Développement d'un outil web d'extractions automatique de métadonnées des données produites dans l'unité

Sophia Solignac

INRAE

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

Rapport de stage 2022

IUT CLERMONT-FD
UNIVERSITÉ
Clermont Auvergne

IUT de Clermont-Ferrand
Département Informatique

Professeur tuteur : PROVOT Laurent

Maître de stage : TRAORE Amidou

Sophia SOLIGNAC

J’autorise la diffusion de mon rapport sur l’intranet de L’IUT

Remerciements

Je tiens à remercier M. Amidou Traore, mon maître de stage pour m'avoir accompagné lors de mon insertion, ainsi que pour ses conseils et l'expertise qu'il m'a apporté dans le domaine de la recherche et de l'imagerie RMN.

Je remercie aussi toute les membres de l'équipe AgroRésonance, plus particulièrement Jean-Marie Bonny, directeur de recherche, Guilhem Pagès, ingénieur de recherche, Cécile Keller, technicienne de recherche et Magali Nuixe, doctorante pour leur accueil chaleureux au sein de l'équipe.

Je remercie tout particulièrement Magalie Weber de l'unité BIA (INRAE, Nantes) qui porte le projet TransformON qui utilisera un des livrables de mon stage et surtout qui assure le financement de celui-ci.

Je tiens aussi à remercier Véronique SANTE-LHOUTELLIER, directrice de l'unité QuaPA pour son accueil et son accompagnement lors des démarches administratives, mais aussi et surtout pour m'avoir offert l'opportunité de ce stage.

Enfin, je remercie Laurent Provot mon enseignant référant, pour sa disponibilité tout au long de mon stage et le soutient qu'il m'a apporté.

Sommaire

Introduction

Pour ma deuxième année de DUT Informatique, j'ai eu à réaliser un stage d'une durée de 10 semaines, du 27 Juin au 2 Septembre 2022**. Mon organisme d'accueil était le centre Clermont-Auvergne Rhône-Alpes de l'Institut National de Recherche pour l'Agriculture, l'alimentation et l'Environnement situé à Theix.

L'INRAE est un centre de recherche publique dont la vocation est de produire et diffuser des connaissances scientifiques ; de mobiliser ces connaissances au service de l'innovation, de l'expertise et de l'appui aux politiques publiques. En tant qu'organisme de recherche publique, il est pleinement engagé dans le mouvement de la science ouverte qui vie à rendre disponible gratuitement les données et résultats de la recherche. Un des éléments de base dans la gestion de données et leur partage est le Plan de Gestion de Données (PGD), document formalisé qui décrit, via les **métadonnées*** l'ensemble du cycle de vie de la donnée. Les chercheurs ont recours à des Systèmes d'Information d'aide à la saisie pour générer les PGD. La plateforme AgroResonance a développé un SI qui répond à ses exigences (ISO 9001, autonomie, traçabilité).

Mon stage a pour objectif de développer des scripts d'extraction automatique des métadonnées à partir des fichiers descripteurs des données brutes générés par les spectromètres RMN. Ceci afin d'une part de faciliter l'utilisation du SI par l'auto-renseignement des formulaires du PGD et d'autre part structurer les métadonnées pour alimenter la construction d'une ontologie, vocabulaire permettant de décrire les produits et processus du domaine de recherche de la plateforme AgroResonance. En effet, un des livrables du projet sera utilisé dans le cadre du projet TransformON. L'objectif de ce dernier est de construire une une ontologie sur les procédés alimentaires et non alimentaires pour permettre une meilleure interopérabilité des données du département TRANSFORM

Je vais donc commencer par expliquer plus en détails le fonctionnement de l'INRAE, plus particulièrement la plateforme AgroResonance dont je faisais partie. Je décrirais ensuite la réalisation de ma mission de stage, et terminerais par vous donner le bilan technique de mes réalisations au sein de ce projet.

**Ma recherche ayant été compliqué dû aux conditions sanitaires actuelles, ma soutenance a lieu afin la fin de ce stage. Ce rapport est donc rédigé durant le stage.

Présentation de l'Institut

L'INRAE (l'Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement), fusion depuis 2020 entre l'INRA (Institut national de la recherche agronomique) fondée en 1946 et l'IRSTEA (Institut national de recherche en sciences et technologies pour l'environnement et l'agriculture) est un Institut de recherche publique sous la tutelle des ministères de la recherche et de l'agriculture.

Premier institut de recherche Agronomique en Europe et Deuxième dans le monde en termes de publications en sciences agricoles et en sciences de la plante et l'animal, l'INRAE mène des recherches dans l'objectif de développer des modes de production, animal ou végétal, ainsi qu'une alimentation plus saine et de qualité dans une idée d'agriculture durable et de préservation de l'environnement.

Ses objectifs pour l'environ 2030 sont les suivants :

I) <u>Objectifs scientifiques</u>

1. Répondre aux enjeux environnementaux et gérer les risques associés
2. Accélérer les transitions agroécologique et alimentaire, en tenant compte des enjeux économiques et sociaux
3. Une bioéconomie basée sur une utilisation sobre et circulaire des ressources
4. Favoriser une approche globale de la santé
5. Mobiliser la science des données et les technologies du numérique au service des transitions

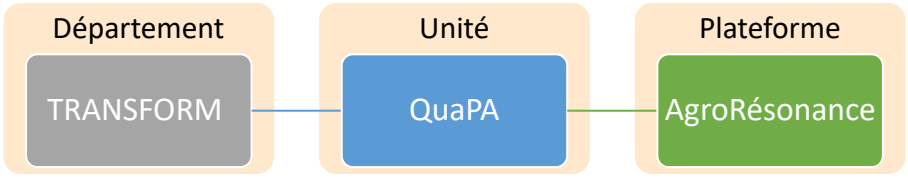
II) <u>Objectifs de politique générale</u>

1. Placer la science, l'innovation et l'expertise au cœur de nos relations avec la société pour renforcer notre culture de l'impact
2. Être un acteur engagé dans les sites universitaire en France et un leader dans les partenariats européens et internationaux

Etant une structure de très grande envergure, avec 17 centres de recherche en France métropolitaine et un en Guadeloupe, et disposant d'un effectif de plus de 8000 titulaires et 2800 contractuels, l'INRAE est décomposé en 14 départements de recherches, régis par 3 directions scientifiques.

Environnement

Pour mon stage, j'ai donc été assigné à la plateforme AgroResonance, de l'unité QuaPA appartenant au département TRANSFORM : Aliments, produits biosourcés et déchets. (Voir schéma ci-dessous, réduis au minimum pour des questions de lisibilité)



Cette unité comprend 36 permanents et accueille une vingtaine d'étudiants et CDD par an, dont environ une dizaine de doctorants et post-doctorants. Elle est structurée en trois équipes de recherche disciplinaires : Biochimie et Protéines du Muscle (BPM), Imagerie et Transferts (IT), Micro contaminants, Arômes et Sciences Séparatives (MASS) et deux plateformes : AgroResonance et Protéomique.

