des collaborateurs est possible sur GitHub. Le répertoire crée peut être "ouvert" c'est à dire que toute la communauté GitHub peut avoir accès au répertoire mais seuls les collaborateurs désignés peuvent modifier, supprimer ou ajouter des fichiers. Le répertoire peut être aussi configuré comme "privé", dans ce cas le répertoire est personnel, il n'est visible par aucun autre utilisateur et ne peut être partagé à d'autres.

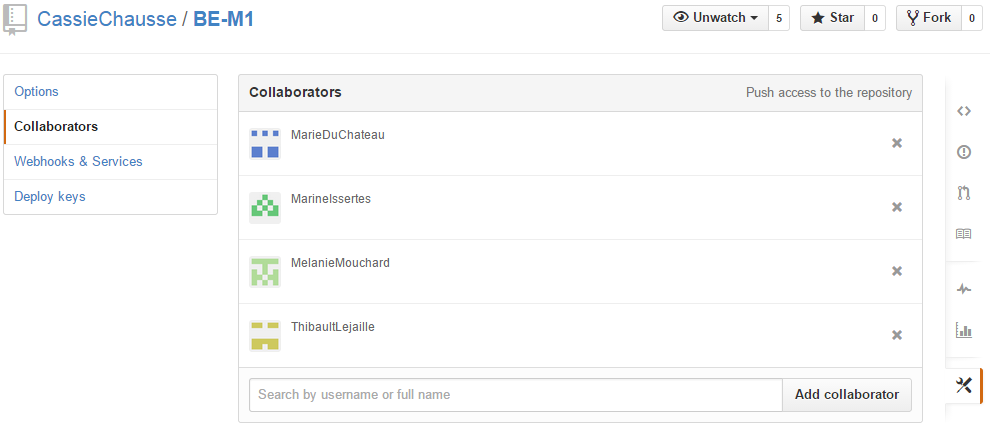
Dans notre cas le répertoire "BE-M1" est ouvert et partagé aux membres de l'équipe (Marine, Mélanie, Marie, Thibault et Cassie). Chaque répertoire nécessite une branche dite "master". En général, c'est l'utilisateur qui a créé le répertoire qui est considéré comme Master. Il a l'avantage de pouvoir gérer l'ajout ou la suppression de collaborateurs. De plus, c'est le seul à pouvoir gérer les conflits entre fichiers et versions des fichiers. C'est donc le responsable de configuration de notre groupe, Cassie, qui a ce statut.

Figure 4 Interface sur le Web - Gestion des collaborateurs

Lorsque plusieurs collaborateurs modifient un fichier et le synchronisent en même temps, Github détecte un conflit et affiche les différences entre les deux fichiers. Ce conflit est envoyé au Master qui devra choisir la bonne version du fichier.

L'outil GitHub nous a été d'une grande utilité tout le long du projet. Nous synchronisions principalement :

* Les différentes bases de données récupérées
* Les évolutions de programmes créés par notre responsable développement ainsi que les autres membres du groupe
* Les éléments à ajouter au rapport de projet (SADT, Diagramme de GANTT, PERT,...)
* Les analyses évolutives des données

En résumé, GitHub est un outil nécessaire et suffisant pour gérer toutes les étapes d'un projet.

# Assurances et contrôle qualité

## Revues

Formulaire de suivi pour chaque réunion

Revue :

* 29/01 : choix du sujet avec Mme Bahsoun
* 02/02 : définition du sujet, de la problématique et recherche des données sur les différentes bases de données en ligne avec Mme Bahsoun
* 23/02 : lecture du plan et du rapport avec Mme Bahsoun
* 09/03 : Point sur l’avancée du projet entre les membres de l’équipe. On a dit ce qu’on a fait et on s’est reparti le travail à faire pour la suite

## Contrôle qualité

Actions mises en place durant le projet

Faire les fiches de tests pour chaque fonction + Cahier de recettes

# Bilan

## Bilan par rapport au client

## Bilan pour l’équipe

# Sources

# Annexe

PubPsych

PubMed ???

Web of science ???

Cairn ??

Cahier des charges complet (feuille qu'elle nous a donné)

CODES REALISES !!!!!!!!!!!!!!!!!!!!!!!!