INGÉNIEUR EN MATHÉMATIQUES APPLIQUÉES ET MODÉLISATION SPÉCIALITÉ GESTION MULTIMÉDIAS DES DONNÉES MASSIVES

RAPPORT DE PREMIER SEMESTRE D'APPRENTISSAGE

SUJET
Valorisation des méthodes de Data Science pour la Caisse Nationale des Allocations Familiales (CNAF)

Apprenti: Nicolas TURPAULT (MAM5 - GMD)

Maître d'apprentissage : Clémentine NEMO

Tuteur Enseignant : Lionel FILLATRE

Entreprise : ATOS

19/09/2016 - 20/02/2017

Résumé

Lors de ce rapport, on discutera des idées ainsi que du travail mis en place pour valoriser l'open data de la CNAF. Ceci passera tout d'abord par un travail de visualisation des données de la CNAF, suivi des présentations de plusieurs idées devant différentes parties qui se conclura par le développement du projet de valorisation de l'Open data grâce au web sémantique.

Remerciements

Premièrement, j'aimerais remercier l'entreprise ATOS de Sophia Antipolis de m'avoir accueilli lors de ces cinq premiers mois d'apprentissage. Merci à tous les collaborateurs qui ont participé au bon déroulement de cet apprentissage. Tout le monde a été très accueillant, prêt à aider et cela m'a permis de découvrir de façon très positive la vie au sein d'une multinationale comme Atos.

Un remerciement particulier à tous les membres de l'équipe travaillant sur le portail partenaire de la CNAF qui m'ont intégré de façon très rapide au sein de leur équipe.

Je tiens aussi à souligner la disponibilité et l'écoute de mes deux encadrants d'apprentissage qui sont Kévin ROYERE et Clémentine NEMO. Ils m'ont été d'une aide précieuse au bon déroulement de cet apprentissage par leur suivi et leurs critiques positives et négatives à l'égard de mon travail et de mes idées. Ils m'ont aussi permis de découvrir de nouveaux domaines et d'acquérir de nouvelles compétences très diverses comme des compétences fonctionnelles, l'aisance à l'oral, l'adaptation à l'auditoire, la place à prendre au sein des projets, etc.

Un remerciement particulier à Alain CIVEL, directeur technique du site, qui m'a donné quelques travaux d'analyses de données simples ainsi que la mise en place de réunion avec le client afin de présenter mes compétences et travaux. Il a été un moteur durant ces 5 derniers mois par ses contacts et ses connaissances qui ont permis un regard particulier sur mes travaux.

Finalement, je remercie Gabriel KEPEKLIAN (Directeur en Recherche et Développement sur le site d'Atos à Bezons) et Olivier CORBY (enseignant chercheur à l'INRIA) sans qui ce projet n'aurait pas été possible. Ils m'ont permis la compréhension et l'acquisition de nombreuses compétences en web sémantique ainsi que l'accès à leurs outils respectifs (Datalift et Corese) accompagnés de conseils très pertinents.

Plan du rapport

Introduction

Atos Sophia est au cœur d'une multinationale, elle offre des services informatiques aux entreprises.
Dans le cadre d'un contrat avec la CNAF, ATOS prend en charge la maintenance des applications du domaine « Gestion de la relation allocataires et échanges partenaires ». Plus de 50 personnes travaillent à Atos sur ce domaine.

Les 5 premiers mois de mon apprentissage sont découpés en deux tâches (menées en parallèle) :

- **Analyse fonctionnelle, encadrée par Kevin ROYERE, Analyste dans l'équipe « Portail CNAF »**

Le portail partenaire est un portail qui permet aux partenaires de la CNAF (mairies, collectivités territoriales, crèches, …) de faciliter les transactions/inscriptions/suivis. Il se compose de nombreux services comme Offre bailleur[1] ou CDAP.
L'équipe « CDAP-HABPPS » d'Atos réalise le service CDAP qui signifie Consultation du Dossier Allocataire par le Partenaire et le service d'habilitation au Portail Partenaire Sécurisé (HABPPS).

Dans ce contexte j'ai intégré l'équipe CDAP-HABPPS. L'équipe est composée de 8 personnes. J'ai effectué les tâches d'analyse fonctionnelles telles que relecture et mise à jour des cahiers des charges, guide utilisateur et spécifications fonctionnelles détaillées ou encore la création de « User Stories » étant donné que nous travaillons en méthode agile.

Ces tâches m'ont permis de comprendre le métier d'analyste fonctionnel qui demande une bonne compréhension des besoins du client et la retransmission de façon simple et complète aux développeurs. Ceci m'a permis de mieux comprendre ce que délivre Atos comme prestation puis d'intégrer une équipe. La difficulté la plus importante rencontrée fut l'absence de réponse exacte aux problématiques. Le métier d'analyste fonctionnel dépend de la compréhension et de l'analyse des besoins du client, il y a donc rarement de réponse exacte et on doit sans cesse remettre son travail en cause afin de répondre aux besoins du client.

La suite de ce rapport ne décrit pas ces travaux.

- **Valorisation des méthodes de data science à Atos, dans le contexte CNAF, encadrée par Clémentine NEMO, responsable fonctionnelle du projet ASTRID (client Air France)**

La data science n'est pas un domaine de compétence développé par l'agence Atos de Sophia-Antipolis. Dans ce contexte mon apprentissage est l'occasion pour Atos de

[1] Permet la dématérialisation des déclarations des bailleurs

fédérer les collaborateurs à la data science et de démarcher les clients sur de nouveaux axes de réflexion.

Dans le contexte du marché avec la CNAF, le premier sujet de travail était sur la valorisation des données de l'open-data CNAF. Mes 5 premiers mois m'ont permis d'aborder ce sujet mais aussi de faire d'autres propositions plus intéressantes pour Atos et la CNAF (architecte, commerciaux, …).

Ce rapport présente la démarche de recherche des différents sujets et les conclusions auxquelles j'ai abouti.
Notons que ce travail a été mené techniquement en autonomie. Bien que mon encadrante ne soit pas sur le même site que moi, elle était très disponible pour répondre à mes interrogations.

Description du travail proposé

Mon sujet pour ces 5 premiers mois d'alternance était:

"Analyse de l'Open Data de la CNAF"

Atos est prestataire de service pour la CNAF depuis de nombreuses années. Afin de garder la relation commerciale et face à la concurrence, il est vital d'être force de proposition. C'est dans ce cadre que mon alternance a été proposée, afin d'apporter un ace « data science » qui n'a jamais été proposé.

Pour mener à bien ce travail, des travaux de réflexion, de recherche, de présentation et d'implémentation ont dû être réalisés. L'ensemble de ces parties sera développé en s'appuyant sur la démarche suivie lors de mon apprentissage.

Présentation de l'Open Data

On rappelle la définition :

"Un Open Data est une donnée numérique dont l'accès et l'usage sont laissés libres aux usagers."

Pour rappeler le contexte, le gouvernement ainsi que des entreprises privées acceptent de prendre part à la mission Open Data de plus en plus à partir de 2011.

C'est dans ce contexte que l'Open Data de la CNAF a été ouvert en 2014. Dans celui-ci, il y a un Open Data interne et un Open Data accessible à tous. Pour ma part, je n'ai travaillé que sur l'Open Data accessible à tous.