Les objectifs scientifiques de l'unité s'inscrivent dans les défis sociétaux, à savoir anticiper les adaptations nécessaires à la transition alimentaire mondiale, en particulier aux besoins en protéines et à un meilleur équilibre entre les différentes sources de protéines, à la durabilité des systèmes de production et des procédés de transformation et de conservation, et aux besoins de certaines populations ciblées (séniors, sportifs…). Pour répondre à ces objectifs, l'unité QuaPA concentre ses recherches selon deux axes :

(1) La conception raisonnée d'aliments carnés spécifiques. Ceci demande de tenir compte de leur acceptabilité et de leur devenir après ingestion selon des approches système modèle et directement chez l'Homme. La conception raisonnée de nouveaux aliments intègre le développement de procédés ad hoc avec un focus sur la structuration des assemblages de protéines de différentes sources et leur modélisation in silico.

(2) L'amélioration de la durabilité des systèmes alimentaires dans un contexte de transition. Cette amélioration est réalisée au travers des recherches de l'unité sur la sécurité chimique des produits, la limitation des pertes en produits d'origine animale au cours des procédés et la valorisation de la biomasse issue des coproduits animaux.

La responsabilité de la plateforme AgroResonance est assurée par Jean-Marie Bonny. L'équipe comprend actuellement 7 agents INRA, une ingénieure du CHU, une ingénieure de l'Université Clermont Auvergne, une post doctorante et cinq doctorants. Les domaines d'expertise de la plateforme recouvrent donc l'imagerie optique, l'imagerie et la médecine nucléaire, la radiologie et l'imagerie interventionnelle ainsi que l'imagerie par résonance magnétique

Existant

La plateforme AgroResonance met à disposition de la communauté scientifique les méthodes d'Imagerie et de Spectroscopie RMN pour répondre à des questions de recherche dans les domaines de l'agronomie, l'aliment et de la santé. Elle est équipée de trois imageurs complémentaires à très haut champ magnétique (4,7T, 9,4T et 11.7T), et développe une IRM portable (NMR MOUSE®) pour faire des mesures hors du laboratoire.

La plateforme est dans une démarche qualité active dans tous ses programmes de recherche pour laquelle elle est certifiée ISO 9001. Dans cette optique, elle doit veiller à la qualité de la chaine de mesures (ressources techniques) et la traçabilité des données. De plus, AgroResonance a également une démarche d'ouverture et de partage de données qui s'inscrit parfaitement dans la politique de Science ouverte encouragée par l'INRA.

Science Ouverte : La science ouverte consiste à rendre « accessible autant que possible et fermé autant que nécessaire » les résultats de la recherche, issus en majorité des fonds publics.

Très engagé dans le développement de la science ouverte depuis de nombreuses années, l'INRAE est le premier institut français à se doter (dès la fusion en 2020) d'une Direction pour la science ouverte (la DiPSO) dont l'objectif est de répondre aux enjeux liés à l'ouverture de la recherche dans un contexte de développement du numérique (loi pour une République numérique et Plan National pour la Science Ouverte) et d'attentes de plus en plus fortes de la société. Pour accompagner les chercheurs dans ce domaine, la DiPSO développe des outils pour la gestion des données, des publications, des vocabulaires…ainsi que des formations associées : on peut citer :

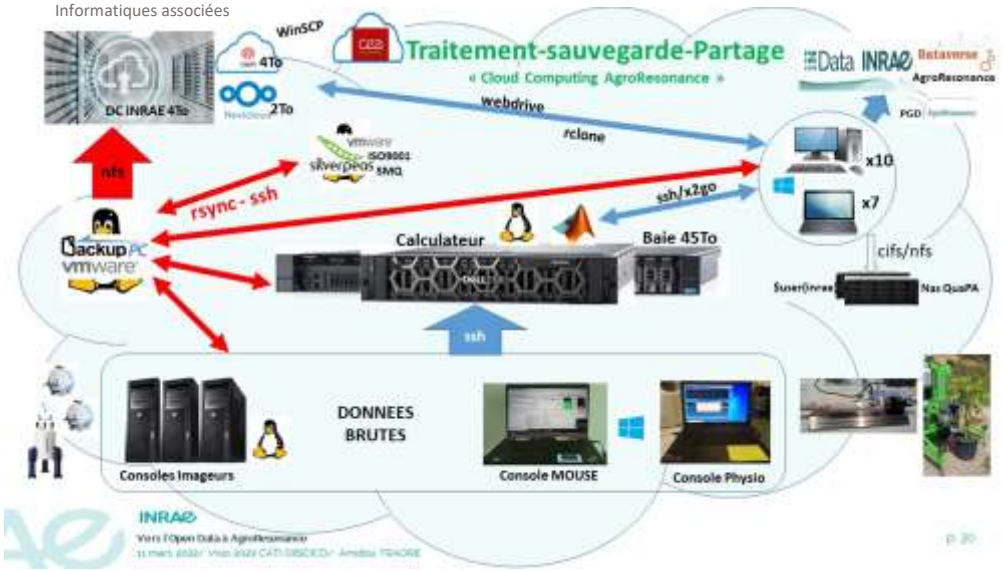
- HAL INRAE, l'archive ouverte d'INRAE : destinée au dépôt et à la consultation des travaux scientifiques d'INRAE. Elle contribue au développement du libre accès à l'information scientifique et technique, à l'ouverture de la science sur l'ensemble des thématiques de l'institut et à l'amélioration de la visibilité des recherches d'INRAE.
- Data INRAE, l'entrepôt des données d'INRAE devenu entrepôt national DATA.GOUV.FR : cet entrepôt de données offre de nouveaux services pour faciliter la gestion, le partage et la recherche des données de l'Institut.
- Datapartage, le site sur la gestion et le partage des données scientifiques : présente les services, outils et bonnes pratiques recommandées par INRAE.
- Thésaurus INRAE : référentiel thématique couvrant les domaines de recherche INRAE, est un outil pour faciliter l'accès aux objets numériques scientifiques et mettre en œuvre l'interopérabilité des systèmes d'information au sein de notre organisation.
- *L'attribution de* **DOI (Digital Object Identifier)** : identifiant unique et pérenne (URI) afin de facilité la gestion à long termes des ressources numériques (publications, données …) avec leurs métadonnées associées.

Figure 1 : Outils utilisés au cours du projet au sein de l'INRAE

On peut également signaler que la gestion des données selon les principes FAIR (Findable, Accessible, Interoperable and Reusable) est maintenant exigée par l'ensemble des financeurs publics de la recherche (ANR, Europe …). Ces organismes exigent à minima un ou plusieurs versions de PGD pour tous les projets qu'ils financent.

Gestion des données à AgroResonance : En tant que plateforme nationale labellisée et certifiée ISO 9001, AgroResonance a depuis de nombreuses années une politique de gestion et de sécurisation des données qui répond aux exigences de traçabilité des organismes de labellisation et de certification. La figure ci-dessous montre le flux des données au sein de la plateforme.

Figure 2 : Cheminement des données à l'INRAE et plateformes Informatiques associées



Conformément à sa démarche qualité, chaque projet de recherche conduit au sein de la plateforme est associé à une Fiche Projet qui contractualise la demande, les ressources humaines et matérielles allouées ainsi que les livrables. Une rubrique de la FP concerne également la gestion des données pendant et à l'issu du projet. Afin de se conformer aux exigences de la science ouverte, un PGD est maintenant également rédigé pour chaque projet, synthétisant ces informations dans le but de permettre leurs réutilisations dans le cadre de différents projets ou recherches futures, et toujours dans une optique de partage public de la Science

Pour ce faire, un SI d'aide à la saisie et de gestion centralisée des PGD a été développé par mon maitre de stage, A. Traoré chercheur de la plateforme, permettant de décrire l'ensemble des étapes du cycle de vie des données qu'elle génère afin de faciliter la gestion et le partage selon les principes FAIR.

Pour faciliter la rédaction de ces PGD, le SI développé permet une mise en page automatique de ces PGD après remplissage manuel de formulaires par les chercheurs, ainsi que le stockage des PGD et des informations des différents projets dans une base de données MySQL. Ce site utilise PHP pour traiter les formulaires ainsi que les insertions dans la base.

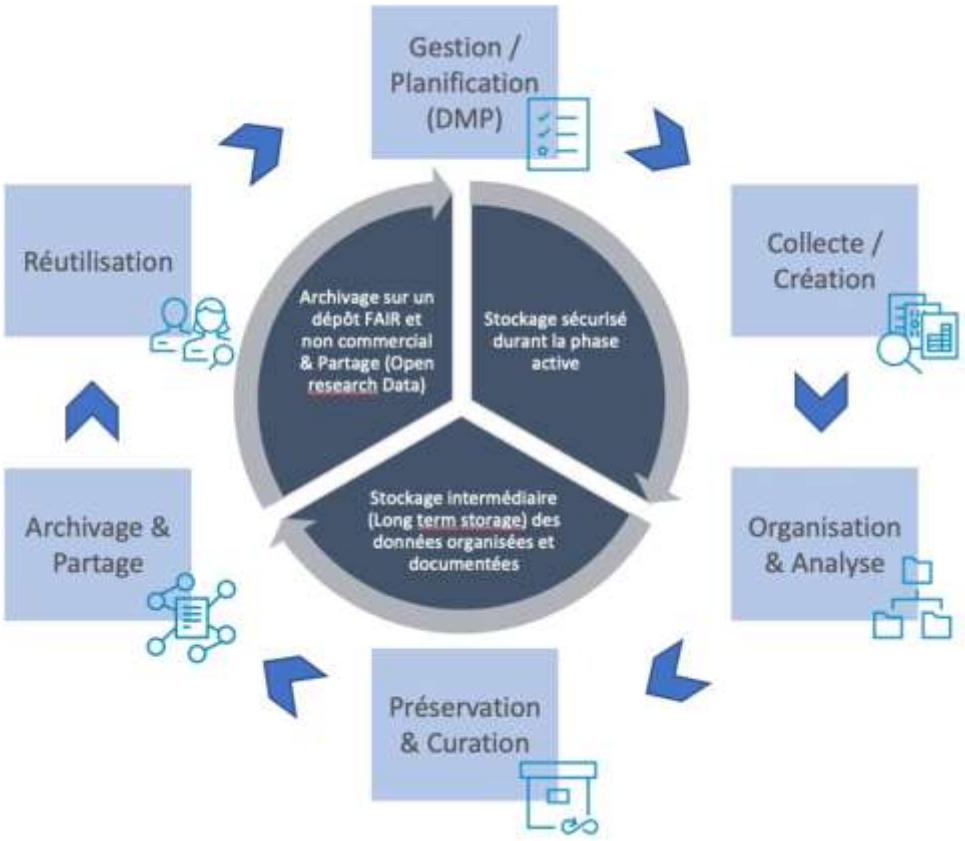


Figure 3 : Cycle de vie des données

Données produites : La plateforme est spécialisée en Imagerie et Spectroscopie RMN (Résonance Magnétique Nucléaire). Elle produit donc des images et/ou spectres RMN pour répondre à des questions scientifiques des chercheurs. Les données brutes générées par ces spectromètres-imageurs sont formatées avec une arborescence hiérarchique des fichiers descripteurs contenant des métadonnées (e.g., nom du sujet, appareil, lieu, paramètres d'acquisition, dates d'acquisition, format de la donnée…) nécessaires à leur exploitation (traitements, reconstruction, quantification, …).

Mis en forme : Police :Italique

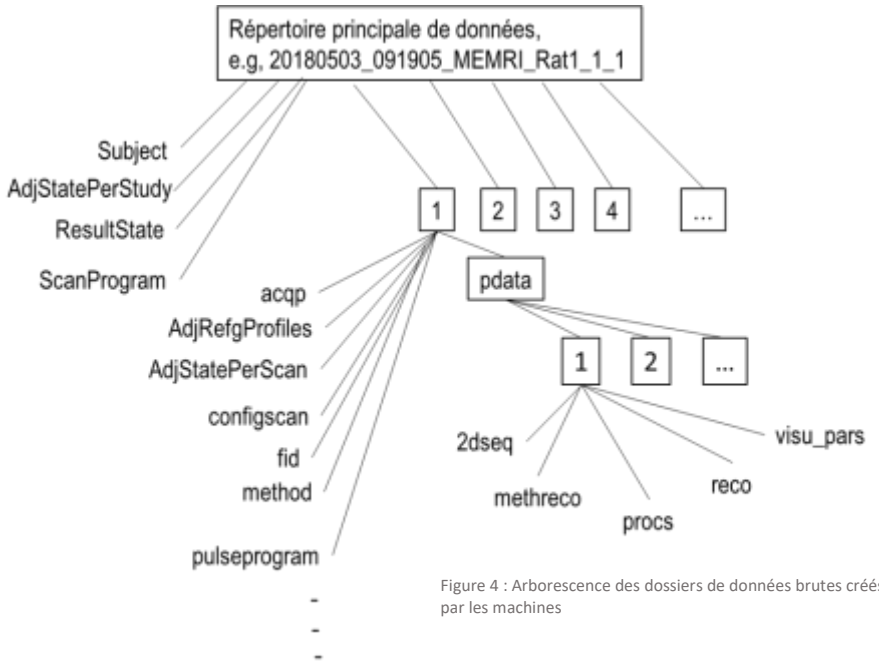


Figure 4 : Arborescence des dossiers de données brutes créés par les machines

Une grande partie de ces métadonnées sont identiques (ou permettent la déduction) des informations requises pour la documentation des données dans un PGD.

Objectifs

Ma mission principale était donc la suivante : Développer un module PHP attaché au site web existant afin de permettre l'extraction des données contenues dans les fichiers d'acquisition et de paramétrage machines dans le but de préremplir les formulaires de création de PGD notamment ceux relatifs à la production des données.

Cette tâche était ainsi décomposée en 3 parties : Extraction des fichiers de métadonnées brutes dans des formats exploitable (JSON, CSV…), Insertion de ces métadonnées dans une base de données structurée et enfin pré-remplissage des formulaires de rédaction de PGD.

- Extraction des fichiers de métadonnées

  Les fichiers de données étaient écrits dans un format textuel JCAMP-DX (http://www.jcamp-dx.org/) qui est un format standard développé au départ pour l'échange des spectres Infrarouge lisibles par la machine. Ce format est de nos jours étendu aux méthodes spectroscopiques et surtout à la RMN. Dans ces fichiers, il y a différentes balises de formatage (commentaire, paramètres, valeurs …). Il a donc fallu, depuis l'import de ces fichiers par l'utilisateur sur le site d'origine, extraire ces données afin de trier et normaliser leurs écritures, simplifiant ainsi leur lecture et leur utilisation.

- Insertion en Base

  Après avoir traité les données, il fallait choisir lesquelles étaient intéressantes, non seulement pour la rédaction des PGD, mais aussi pour les chercheurs de la plateforme qui auront accès à cette base. Il a donc fallu construire une base de données afin de stocker ces données de manière efficace.