L'Open data a pour vocation le libre accès mais surtout la réutilisation des données. En effet, avec la mise en place des Open Data, on voit de nouvelles réutilisations et applications apparaître régulièrement. Ceci favorise donc les usagers qui ont accès aux données, mais aussi les entreprises ou organisations qui apportent une plus-value à leurs données et qui peuvent bénéficier d'idées et d'applications qu'ils n'auraient sans doute pas développer (par manque de temps, de moyens et parfois d'idées). L'idée de l'Open Data est donc souvent bénéfique. Cependant on voit de grandes disparités de réutilisations entre les différents Open Data, on peut penser que le domaine d'intérêt, la forme des données et leur accessibilité jouent un rôle important dans cette disparité.

Mise en place du sujet

Les travaux de réflexion et de recherche sont parties intégrantes de mon apprentissage car ils m'ont permis de proposer et décrire le travail que j'allais réaliser par la suite.

Le sujet étant assez vague, le premier travail à réaliser est de s'imprégner de l'Open Data de la CNAF et comprendre les données qui sont exposées dans celui-ci.

L'Open Data de la CNAF est un ensemble de 82 jeux de données regroupées par catégories: "aides diverses aux familles", "engagement de service", etc. Il contient 324 fichiers au format CSV qui représentent pour la plupart un nombre d'allocataires ou chiffres relatifs aux services (courriers, visites, taux d'appels, …). Ce ne sont que des chiffres par

communes ou par Caisses d'Allocations Familiales (CAF) correspondant quasiment[2] aux départements français avec une granularité temporelle trimestrielle.
L'adresse de l'Open Data est : http://data.caf.fr et une grande partie des données est disponible sur http://data.gouv.fr.

Ma a permis d'identifier plusieurs sujet à traiter :

1. **Compréhension des données par une visualisation utile aux présentations ATOS-CNAF**

Le domaine « Gestion allocataires et échange partenaires » sur lequel se positionne le contrat Atos, adresse (entres autres) les problématiques de réduction de la charge de travail des collaborateurs de la CNAF et d'optimisation des nouveaux services partenaires et utilisateurs.

Dans ce contexte, seuls les jeux de données relatifs aux travaux que mènent les équipes Atos sur le marché CNAF sont utilisés : visites à l'accueil, nombre d'appels téléphoniques, utilisation du site web, …
Un travail de visualisation de 21 jeux de données est donc réalisé et sera détaillé plus tard. Ils décrivent seulement les CAF (données au niveau départemental).
Suite à cette visualisation, je proposerai l'idée de regrouper les 102 CAF existantes en groupes.

2. **Valorisation de l'Open Data de la CNAF**

Suite aux difficultés rencontrées lors de la réalisation de la visualisation des données de la CNAF, l'idée est de rendre cet Open Data plus facile d'accès et de réutilisation.

Pour cette partie, je m'appuie sur l'utilisation des technologies de web sémantique afin de valoriser l'Open Data de la CNAF. La mise en place d'outils améliorant l'Open Data permettrait à un utilisateur quelconque (pouvant ne pas connaître l'analyse de données) de profiter de la connaissance contenue dans l'Open Data.

3. **Analyse des tweets mentionnant les CAF**

Ma dernière proposition propose d'analyser les tweets en rapport avec les CAF afin d'analyser les ressentis à propos de la CNAF. Ceci ne concerne pas l'Open Data directement mais pourra sensibiliser les collaborateurs au fait que certaines données sont accessibles et peuvent être utilisées pour de quelconques applications.

[2] 2 CAF pour les Pyrénées Atlantiques + une CAF appelé caisse nationale maritime

Description du travail réalisé

Comme on a pu le constater précédemment, le travail réalisé est directement lié aux idées et propositions apportées. C'est pour cela que l'on détaillera le travail réalisé suivant les idées en indiquant les difficultés rencontrées pour chacune d'entre elles. On apportera une description chronologique ainsi que les retours des différentes présentations afin de mieux comprendre pourquoi le travail s'est focalisé sur la mise en place d'un outil utilisant le web sémantique.

Compréhension des données par une visualisation utile aux présentations ATOS-CNAF

Afin de m'imprégner de l'Open Data de la CNAF et valoriser ce temps pour Atos, je décide tout d'abord de faire des visualisations sur les données relatives aux services proposées par la CNAF et à la charge de travail de ses agents.

Pour ce faire, voici les étapes réalisées:

- Récupération/Téléchargement des données (21 jeux de données, 139 fichiers CSV)
- Nettoyage des données
- Concaténation des fichiers
- Affichage/Création de visualisation

Voici un exemple de résultat obtenu:
Fichiers utilisés:

- Amplitude horaire à l'accueil physique de 2009 à 2015 (7 fichiers, 1 par année)
- Amplitude horaire à l'accueil téléphonique de 2009 à 2015 (7 fichiers, 1 par année)

Ici, on représente la moyenne sur l'ensemble des CAF

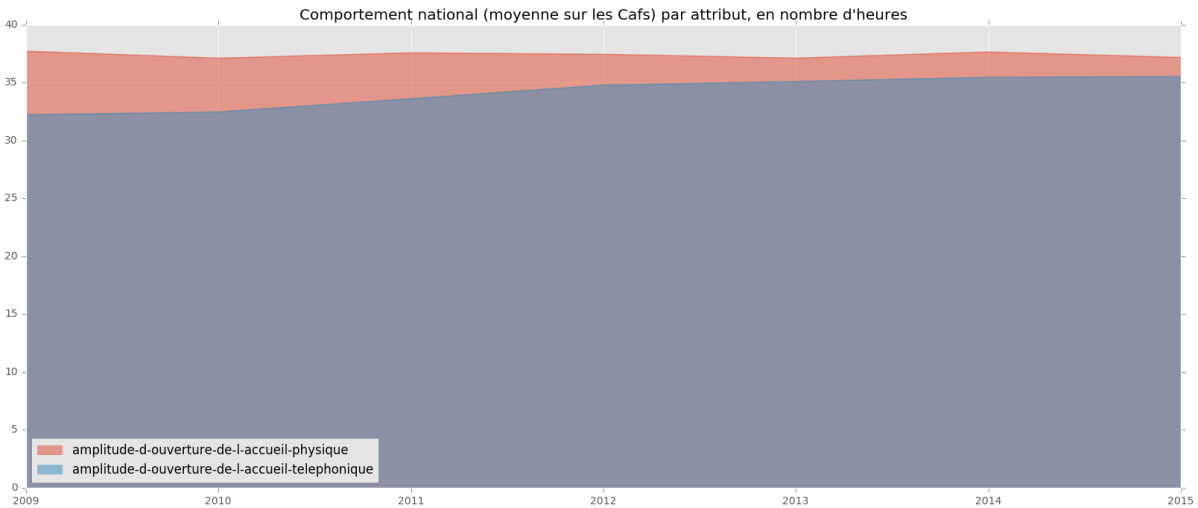

Figure 1 Comportement moyen des amplitudes horaires des CAF

Ceci n'est qu'un exemple de visualisation. Lors de ma présentation, je fournis 8 visualisations résumant les 21 jeux de données utilisées sous différentes formes (diagrammes en bar, camemberts, diagramme en boîte, ...) complétées par des

commentaires qui permettent d'analyser les informations contenues dans ces jeux de données.[3]

Une fois cette visualisation réalisée, je constate la disparité des données en fonction des CAF. C'est pour cela que je décide de les classifier grâce à un algorithme de « clustering » appelé « K médoïdes ». Sans entrer dans le détail de cet algorithme, l'intérêt de celui-ci dans ce cas est de pouvoir regrouper les CAF par groupe (ici 8 après étude des données) en évaluant la similarité de leurs attributs (sur les 21 jeux de données fournis) et de trouver la CAF la plus représentative de chaque groupe (elle correspondra au « médoïde » du groupe).
Les applications ATOS étant testées par certaines CAF avant d'être mises service, on pourrait proposer ces CAF comme « testeuses » afin d'être représentatif de l'ensemble des CAF. En effet, les CAF d'un groupe ayant des fonctionnements similaires au point de vue des données utilisées, une CAF permettant de représenter l'ensemble du groupe semble intéressant.
Voici la représentation en groupe des CAF:

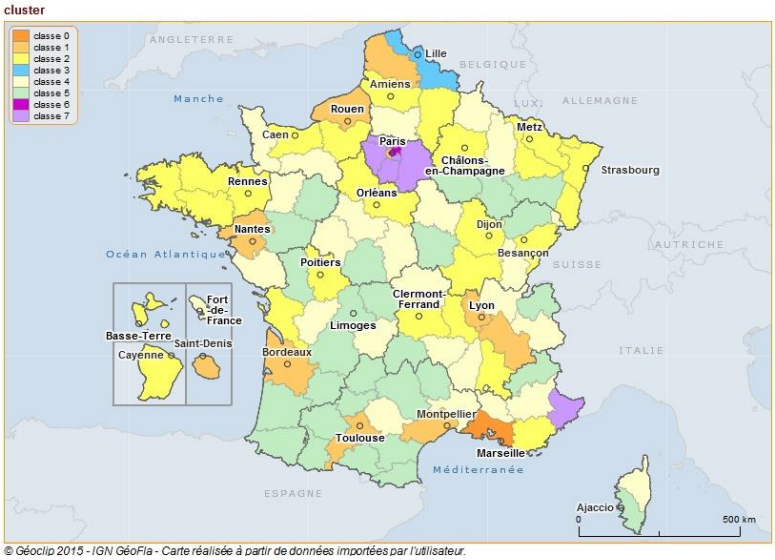

Figure 2 Représentation des 8 groupes de CAF

On voit la répartition des 8 groupes sur la carte de France. Les représentantes de chacun des groupes sont : (par ordre de la classe 0 à la classe 7)

- caf des Bouches du Rhône
- caf de l'Hérault
- caf du Loiret
- caf du Nord
- caf de Béarn et Soule (Pyrénées Atlantiques[4])
- caf des Hautes Pyrénées
- caf de Paris
- caf de l'Essonne

Je n'ai pas trouvé le moyen de les représenter de façon pertinente et simple sur la carte.

[3] Autre visualisation représentée dans le cadre d'une présentation en annexe
[4] Le département des Pyrénées atlantiques étant le seul à avoir gardé plus d'une CAF dans son département: Caf de Béarn et Soule et Caf du Pays Basque et du Seignaux.

Les principales difficultés rencontrées à cette étape sont:

- L'absence de fichier contenant les ressources (URL) afin d'automatiser le téléchargement
- Le temps de nettoyage qui est plus important que prévu. En effet, l'élément commun de chaque fichier à concaténer est le nom de chaque CAF. Cependant ce nom comporte régulièrement des erreurs typographiques (accents, …), d'intitulé (certains noms ont changé avec les années) et d'espaces. Il faut donc tous les traiter afin d'avoir une concaténation possible
- C'est une nécessité de représenter les groupes sur une carte afin d'avoir une représentation visuelle, compréhensible facilement. Les outils à ma disposition ne me permettent de représenter le centre des clusters (par une étoile par exemple) sur la carte.

Lors d'une analyse de données, on indique souvent que le temps de "visualisation" (regroupant le nettoyage, téléchargement, etc…) correspond à environ 80% du temps de l'analyse. Les 20% restant sont divisés entre l'analyse et la classification ou prédiction. Dans ce cas, je n'ai pas dérogé à la règle car mon temps s'est divisé de cette façon.

Valorisation de l'Open Data de la CNAF par l'utilisation du web sémantique

1. Présentation du Web Sémantique

Le web sémantique est la mise en place d'un format standardisé des données (représentation en triplets) permettant l'application de requêtes. La puissance du web sémantique se fait par la création d'un vocabulaire associé permettant de structurer et d'apporter de la connaissance aux données. Cet outil uniformisant les données a pour but global de pouvoir réutiliser l'ensemble des données libres sur le web de façon simple et de regrouper les informations de données indirectement liées sur le web (ici, par exemple regrouper directement les données de l'Open Data de la CNAF et l'Open Data gouvernementale, même si ces données sont sur des plateformes différentes).
C'est dans ce cadre que le web sémantique est l'outil adapté à l'Open Data. En effet, comme il a été rappelé précédemment, l'Open Data a pour but la facilité d'accès et de réutilisation. Le web sémantique offre ces dispositions grâce à l'uniformisation des données et l'apport de connaissances qui facilite la réutilisation à une personne extérieure (les données nécessitant souvent l'apport de données métiers pour permettre la réutilisation[5]).

2. Travail réalisé

Afin de proposer l'ensemble de ces idées, un travail de présentation devant différents acteurs m'a été nécessaire.[6]

Voici les étapes réalisées lors du développement de l'application:

1. Récupération des données
2. Nettoyage et description des données

[5] Voir le « Venn Diagramm » en annexe
[6] Les présentations peuvent être fournies sur demande

3. Mise en place du vocabulaire représentant les données
4. Transformation des données: du format CSV vers RDF (triplets de données)
5. Mise en place d'une base de données de triplets accessible sur le web
6. Exemple de requête permettant de justifier l'intérêt de la méthode
7. Mise en forme automatique des données sur une carte

1. La récupération des données se fait en améliorant le référentiel d'URL créé pour le téléchargement des données lors de la visualisation.

2-3.Le nettoyage reprend les scripts écrits lors des étapes précédentes et est ensuite amélioré.
La description des données s'est faite en parallèle de la mise en place du vocabulaire. En effet, il a fallu trouver une façon pertinente de décrire les données en les classant et regroupant.
J'ai opté pour 3 classes: la CNAF, les CAF et les communes
Les propriétés étant regroupées par aides (puis composante ou attributs) ou par service.
Ceci permet de créer un vocabulaire de 3 classes et plus de 100 propriétés liées entre elle par héritage, similarité ou influence.

4. Cette étape nécessite de savoir ce qu'est un triplet: il est composé d'un sujet, une propriété et un attribut. Exemple :
"Nicolas a pour nom Turpault" s'écrit: sujet:Nicolas propriété:Nom "Turpault"
Ceci est une explication minimale mais permettant de comprendre que toutes les données doivent être regroupées sous un format "sujet, propriété, attribut". Le format RDF est un format de fichier stockant les triplets.
Afin de faire cette conversion, je regroupe toutes les données par date (c'est à dire par trimestre car c'est la temporalité la plus faible). Ceci me permet de faire un fichier RDF par trimestre. On ne rentrera pas dans les détails techniques, mais cette méthode permet d'accéder à l'ensemble des données ou à seulement celle d'une année, d'un trimestre ou d'une période donnée.[7]

5. Cette étape se grâce à un outil s'appelant Fuseki.

6. Voici la comparaison d'une requête entre l'Open Data actuel et la nouvelle solution:
Le langage de la requête est SPARQL, le langage qui permet les requêtes sur les triplets. Ce qu'il faut comprendre de ce tableau n'est pas la requête, mais le fait que les données peuvent être récupérées en une seule requête :