- Pré-remplissage des PGD

  Enfin, il a fallu faire en sorte d'utiliser les données de cette base pour préremplir les PGD sur le site existant. J'ai donc dû gérer, non seulement l'extraction des données depuis la base, mais aussi faire en sorte que l'utilisateur ai une interface simple et complète lui permettant de choisir une rédaction manuelle ou semi-auto, ainsi que de pouvoir modifier à son choix les champs qu'il souhaite insérer dans son PGD. En effet, étant dans le domaine de la recherche, bien que la grande majorité des cas soient possiblement réalisable automatiquement, il y a de très nombreuse exception de procédures qui ne le peuvent.

Gestion de Projet

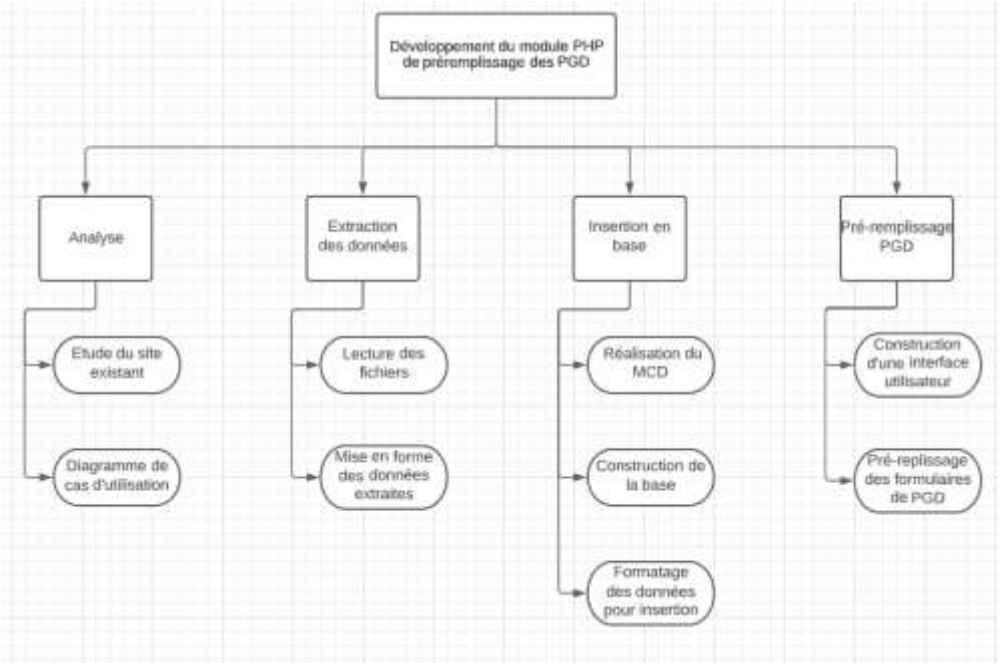
Dans un objectif d'organisation de mon travail, j'ai tout d'abord commencé par appréhender mon sujet grâce à la rédaction de documents de gestion de projet.

WBS :

L'un des premiers outils de gestion de projet que j'ai utilisé est le **Work Breakdown Structure**, plus communément appelé WBS. Ce diagramme permet de découper un projet en différentes tâches. Il est très utile au début d'un projet pour savoir comment organiser son travail et surtout dans quelle direction se diriger. Il permet de dégager les grandes étapes d'un projet.

J'ai donc utilisé ce diagramme pour lister les tâches qui m'étaient confiés afin d'avoir une vision claire de ce que j'avais à faire. J'ai découpé mon travail en 4 parties : L'analyse du site existant, l'extraction des données depuis les fichiers machines, la création de la base de données et l'insertion des données dans celle-ci, puis la mise en place d'une interface utilisateur préremplie avec les données en base pour la rédaction des PGD

Figure 5 : WBS du projet de mon stage



Afin de savoir comment étaient construits les formulaires de remplissage manuel des PGD déjà existant, ainsi que comprendre la structure du site, j'ai donc arpenté les lignes de code. Ce faisant, je me suis rendu compte que le site était fait d'un seul bloc, le coté serveur, client et base de données n'étant que très partiellement séparés, car ce n'était que des scripts s'appelant l'un l'autre. Cependant, le sujet de mon stage n'étant pas sur la refonte complète du site, mais bien l'implémentation d'une nouvelle fonctionnalité, et ayant seulement 10 semaines, j'ai donc dû m'adapter, en faisant en sorte de coller à la structure existante, mais en construisant des scripts spécialisés afin de bien faire cette différence.

Une fois l'analyse effectuée, et avoir pris connaissance de ma base de travail, j'ai commencé à me plonger dans les fichiers que j'allais devoir utiliser afin d'extraire les métadonnées utiles au remplissage d'un PGD. Cette phase était certainement avec la suivante la phase la plus critique de mon projet, car c'est celle-ci qui a déterminé la manière dont j'allais organiser mes données, qui seront utilisé par mon module pour le reste des étapes. Je devais ainsi faire en sorte d'avoir une base bien structurée et faire en sorte de récupérer un maximum de données dans les fichiers machines.

La base de données. C'est ici le cœur de tout le module. Après l'extraction des données, j'ai obtenu un jeu de métadonnée, bien qu'organisé, très large avec de nombreuses informations « doubles ». En effet les fichiers machines ont régulièrement des données équivalentes dans deux fichiers différents, ou bien encore des données qui correspondent à l'avant/après d'un calcul ou enchainement de calcul qui serait très simplement déductible (L'incrémentation successive d'une constante à une valeur, par exemple). De plus, étant dans le domaine très précis de la RMN, une grande partie des données peuvent être utiles aux chercheurs de ce domaines, mais en excédent pour un chercheur extérieur, qui n'aurait pas besoin de tous les détails pour réaliser ses propres recherches. Ainsi, j'ai avec l'aide de mon tuteur et de l'équipe, noté les données les plus importantes afin de structurer une base de données optimisée pouvant les recevoir. Enfin j'ai dû faire en sorte de reformater ces données pour qu'elles collent au format de la base de données (les tableaux en JSON par exemple) afin de les insérer.

Enfin je me suis lancée dans la création des formulaires préremplis des PGD. Afin de rendre plus simple mon travail, et surtout de permettre une expérience fluide et une interface simple à comprendre pour l'utilisateur j'ai remodelé ces formulaires, et ai ainsi aisément inséré les données de la base, prêtes à être envoyées dans un PGD.

MCD :

Pour faciliter la communication avec les chercheurs de l'équipe ainsi que de mettre en visuel l'aspect et l'organisation de la base de données que j'allais créer, j'ai construit un MCD de celle-ci avant toute choses.

Ce document a énormément évolué au cours de mon stage car c'était un aspect primaire de mon projet. En effet, c'est d'ici que viennent toutes les données utilisées pour le pré remplissage des PGD. Il m'a suivi tout au long de la conception du module, de l'extraction jusqu'au remplissage des formulaires.