[7] Les détails techniques sont en annexe

Requête	Avant	Après

Description des aides de ma CAF (ici Alpes Maritimes) en décembre 2015; plus précisément le nombre de foyers allocataires par aide (celle qui ont cette information)	-40aine de fichiers à télécharger
	-Une ligne ou une valeur à récupérer dans chaque fichier
	-Les concaténer et afficher les résultats
		prefix caf: <http://data.caf.fr#>
		select * where {
		graph ?g{
		?y a caf:Caf ;
		caf:DEP "6" ;
		?p [caf:NB_ALLOCATAIRES ?nball] }
		?g caf:year "2015";
		caf:month "decembre" }

Le nombre de visites à l'accueil ces dernières années (les valeurs sont par CAF)	7 Fichiers à télécharger et à concaténer avant d'avoir le résultat	select * where {
		graph ?g{
		?y a caf:CAF;
		?y caf:VISITES ?nbVis .}
		}

7. Grâce à l'application Google Maps et à l'application Corese, on peut afficher ces valeurs sur une carte interactive:

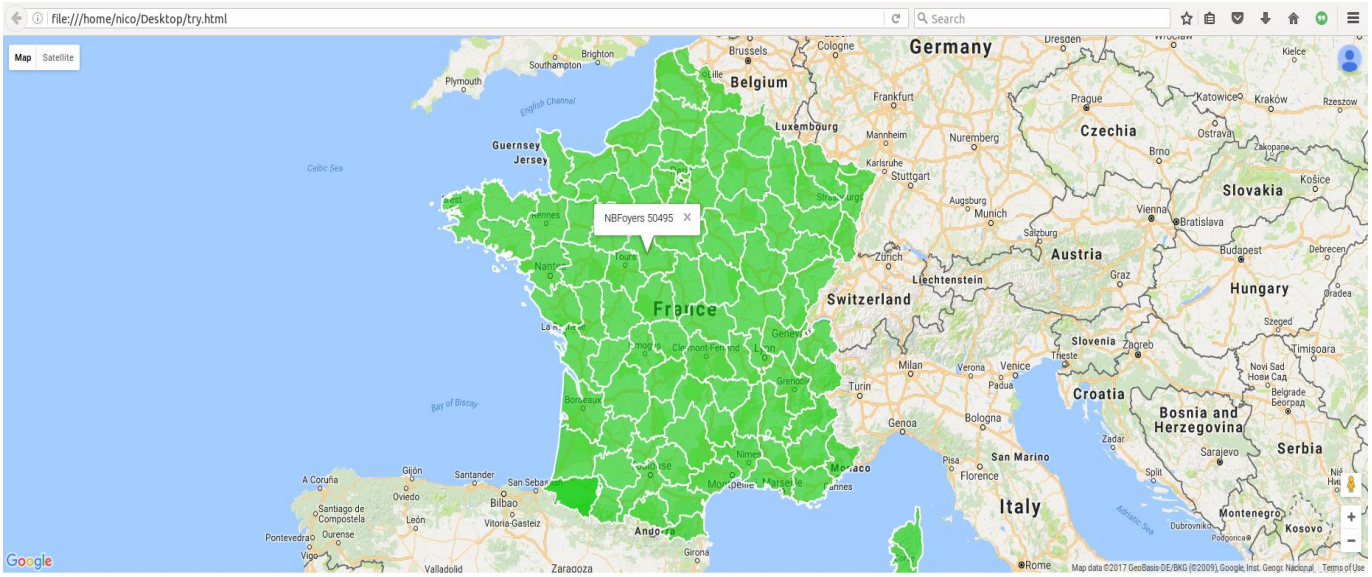

Figure 3 Représentation du nombre de foyers allocataires par CAF

On peut aussi afficher des résultats récupérés sur le "dbpedia" français (une base de données de triplets représentant des données françaises tirées de wikipédia) couplées avec notre base de données (une seule requête)[8]:

[8] Voir la requête en annexe

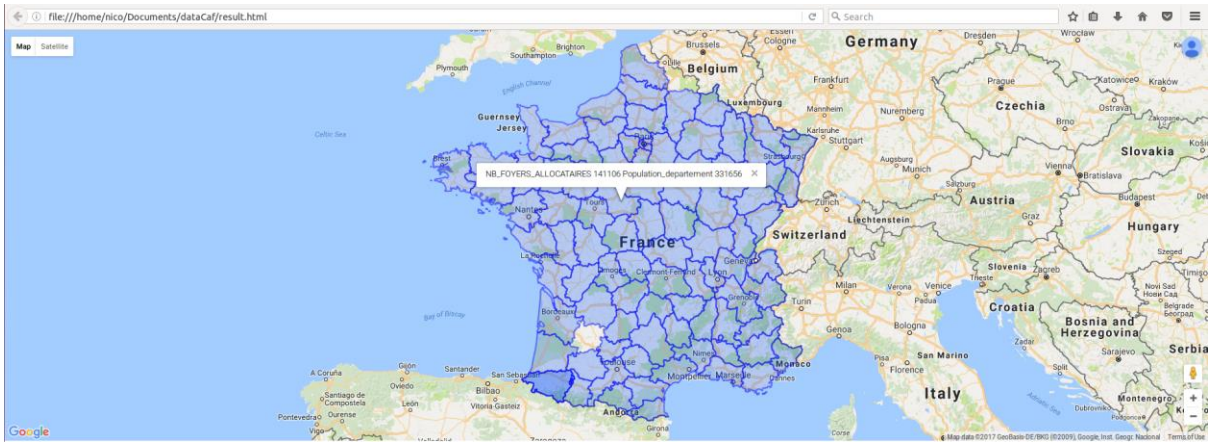

Figure 4 Représentation du nombre de foyers allocataires et de la population de chaque département

On voit ici une anomalie pour les Pyrénées Atlantiques étant donné qu'il y a deux CAF, la similarité ne s'est pas faite (possibilité d'amélioration) et la Corse vu que ses numéros de départements contiennent des lettres et ceux-ci ne sont pas représentés de la même façon dans les deux bases de données (data.caf.fr et data.gouv.fr).

Les difficultés rencontrées lors de cette étape sont:

- Les structures internes des fichiers CSV étant très différentes, le nom des propriétés est à des endroits souvent différents (nom du fichier, nom de colonne, ligne) et m'oblige à adapter les scripts pour chaque structure.
- Les méthodes pour récupérer les âges doivent être adaptées car : les âges étaient représentés parfois par une ligne dédiée, ou le nom de la colonne et régulièrement accompagnés de texte à supprimer
- Impossibilité d'utiliser Datalift (présenté dans « Travail de présentation et d'initiation à la data science ») qui est un outil de conversion automatique pour faire la totalité de la conversion due à la complexité présentée ci-dessus. De plus, on aurait eu un problème de mémoire lors de la transformation de tous les fichiers sur cet outil.
- Impossibilité de faire apparaître l'ensemble des communes (+30 000) sur l'application Google par problème de mémoire, on est donc limité à environ 10 000 communes (représentées et interactives) sur une carte.

Les améliorations possibles sont:

- Faire une carte permettant d'afficher les CAF, puis un détail par commune seulement pour une CAF sélectionnée.
- Mise en place d'une interface pour faire abstraction des requêtes (Afin de laisser libre utilisation à l'ensemble des utilisateurs) en proposant à l'utilisateur des boutons et listes déroulantes simples.
- Finir les conversions de fichiers car seulement 242 fichiers sur 321 ont été convertis à cause des structures différentes nombreuses et un temps imparti.