Pour des raisons de simplicité, je ne vais expliquer que les attributs pouvant porter à confusion ou étant lié au domaine de la RMN

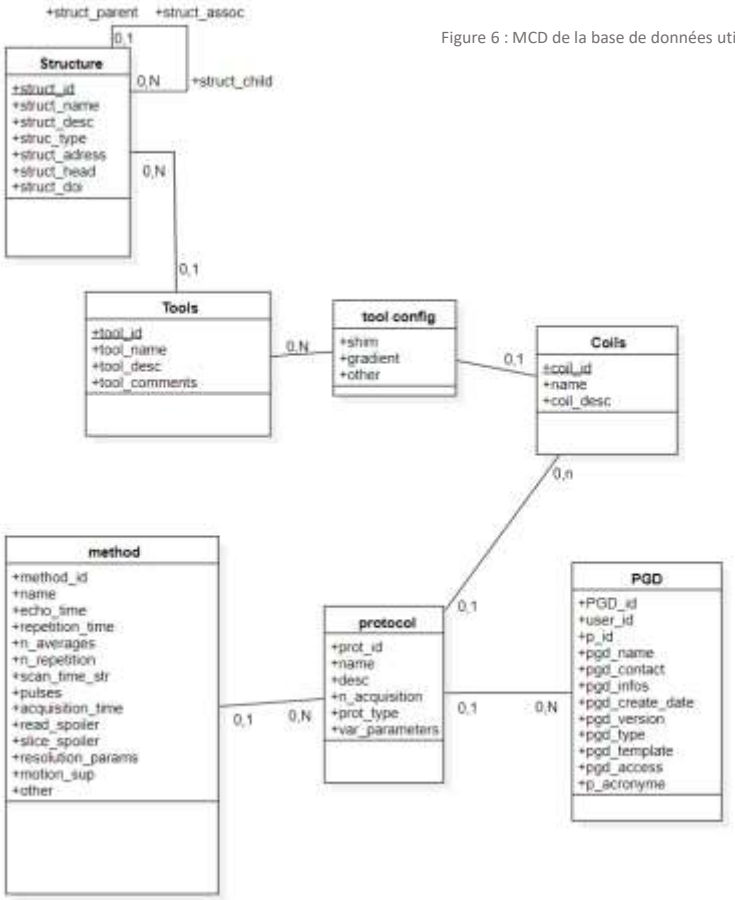


Figure 6 : MCD de la base de données utilisée

Structure : Les structure représente les différentes entités (équipe, unité, plateforme, infrastructure de recherche) au sein de l'INRAE. C'est un élément central dans le SI car elle peut déterminer le modèle (template) de PGD que l'on souhaite rédiger. Chaque structure peut en effet disposer d'un template qui lui sera spécifique, et qu'il faudra donc prendre en compte dans le traitement afin que le PGD résultant contienne le détail des informations conformes à ses besoins. Elle détermine aussi un certain niveau d'automatisation par le fait que les flux de données dans une structure ne varient que très peu d'un projet à l'autre (cf : Figure 2 Cloud Computing AgroResonance).

Struct_type : Le type de structure, représente le niveau dans la hiérarchie (unité, plateforme, départements…)

Struct_head : Le/La dirigent(e) de la structure.

Struct_doi : Digital Object Identifier. C'est l'identifiant unique permettant de retrouver la structure sur le web.

Struct_assoc : Table d'association permettant de hiérarchiser les structures (unité appartenant à un département par exemple).

Tools : Machines utilisée dans l'acquisition des données. Elles se situent toujours dans une structure donnée.

Tool_name : Nom UNIQUE de la machine, donnée en sortie d'usine comportant la puissance de son champ magnétique.

Coils : Antenne d'acquisition utilisés en parallèles d'une machine. Elles sont créées pour un type de machine en particulier.

Coil_config : Table de liaison entre les antennes et leurs machines comportant le paramétrage à entrer dans la machine pour faire fonctionner cette antenne.

Toutes les données dans cette table sont des paramètres de réglages entre la machine et l'antenne.

Protocol :

N_acquisition : Nombre d'acquisition ou d'images effectuées.

Prot_type : Type du protocole. Il nous permet de savoir quelle sera la donnée variable durant les différentes acquisitions de ce protocole (e. g Un protocole MSME_T2_map fera plusieurs acquisitions de la méthode MSME en modifiant le temps d'écho)

Var_parameters : liste des différentes valeurs prisent par le ou les paramètres variables (Ex. Temps d'Echo ci-dessus.)

Method : La méthode d'acquisition utilisée pour un protocole (Ex : MSME). C'est cette méthode qui définit le type d'ondes utilisés pour l'acquisition, ainsi que tous leurs paramètres,

Echo_time : le temps d'écho de cette méthode. Parfois égal au temps d'échantillonnage du signal

Repetition_time : Temps d'attente pour le retour à l'équilibre avant de répéter une acquistion

N_averages : Nombre de répétitions pour augmenter le rapport signal sur bruit

Scan_time_str : Durée Totale de l'acquisition de cette méthode.

Pulses : Présente toutes les ondes Radio Fréquence utilisés avec tous leurs paramètres organisés dans un fichier JSON

Read_spoiler : dispersion des signaux parasites avant lecture du signal d'intérêt

Slice_spoiler : dispersion des signaux parasites avant l'encodage de la coupe

Resolution_params : Paramètres liés à la résolution finale de l'image acquise (Champ de vision initial, Matrice appliqué à l'image initiale et épaisseur d'une couche de l'image)

Motion_sup : suppression des mouvements de l'échantillon

Other : Autres paramètres liés à l'acquisition, spécifiques à certains cas d'utilisation de la méthode

PGD : C'est ici que sont stockés les informations ACTUELLES liées aux PGD. Les fichiers PGD sont stockés à part en gardant trace de chacune des versions.

User_id : Utilsateur ayant créé le PGD

P_id : Projet auquel appartient ce PGD

Pgd_contact : adresse électronique de la personne en charge de ce PGD

Pgd_infos : Informations quant à la création de ce PGD (Objectifs, description générale du sujet)

Pgd_version : Version du PGD. Peut être initiale, intermédiaire (en cours de projet) ou finale.

Pgd_type ; Type du PGD. Peut être soit électronique, soit un fichier (Word ou PDF).

Pgd_template : Le template à utiliser pour ce PGD. Est lié à une structure, car chaque structure peut disposer de son template

Pgd_acces : droits de lecture, écriture du PGD

P_acronym : l'acronyme du projet associé à ce PGD.

Réalisation

Après avoir étudié et assimilé la structure de mon outil de travail, puis rédigé les documents nécessaires au bon déroulement de mon projet, j'ai donc commencé par l'étude des fichiers de métadonnées machines. Parmi ces fichiers j'ai pu relever un format très grandement majoritaire ; les fichiers JCAMP-DX (figure 7).

Figure 7 : Fichier de donnée au format JCAMP-DX

Ces fichiers ont une structure relativement simple : chaque donnée débute par un « tag » permettant d'identifier sa nature. Voici les différents types :

« ## » pour les données relatives au fichier (nom, version de JCAMP-DX, utilisateur l'ayant créé…)

« **##$** » pour les données relatives à l'acquisition. Ce sont ces données qui contiennent ce qui nous intéresse, comme la machine utilisée, la puissance des impulsions, le temps de repos…

« **$$** » pour les données relatives aux variables utilisées dans les x dernières secondes lors de l'acquisition. Elles ne nous sont pas utiles car elles ne font que de répéter en synthétique ce qui était écrit dans les données précédentes.

J'ai donc utilisé ces tags pour reconnaitre mon type de donnée, et les traiter adéquatement. Les données de fichiers étaient relativement simples : elles sont de la forme nom = valeur. Les données « variables utilisées » ne m'étaient pas utiles, j'ai donc simplement ignoré.

Cependant, les données d'acquisition étaient quant à elles, bien plus complexes. Elles peuvent en effet varier du simple nom = valeur, à des formes très complexes, illustré ci-dessous (figure 8).

```
##$ACQ_RfShapes=( 64 )
(<$ExcPulse1Shape>, 66.2678752747411, 0, 0.5, 0, 100 100 100 100 100 100 100
100, 0 0 0 0 0 0 0 0) (<$RefPulse1Shape>, 340.450089390245, 0, 0, 0, 100 100
100 100 100 100 100 100, 0 0 0 0 0 0 0 0) (<>, 0, 0, 0.5, 0, 0 0 0 0 0 0 0 0,
0 0 0 0 0 0 0 0) (<>, 0, 0, 0.5, 0, 0 0 0 0 0 0 0 0, 0 0 0 0 0 0 0 0) (<>, 0,
0, 0.5, 0, 0 0 0 0 0 0 0 0, 0 0 0 0 0 0 0 0) (<>, 0, 0, 0.5, 0, 0 0 0 0 0 0 0
0, 0 0 0 0 0 0 0 0) (<>, 0, 0, 0.5, 0, 0 0 0 0 0 0 0 0, 0 0 0 0 0 0 0 0) (<>,
0, 0, 0.5, 0, 0 0 0 0 0 0 0 0, 0 0 0 0 0 0 0 0) (<>, 0, 0, 0.5, 0, 0 0 0 0 0
0 0 0, 0 0 0 0 0 0 0 0) (<>, 0, 0, 0.5, 0, 0 0 0 0 0 0 0 0, 0 0 0 0 0 0 0 0)
(<>, 0, 0, 0.5, 0, 0 0 0 0 0 0 0 0, 0 0 0 0 0 0 0 0) (<>, 0, 0, 0.5, 0, 0 0 0
0 0 0 0 0, 0 0 0 0 0 0 0 0) (<>, 0, 0, 0.5, 0, 0 0 0 0 0 0 0 0, 0 0 0 0 0 0 0
0) (<>, 0, 0, 0.5, 0, 0 0 0 0 0 0 0 0, 0 0 0 0 0 0 0 0) (<>, 0, 0, 0.5, 0, 0
```

Figure 8 : Donnée complexe

Elles présentaient cependant un paterne constant. J'ai donc choisi d'utiliser des expressions régulières afin d'identifier ces patterns et les extraire efficacement sans avoir à manuellement gérer chacune des variables. J'ai ici utilisé un débugger d'expression régulière (https://regex101.com) pour parvenir à mes fins, car elles sont très vite devenues longues et complexes, à la vue du format des données (figure 9).

Code de champ modifié

Figure 9 : Exemple d'expression régulière complexe et son utilisation

Après avoir réussi à traiter une grande majorité des données, je me suis aperçu que certains paternes qui semblait être les mêmes différaient très légèrement d'un fichier à l'autre. Après beaucoup de réflexion sur la manière la plus simple de gérer ce problème, j'en suis arrivée à la conclusion qu'il me fallait des scripts d'extraction différents selon les fichiers lus. J'ai donc créé différents fichiers d'extractions, définissant chacun une classe pour me permettre de les

faire hériter de mon extraction classique, et définir ainsi des algorithmes adaptés à chaque pat. J'ai ainsi utilisé de l'introspection, grâce aux noms du fichier en cours d'extraction afin de lui associer sa classe d'extraction à l'exécution.

Une fois ces problèmes réglés, je disposais d'un grand tableau contenant séparément les données de chaque fichier. Afin de m'assurer de son intégrité afin d'attaquer la base, j'affichais donc ce tableau sur le site (figure 10).

Figure 10 : Affichage des données après extraction et réorganisation

L'extraction des données terminée en grande partie, je me suis attaquée à la création de ma base de données MySQL, ainsi qu'à repérer avec les chercheurs de l'équipe les valeurs intéressantes, et les liens entre celles-ci. J'ai donc réalisé un premier MCD pour permettre une discussion plus claire avec eux.

Cependant, après les premiers tests d'insertion de données, je me suis rendue compte d'un problème majeur dans la conception de ma base. En effet, il y avait dans cette base d'autres données liées au site, majoritairement celles des projets, des utilisateurs, mais aussi des PGD qui y avaient été enregistrés. Mes données étaient bien reliées comme je le souhaitais, mais rien ne nous permettait de les relier à un projet ou PGD. A ce moment-là du développement, je n'avais pas encore touché au formulaire des PGD, et je ne pouvais de ce fait, pas encore utiliser ces données pour faire le lien entre des métadonnées, et un projet.

J'ai revu ma base afin d'accueillir cette liaison. Ce faisant, nous sommes repassés sur la structure des protocoles et méthodes d'acquisition, car c'est le point central de ces données. J'ai par exemple fait le choix de simplifier certaines valeurs « multiples » dans une seule colonne des tables, en utilisant un format JSON accepté par MySQL, car ce sont des données qui n'ont de sens que dans un tout. J'ai rassemblé, par exemple, les données liées aux différentes impulsions utilisés dans une méthode donnée. Cela m'a donc demandé de traiter indépendamment ces valeurs avant insertion afin de les formater correctement en un ensemble logique (figure 11).

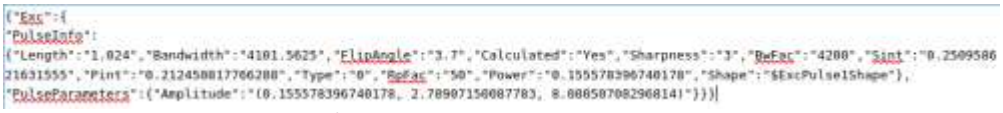


Figure 11 : Fichier JSON comportant les données des impulsions

Après tout ceci, et m'être rendu compte de la complexité des données que j'ai à traiter, j'ai fait en sorte d'être en dialogue constant avec mon tuteur sur les données en base, pour m'assurer d'obtenir une base solide avec ni trop peu, ni trop d'informations.

J'ai à ce moment commencé à travailler en parallèle sur l'interface utilisateur en créant un formulaire d'envoie des fichiers (figure 12). Je me suis cependant rendue compte d'un problème simple lié à l'input de fichier des formulaires en HTML : lorsque l'utilisateur souhaite importer plusieurs fichiers, il ne peut le faire que depuis le même dossier. S'il quitte ce dernier, sa sélection n'est pas prise en compte, et s'il reclique après avoir déjà inséré des fichiers, les précédents sont écrasés. Or dans l'arborescence de fichiers machine à extraire, les fichiers sont organisés dans différents dossiers. De plus, côtoyant les métadonnées se trouve le résultat de l'acquisition. Une ou plusieurs images contenues dans un fichier binaire, qui peut de ce fait être très lourd, mais qui nous est totalement inutile dans la rédaction de PGD. Les images sont stockées sur un serveur à part, qui n'est pas géré directement ici.