Analyse de tweets

Cette étape s'est inspirée de travaux réalisés lors de travaux pratiques au sein du cours de « Data Science » enseigné par Mr Fillatre à Polytech Nice Sophia. Cette idée m'a vraiment inspirée pour une entité telle que la CNAF puisque j'ai souvent entendu parler mes

camarades étudiants parlé de leur CAF en positif ou en négatif (pour le montant de leurs aides, leurs délais, …). Je me suis donc demandé si la population de ne partageaient pas ces idées concernant leurs CAF sur les réseaux sociaux.

L'intérêt de cette étape est donc de montrer une approche différente de l'analyse de données en utilisant des données que l'on ne pense pas toujours à utiliser.
L'analyse de tweets n'a pas été si évidente car des outils très puissants existent et sont disponibles mais la plupart de ceux-ci sont en anglais.
J'ai donc entraîné un algorithme simple sur des tweets en français en utilisant des hypothèses minimalistes afin de montrer les possibilités d'un outil comme celui-ci. Le but n'est donc pas d'avoir des résultats mais de montrer la possibilité d'une telle approche.

Le but de cette approche est de définir si les tweets qui concernent les CAF sont positifs ou négatifs.
L'hypothèse simpliste (et critiquable) utilisée est: Les tweets qui contiennent un émoticône positif sont positifs et ceux contenant un émoticône négatif sont négatifs.
Voici les étapes de mon travail:

- Téléchargement 100 000 tweets contenant un émoticône positif et 100 000 tweets contenant un émoticône négatif
- Entrainement d'un algorithme appelé "Bag of words"
- Prédiction de l'état des tweets concernant les CAF

Pour vulgariser l'algorithme de "Bag of words" :
On compte le nombre d'occurrence d'un mot au sein de chaque classe (positif ou négatif ici) et dans l'ensemble global afin d'attribuer un "poids" à chaque mot lorsqu'il apparaît dans un tweet à prédire (si le tweet à prédire contient un poids de mots annotés positifs plus important que le poids de négatifs, il est positif).
Avec cette approche très simple on obtient ces résultats:
Sur 1000 tweets,

- Tweets positifs: 44%
- Tweets négatifs: 56%

Il y a de nombreuses améliorations possibles pour une approche minimaliste telle que celle-ci et des difficultés à prendre en compte si on veut améliorer l'approche.
Les améliorations possibles sont :

- Avoir une hypothèse beaucoup plus forte pour l'attribution des sentiments
    - utilisation de bases de données annotées de texte par exemple
    - La compléter grâce à une petite application afin de spécialiser l'algorithme aux tweets (peut aussi permettre de tester notre algorithme)
- Définir plus de classes d'émotions (haine, colère, joie, neutralité, …)
- Créer un outil de classification en temps réel afin d'afficher les tweets positifs sur le site de la CNAF

Voici un aperçu d'une application créée pour valider mes résultats (mon algorithme n'étant précis qu'à 60% après ces tests):

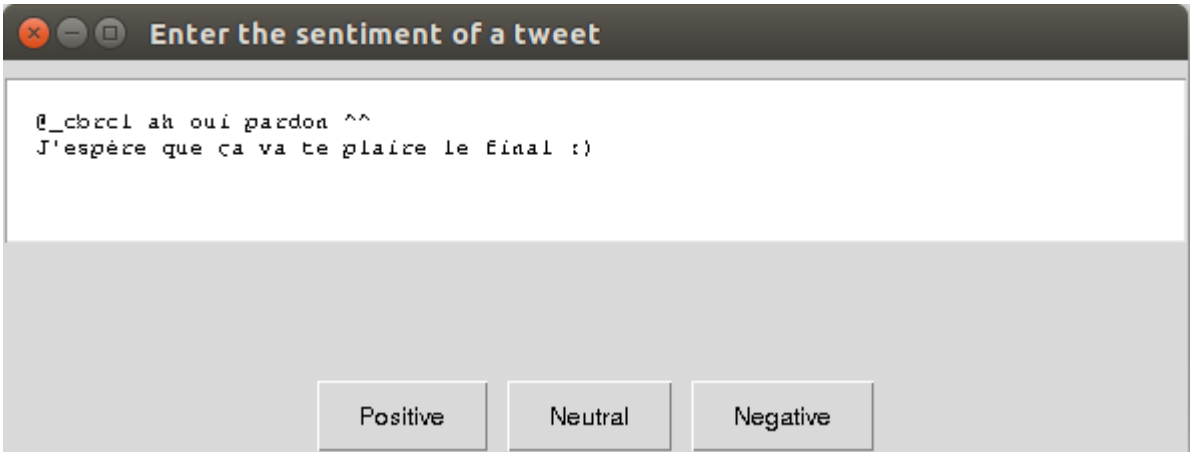

Figure 5 Application permettant de classifier un tweet

Les difficultés qui sont ou pourront être rencontrées sont:

- Pas d'algorithme pré-entraîné disponible pour des tweets en français
- Peu de bases de données textuelles annotées en français
- Peu de tweets concernant les CAF (moyenne entre 3 et 4 par jour)
- Nécessité d'utiliser une infrastructure plus importante pour avoir un algorithme assez robuste entraîné sur une quantité suffisante de données
- Algorithmes nécessitant un volume de données très important

Travail de présentation et d'initiation à la data science

Afin de présenter la data science et présenter mes travaux, j'ai réalisé plusieurs présentations.

- **Présentation aux collaborateurs d'Atos travaillant dans l'équipe CNAF**

La première présentation se déroule après avoir accompli les travaux de visualisation, d'analyse tweets, et une brève approche pour la valorisation de l'Open Data (Décembre).
2h me sont attribuées pour présenter l'ensemble de mon travail.
Les collaborateurs présents lors de cette présentation sont:

- le directeur du marché CNAF à ATOS (Luc CHAMPALLE)
- le directeur technique du site, (Alain CIVEL)
- l'architecte logiciel du projet "Portail Partenaires" (Sébastien RENE)
- mes deux tuteurs d'alternance (Kévin Royère et Clémentine NEMO)

Je présente mon domaine, mes compétences ainsi que mes travaux réalisés. Lors de cette présentation, la décision de poursuivre l'idée d'un outil de web sémantique se met en place.
La visualisation et l'analyse de tweets restent donc à l'état présenté dans ce rapport.
Pour la mise en place d'un outil de web sémantique afin de valoriser l'Open Data, je montre qu'on peut le faire sur une partie des données de façon très simple en utilisant l'outil "Datalift" développé par Gabriel KEPEKLIAN (Directeur en Recherche et Développement sur le site d'Atos à Bezons) :