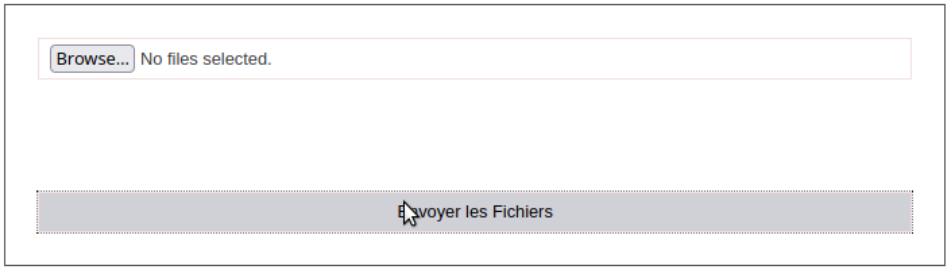


Figure 12 : Formulaire d'envoi de fichiers

J'ai donc attaché à mon input une fonctionnalité, à chaque fois qu'un ou plusieurs nouveaux fichiers sont insérés. J'ai utilisé JavaScript pour cela, car cela m'a permis de gérer ces problèmes avant même l'envoi des fichiers au serveurs, gagnant ainsi en performance. Mon input a donc les fonctionnalités suivantes.

Tout d'abord, à chaque insertion de nouveaux fichiers, les tailles, noms ainsi que le nombre de fichiers total sont vérifié. On évite ainsi des fichiers trop volumineux, un trop grand nombre de fichiers importés d'un coup, la présence de ce fichier dans ceux déjà choisi, ou bien un fichier « interdit » (Comme les images acquises). Etant des données liées à un protocole particulier, il n'y a jamais la nécessité d'envoyer un grand nombre de fichiers. Plusieurs extractions peuvent être effectués les unes à la suite des autres. Dans le cas ou ceci n'est pas respecté, un message d'erreur est affiché (figure 13 et 14).

Figure 13 : Erreur nombre de fichiers

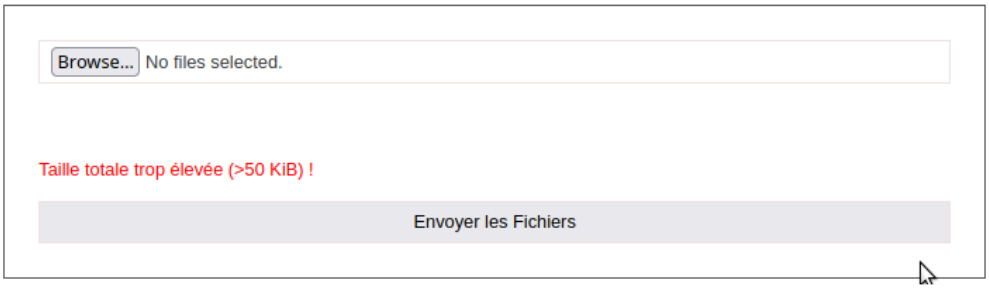


Figure 14 : Erreur taille totale des fichiers

Après l'insertion des fichiers, un nouvel input est créé afin de permettre à l'utilisateur de choisir de nouveaux fichiers, appartenant à un autre dossier (figure 15). Il est soumis aux mêmes conditions que les autres, car elles sont vérifiées sur l'ensemble des input présents.

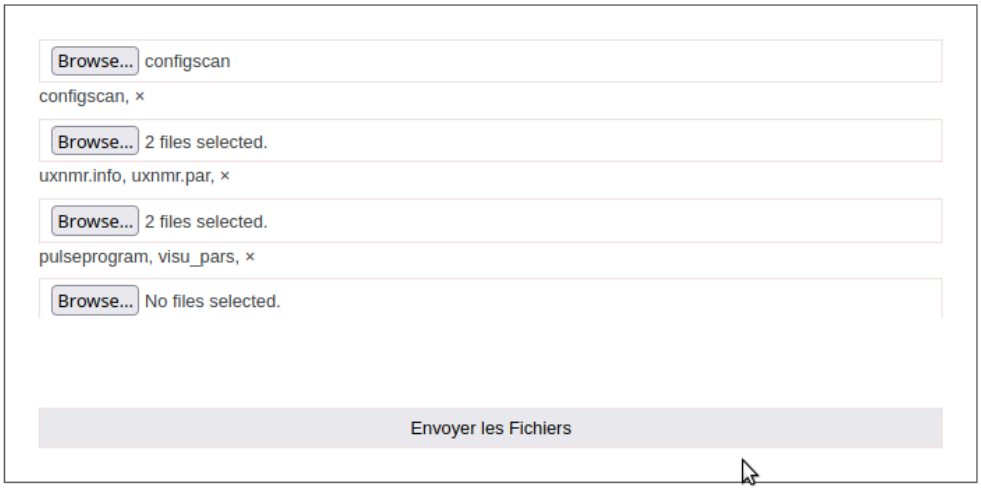


Figure 15 : Exemple d'un envoi de fichiers appartenant à différents dossiers

Il peut évidemment s'il le souhaite supprimer des fichiers grâce à la croix situé à côté de l'input en question. Il peut ensuite les envoyer au traitement, où la classe adaptée se chargera de son extraction.

Après avoir bien organisé la base de données, et géré cette première interface utilisateur, j'ai commencé à implémenter mon travail au sein des formulaires déjà existants de création de PGD. La page est en réalité un grand formulaire généré par un script qui s'occupe de la mise en page et du nettoyage des données du formulaire pour des raison de sécurité. Les données de base du formulaire lui sont données, et il renvoie puis affiche le formulaire construit.

Pour bien expliquer ou se situe mon travail sur ce site, je vais donc décrire l'utilisation attendu du site par un chercheur. Il commence tout d'abord par la page « Projet ». Il y rentre les informations relatives au projet pour lequel il a fait ces acquisitions (figure 15).

Figure 15 : Formulaire de création d'un projet

Une fois ces infos remplies, ou si le projet existait déjà, il va donc créer son PGD. Cela va insérer dans la base les informations principales de ce PGD, ainsi que le projet auquel appartient ce PGD (figure 16). Il devra aussi choisir un modèle pour le PGD. Ces formats appartiennent toujours à une structure, et les informations relative à celle-ci s'affiche afin de lui permettre de les vérifier et de les modifier à sa guise. S'il les change, elles ne sont pas modifiées dans la base, mais uniquement pour ce PGD.

Figure 16 : Formulaire de création d'un PGD

Enfin, il va passer sur la description des données. C'est ici qu'il doit avoir le choix entre le remplissage manuel ou semi-automatique du PGD. Il devra donc choisir quel PGD remplir, puis, le mode de remplissage. S'il choisit le remplissage manuel, des champs vierges apparaissent, lui laissant remplir toutes les données qu'il souhaite (figure 17).

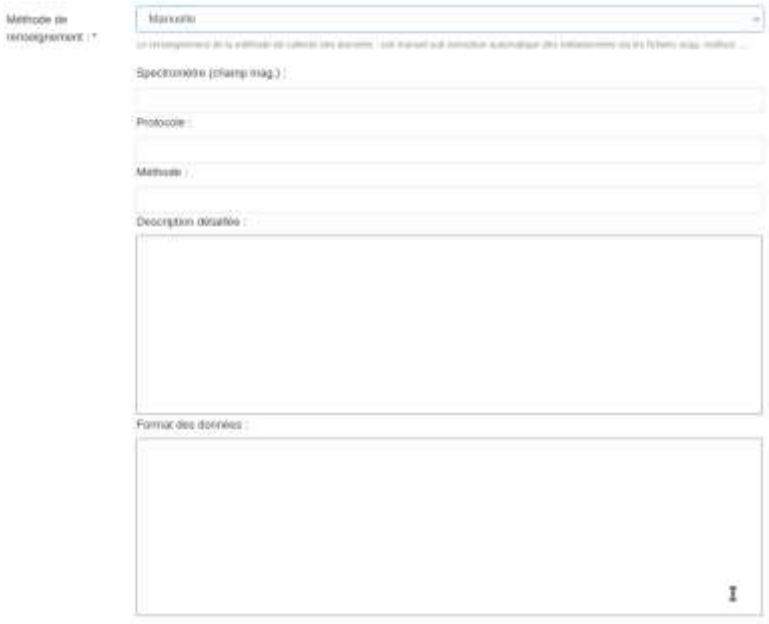


Figure 17 : Formulaire de remplissage manuel du PGD

S'il choisit le remplissage automatique, il tombera d'abord sur le formulaire d'extraction de fichiers décrit précédemment. Les données seront donc extraites des fichiers insérés dans la base puis, les mêmes champs que pour l'insertion manuelle s'afficheront, préremplis lorsque possible avec les données extraites des fichiers (figure 18), puis envoyer le formulaire afin de remplir son PGD dans la base.