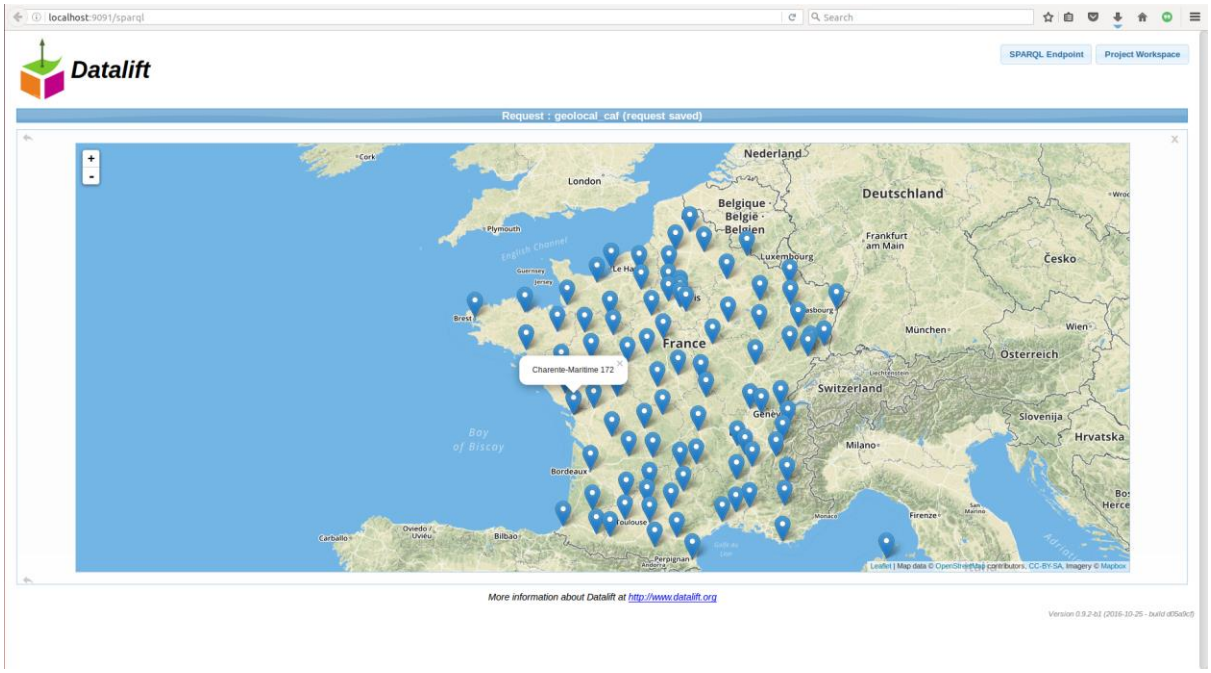

Figure 6 Représentation des Sièges de CAF avec leur numéro en utilisant DATALIFT

Sur l'image ci-dessus on voit le travail de visualisation réalisé avec Datalift sur une partie des données. L'avantage de cet outil est d'avoir une application qui fonctionne de façon très simple (l'abandon de cet outil est dû à la complexité des données de la CNAF).

Les conclusions de cette réunion sont:

- Le travail de visualisation et d'analyse de l'Open Data est intéressant, cependant la plus-value de celle-ci n'est pas très importante pour Atos
- La quantité de tweets concernant les CAF n'est pas assez importante pour permettre de développer une application qui apportera des résultats percutants à la CNAF. Cependant, l'analyse de texte pourrait surement être appliquée à un autre support que les tweets concernant les CAF.
- L'idée de proposer un web de données est bonne. On la retient afin de la présenter à un panel d'auditeurs plus important. Le fait de pouvoir prendre contact avec un expert (Gabriel KEPEKLIAN, créateur de Datalift) en web sémantique au sein d'Atos fut un atout au développement de ce projet.

- **Présentation aux responsables de projet**

Ma deuxième présentation s'est déroulée durant le comité opérationnel (ensemble des directeurs de projets du site Atos Sophia) de Janvier afin de leur présenter mon domaine d'activité et mes compétences dans le cas où ils auraient besoin de compétences telles que celle-ci au sein de leur projet.
Lors de cette réunion, une quinzaine de personnes étaient donc présentes. Cette présentation fut vraiment différente de la première, car j'avais seulement 15 minutes de parole qui m'ont permis de présenter mon domaine, mes compétences et d'énoncer une simple description de ce que j'avais fait à ce moment.
Les conclusions de cette réunion sont :

- Le domaine est intéressant, il faut poursuivre

- Mes travaux sur l'Open Data peuvent être intégrés dans le catalogue permettant de proposer de nouvelles idées aux clients
- Questionnement sur l'aspect commercial du travail sur l'Open Data

- **Présentation en clientèle à la CNAF**

Alain CIVEL ayant de nombreux contacts à la CNAF, il m'a permis de rencontrer et présenter mon travail à Laurence SALAGNAT (responsable fonctionnelle sur Sophia) et André DEWERDT qui était lui à distance (expert CNAF sur l'aspect données à Lyon).
Lors de cette réunion, j'ai eu 1h afin de présenter mon travail et les différentes propositions possibles.
Cette réunion s'est conclue par:

- Possibilité de présenter mes travaux sur l'Open Data aux personnes gérant l'Open Data au sein de la CNAF
- Mise en place d'un nouveau sujet pour la deuxième partie de mon alternance qui est l'analyse des rapports d'erreur sur l'ensemble de leurs applications étant donné le nombre journalier qu'ils en reçoivent.
- Nouvelle réunion prévue le 1 mars

Le développement de l'application utilisant le web sémantique m'a permis d'identifier certaines difficultés commerciales dont je ne connaissais pas l'existence. La mise en place de nouvelles idées nécessitent l'appui de responsables en interne et chez le client. De plus, même si un intérêt certain est identifié lors d'une présentation, il faut souvent du temps avant qu'elle ne s'accompagne de propositions et réalisations concrètes. On voit l'exemple de l'espacement des différentes réunions qui se concluent souvent par la mise en place d'une autre réunion.

Planning du semestre

Date	Commentaires
20/09/2016	Présentation, mise en place, début de lecture du cahier des charges
21/09/2016	Lecture et retour sur cahier des charges HabPPS - Découverte de l'Open Data CNAF
22/09/2016	Lecture et retour sur les spécifications fonctionnelles détaillées d'HabPPS
23/09/2016	Compréhension des "cascades" des SFD + point avec Kévin
28/09/2016	Fin des SFD + CDC, rendu des commentaires à Kévin - Relecture du Guide utilisateur
05/10/2016	Compréhension fichier Interne + PPT conventions parallèles
10/10/2016	PPT pour conventions en parallèles + début des User stories HabPPS
11/10/2016	Création des User Stories (US) HabPPS
12/10/2016	US HabPPS
17/10/2016	Fin de la première version des User stories
18/10/2016	Revue des users stories et sfd + début d'analyse sur l'Open Data
19/10/2016	CNAF data: préparation des fichiers
24/10/2016	CNAF data: préparation des fichiers + concaténation des fichiers
25/10/2016	CNAF data: tri des fichiers pour faire un dataset, premières lectures
26/10/2016	Premières analyses sur le notebook niveau national
31/10/2016	Clustering cafs et cartographie
01/11/2016	Open data, essai open refine
08/11/2016	Médecine du travail
09/11/2016	Cnaf data - utilisation d'Open Refine et comparaison
14/11/2016	Cnaf data, remise en forme, finitions du notebook - abandon d'open refine
16/11/2016	SFD, modification de la convention pending
21/11/2016	Fin des sfd habpps convention pending + datacaf carte + début diapo
21/11/2016	Data cnaf, ppt intro

22/11/2016	visualisation + tweets exemples
23/11/2016	Finalisation de présentation + montée en compétence sur maquettage
24/11/2016	tutos maquettage + finalisation cartographie, clustering
25/11/2016	Présentation offre bailleur fonctionnel + Présentation PPT à Clémentine + nouvelles idées à développer
28/11/2016	Exploration datalift + modifications des SFD
29/11/2016	Amélioration diapo + essais tweets
30/11/2016	définition des besoins pour tweets et datalift
01/12/2016	Tweets téléchargements
02/12/2016	Tweets code fonctionne mais résultats insatisfaisant, possibilité de faire des requêtes sur un rdf créé à partir de csv sur datalift
05/12/2016	Finition présentation, contact école + stéphane giovannangeli propose la présentation au comité opérationnel
07/12/2016	Retour diapo clémentine + revue user stories
08/12/2016	Revue user stories + Use cases
09/12/2016	Présentation PPT CNAF data à Kévin
12/12/2016	Présentation de mes travaux à certains responsables du projet CNAF + US HabPPS
14/12/2016	US HabPPS
15/12/2016	US HabPPS
16/12/2016	Réunion tuteur apprenti
19/12/2016	Modification US
20/12/2016	Modification US et finition
21/12/2016	Datalift travail de découverte
22/12/2016	Essai de cartes après découvertes des nouvelles données
23/12/2016	Blocage sur Datalift pour certains jeux de données et taille des données
02/01/2017	Implémentation en python de concaténation des fichiers
04/01/2017	Implémentation en python du changement de nom (nettoyage)
05/01/2017	Ontologie (vocabulaire)
06/01/2017	Ontologie (vocabulaire)
09/01/2017	Ontologie + début convertion csv to rdf
11/01/2017	Fichier sur Drive pour montrer les différents RDF à faire + reflexion du temps représenté en rdf
12/01/2017	Solution pour le temps en rdf: par fichier + update google sheet avec url
13/01/2017	Réflexion CSV vers RDF
16/01/2017	CSV vers json des services CAF
18/01/2017	Classes pour fichier avec composantes
19/01/2017	Généralisation aux fichiers CAF
20/01/2017	CSV vers json
23/01/2017	CSV vers json - Présentation de mes travaux lors du comité opérationnel (directeurs de projets)
25/01/2017	Questionnaire ACCOS (analyse de données pour Alain)
26/01/2017	MAJ du fichier sur les propriétés et URLS + conversions csv vers json
27/01/2017	Arrêt conversion csv-json (2/3 fait) et début application
30/01/2017	Json vers RDF de tous les fichiers + récupération des données géographiques
01/02/2017	données géo + sttl transformation afin de créer un fichier lisible par google maps automatiquement
02/02/2017	Corese requête et javascript + html carte
03/02/2017	Présentation de mes travaux à deux personnes de la CNAF avec Alain
08/02/2017	Réussite carte, mise en place d'un sttl plus puissant permettant plus d'affichage de communes (jusqu'à 10 000 communes contre 3 000 avant)
10/02/2017	Mise en place d'un serveur fuseki local pour accéder aux données via le web
15/02/2017	Possibilité de lier les données avec dbpedia - finalisation des requêtes montrant la puissance de l'outil
16/02/2017	Début de rapport de mes travaux pour la CNAF
17/02/2017	Retour rapport alternance + modifications

Conclusion

Cette première partie d'alternance fut très enrichissante autant du point de vue technique, que commercial ou même humain.
Il m'a tout d'abord fallu m'adapter au contexte d'une grande entreprise travaillant avec une institution nationale. On peut par exemple constater que la présentation de mes premiers travaux était prête en décembre, et il m'a fallu attendre un mois avant de le présenter aux directeurs de projet, et deux mois avant de le présenter au client.
De plus, la multitude de services au sein de la CNAF ne m'a pas permis de présenter mon application directement aux bons collaborateurs (les responsables de l'Open Data). Il faut d'abord contacter des collaborateurs connus avant d'atteindre un nouveau groupe de collaborateurs au sein de l'institution.
L'absence de référent technique au sein d'Atos fut parfois un problème. Heureusement que Mr CORBY (professeur de web sémantique) et Mr KEPEKLIAN (collaborateur Atos basé sur Lyon, expert en web sémantique) m'ont permis de répondre à mes différentes interrogations et m'ont même parfois donné quelques conseils pour éviter d'être trop technique lors d'une présentation. La littérature m'a aussi été d'une aide indispensable.
Le fait d'être intégré dans une équipe a été un atout pour mon alternance. Ceci m'a aidé à comprendre ce qui se fait au sein d'Atos et trouver ma place.
Cette première partie d'alternance a été très positive dans l'ensemble et m'a permis de rencontrer un nombre important de collaborateurs qui pourront m'être d'une aide importante pour la suite. L'intérêt de pouvoir s'adresser au bon collaborateur est indéniable.
J'ai vu l'intérêt de proposer de nouvelles idées et les implémenter mais savoir les expliquer et les présenter fut ici le point clé de mon alternance.
Ma deuxième partie d'alternance va se dérouler sur un sujet différent. Alors que la première partie m'a permis de développé mes compétences en web sémantique, la deuxième me permettra de proposer des solutions en Machine Learning qui est un domaine dans lequel j'ai un peu plus d'expérience dû à mes stages et différents projets précédents.

Abstract

My apprenticeship was about data science and semantic web. I was working for Atos and CNAF (National Family Allowances Office) which is Atos' customer. I had to propose new ideas to, first, reinforce the partnership between CNAF and Atos, and then bring new abilities into Atos Company. To do it, I have done some technical work which has been concluded by presentations.
Presentations were key points of my apprenticeship because they decided what would be the following steps of it. My abilities to present and explain in a simple way my domain have been improved thanks to it. I have faced some issues that I needed to manage one by one asking questions or searching the solution. The main idea that has been developed in this report was: improving CNAF Open Data. This idea has been proposed to my managers and then developed before being presented to CNAF employees.

Bibliographie

Liens les plus importants :
http://data.caf.fr/site/
http://www.data.gouv.fr/fr/
http://wimmics.inria.fr/corese
http://www.datalift.org/
http://www.caf.fr/sites/default/files/cnaf/Documents/DCom/Quisommesns/Presentation/Rapport_dactivite/RA%202015.pdf
http://docplayer.fr/782250-Jdev-atelier-datalift.html
https://www.kaggle.com
The Semantic Web : Semantics and Big Data – Philipp Cimiano, Oscar Corcho, Valentina Presutti, Laura Hollink, Sebastian Rudolph

Principales newsletters (veille technologique) :
www.datascienceweekly.org
http://roundup.fishtownanalytics.com

Annexes

Présentation d'Atos

Atos est une entreprise internationale spécialisée dans les services informatiques et fait partie des 10 plus grandes SSII au niveau mondial. Elle est également le leader du paiement en ligne sécurisé pour les entreprises en France avec le système SIPS. L'entreprise a réalisé un chiffre d'affaire de 10 686 millions d'euros en 2015, et elle collabore avec environ 77 100 personnes dans 48 pays différents.

Figure 7 : Atos chiffres mondiaux

Atos fournit aux clients du monde entier des services transactionnels de haute technologie, des solutions de conseil et de services technologiques, d'intégration de systèmes et d'infogérance. Grâce à son niveau d'expertise technologique, Atos est présent dans différents secteurs d'activités : secteur public, secteur financier, industrie, énergie et télécom.