Figure 18 : Formulaire de remplissage semi-automatique du PGD

Bilan Technique

Au cours de mon stage j'ai ainsi réalisé la tâche principale qui m'as été confiée. Je suis parvenue à extraire les métadonnées des fichiers machines et les organiser de manière à les insérer dans la base de (méta)données préalablement construite. Ces métadonnées peuvent ensuite servir, selon le choix de l'utilisateur, au remplissage automatique des formulaires du PGD. De même, ces données peuvent être exportées au format CSV ou JSON pour la construction de vocabulaire. J'ai aussi fait en sorte d'agrémenter mon module d'une interface simple et intuitive, permettant d'améliorer l'expérience d'un utilisateur l'utilisant.

Cependant, beaucoup de cas spécifiques sont encore manquants, comme la gestion de protocoles particuliers, qui ne répondent pas aux mêmes formats que ceux que j'ai pris pour exemple, ou encore la gestion des fichiers d'autres logiciel ou format. En revanche, j'ai construit ce module comme un échafaudage pouvant être déplacé ou adapté à sa guise. Les différents formats de fichiers pourront être traités simplement en créant des classes d'extraction fille de celle de base, et en modifiant ce qui sera nécessaire.

La durée (moins de 10 semaines) de ce stage n'était pas suffisante pour prendre en compte la totalité des cas d'utilisation possible de mon module. La complexité des informations à traiter et les relations entres elles m'ont demandé beaucoup plus de temps que je ne l'avais prévu, mais j'ai dû apprendre à surmonter ces difficultés en communiquant un maximum avec l'équipe. Cela m'a apporté des connaissances que je n'avais pas, notamment en RMN, et je trouve très intéressant d'avoir pu ainsi explorer d'autres domaines lors de mon stage.

Conclusion

Ce stage m'a apporté ma première expérience professionnelle dans le domaine informatique. J'en suis très satisfaite, car ce fut une expérience des plus enrichissante. J'ai pu découvrir réellement de métier de développeur ainsi que me frotter directement a des problèmes communs du développeur comme la communication avec le « client », ou collègues dans mon cas, venant d'un milieu différent.

En effet, c'est sur cet aspect que je me suis le plus développée, car il m'a fallu non seulement simplifier mon vocabulaire et ma logique pour permettre à mes collègues de comprendre quelles étaient mes pensées, mais aussi car j'ai fait partie d'un domaine relativement complexe, ayant une manière de penser bien différente de la mienne, et comprenant de nombreux termes techniques qu'il m'a fallu assimiler.

Ce stage m'as aussi permis d'approfondir les connaissances acquises en classe, tout en ayant une approche bien différente de ces dernières. En effet le milieu scolaire dispose d'une approche bien plus linéaire que le professionnel, ou l'on doit sans cesse se poser les bonnes questions et rechercher toujours plus loin afin d'être efficace dans son travail.

Enfin, ce qui m'as le plus satisfait durant ce stage était que j'ai eu à utiliser des langages et des technologies qui m'étaient inconnus ou que nous n'avions pas ou très peu étudié en classe, tel que le JavaScript. J'ai pu confirmer mon autonomie dans la recherche des informations nécessaires tout en apprenant quelque chose qui me sera très certainement utile dans le futur.

Résumé en anglais

As part of my school degree in Computer Science at the University Institute of Technology UCA in Clermont-Ferrand, I had to do a 10-week internship in the public research institution of INRAE located at Theix, Auvergne.

I had to develop alongside the researchers a PHP module designed to extract and structure metadata from machine files in order to pre-fill a document called a DMP (Data Management Plan). It is a quite long document designed to describe the data's life cycle and method of acquisition, and it can be quite hard and long to fill-in manually, so pre-filling it with data that can be extracted from the acquisition files makes it way more manageable.

This module will be implemented on a website designed by my supervisor to help researchers with the filling of DMPs.

My objectives were as follows: Extract the data from machine files and structure it, because it is not always formatted well, and those files are not easily human-readable. Then I had to create a database to store meta data of interest in a structured and optimized way, along with sorting useful from useless data. Finally, I had to use this data to fill a form on the website that can be modified as the user wishes.

I started by reading the existing code of the website in order to understand its core structure, as I was going to work on top of it. When this was done and I had a great understanding of it all, I started working on the extraction of data from machine files. My first issue was the discrepancies between the different metadata files. They all used the same base format (i.e, JCAMP-DX), with each data starting by a "tag" describing what kind of data was following but had different organizations for its data afterward. I solved the problem by having different classes all based on the same algorithm but presenting differences for each file that could be treated.

Afterwards, I began to work on the database welcoming all this extracted data. First of, I talked with the researchers to learn which metadata is more important and deserved to be stored. Once that was done, I created the database and started formatting my previously extracted metadata then inserting it in the base, ready to be used. Finally, I simply used it to fill out the DMP forms on the website, improving the user interfaces along the way.

Overall, my internship was a great and rewarding experience as I got to learn a lot of how a professional and specifically a public research institute works from the inside, as well as the importance of communication when need be.

Even thought there is still a lot to do with my module, I worked aware of the future and tried to build a module as improvable as I could, and it did allow me to discover the profession of developer in a non-scholar environment.

Webographie

https://www.php.net/docs.php

https://regex101.com

https://devdocs.io/javascript/

https://www.w3schools.com

Code de champ modifié

Lexique

PGD : Plan de gestion des données. C'est un document normé décrivant le cycle de vie des données utilisé dans un but de partage efficace et publique des recherches.

PHP : HyperText Preprocessor. C'est un langage web serveur utilisé pour traiter les requêtes envoyées par les clients.

JavaScript : C'est un langage web client utilisé pour dynamiser et embellir les pages web, ainsi que d'effectuer des actions possibles directement sur le navigateur client.

MCD : Modèle Conceptuel de Données. C'est un document permettant de visualiser l'organisation d'une base de données.

Science Ouverte : La science ouverte consiste à rendre « accessible autant que possible et fermé autant que nécessaire » les résultats de la recherche, issus en majorité des fonds publics.

Ontologie : c'est un modèle de donnée ou une forme de représentation des concepts clés d'un domaine et les relations entre eux. Une ontologie (En informatique et en science de l'information) est l'ensemble structuré des termes et concepts représentant le sens (sémantique) d'un champ d'informations, que ce soit par les métadonnées d'un espace de noms, ou les éléments d'un domaine de connaissances.

Métadonnée : Donnée servant à la description ou la définition d'une autre donnée

RMN : Résonance Magnétique Nucléaire. C'est une technique qui exploite les propriétés magnétiques de certains noyaux atomiques.