Complément d'information sur la CNAF

La sécurité sociale en France est divisée en quatre branches :

- Maladie
- Vieillesse
- Recouvrement
- Famille

La CNAF est la branche « Famille » de la sécurité sociale française et elle s'occupe de tout ce qui concerne les allocations familiales. Elle est présente sur tout le territoire au travers de 103 Caisses d'Allocations Familiales (CAF)

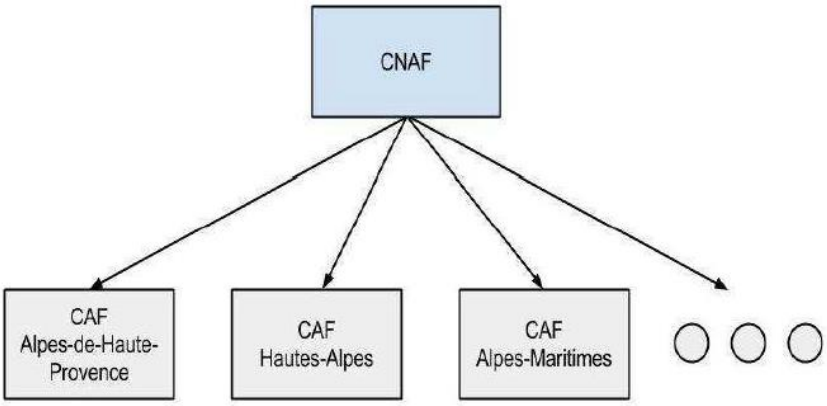

Figure 8 : Diagramme organisationnel CNAF et CAF

Chacune de ces CAF est un organisme de droit privé à compétence territoriale. Elles sont chargées de verser aux particuliers des aides financières à caractère familial ou social, dans des conditions déterminées par la loi, dites prestations légales. Chaque CAF est donc en partie indépendante et mène des actions sociale (compléments de revenus, de suivis, de conseils) qui peuvent différer des autres CAF.
Les particuliers percevant ces aides sont appelés « allocataires » et ils sont un peu plus de 11 millions.
Le site caf.fr a donc pour mission de permettre à la CNAF de mettre à disposition de ses millions d'allocataires des informations les concernant que ce soit au niveau national (CNAF) ou départemental (CAF).

Data Science

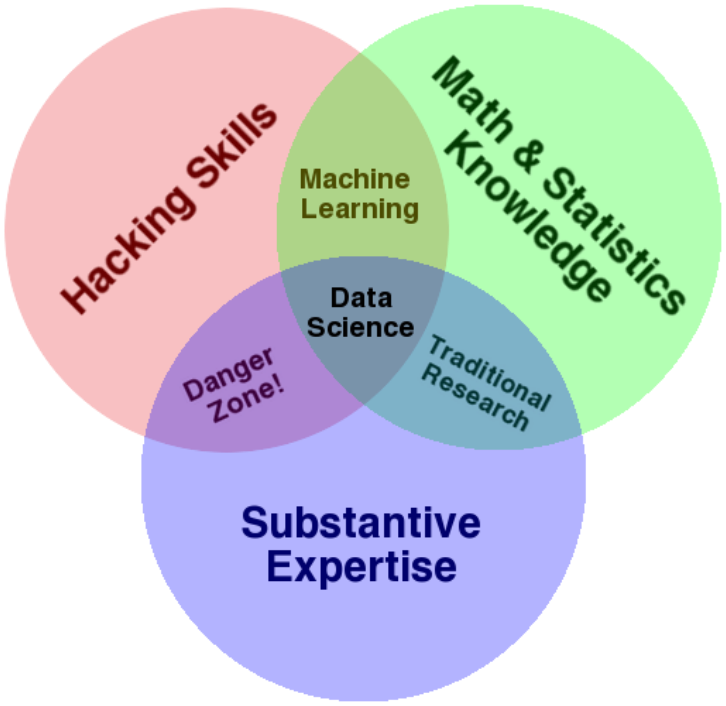

Figure 9 Venn Diagramm

Le « Venn Diagramm » représente les compétences nécessaires à l'utilisation de la « Data Science »

Visualisation

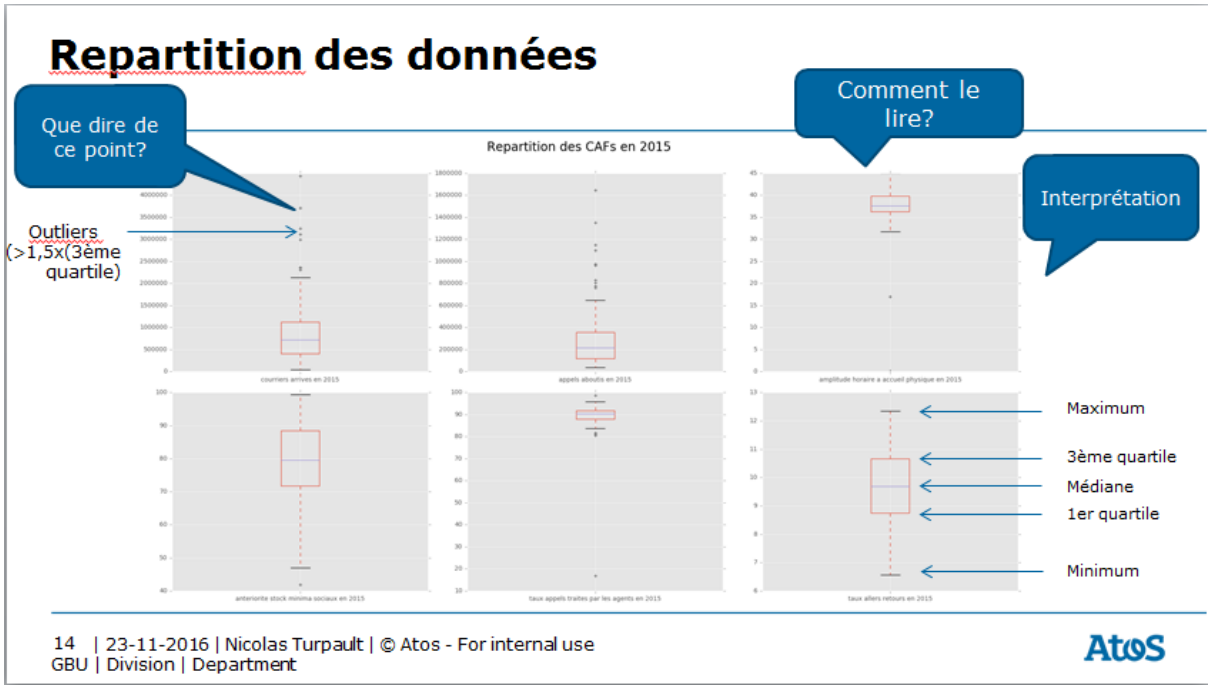

Figure 10 Exemple de diapositive présentant la répartition des CAF suivant certains attributs + explication du diagramme en boîte

Requête de web sémantique (SPARQL)

Requête permettant de regrouper les informations de notre base de données et celle de wikipédia:

```
select ?y ?nball ?pop ?coor
where
{
select ?y ?pop ?nball ?coor where {
  graph ?g{
          service <http://fr.dbpedia.org/sparql> {
          select distinct ?x ?val ?pop where {
          ?x a ont:Department ; prop:insee ?val ; ont:populationTotal ?pop.
          }
          }
  ?y a caf:Caf ; caf:DEP ?dep ; caf:NB_ALLOCATAIRES ?nball .
  filter(str(?dep) = str(?val))
  }
  ?g caf:year "2015"; caf:month "decembre" .
  ?z caf:dep ?dep
  ?z caf:geometry ?coor .
}
}
```

Apport technique

Le langage utilisé pour faire l'ontologie est OWL.
On a utilisé des règles d'inférences pour parvenir à lier les données des communes avec les données départementales.
Les règles d'inférences ont été utilisées au sein du logiciel Corese.
Les données géographiques ont été récupérées sur un site annexe en format KML pour avoir les contours de chaque département et les afficher sur une carte.
On crée un fichier HTML intégrant les données par département aux coordonnées en utilisant STTL signifiant Sparql Template Transformation Language. Il permet d'appliquer un 'template' aux données sous forme de triplet. On crée un fichier HTML complet grâce à ce langage en indiquant où doivent être les données de chaque triplet dans celui-ci.

L'intérêt d'un triple store comme Fuseki cité dans ce rapport est de pouvoir accéder aux données avec n'importe quel logiciel de requête Sparql (corese, datalift en faisant partie).