Université Paris Nanterre
UFR PHILLIA

MONDECA

Julien CHANSON

Master TAL-IL

RAPPORT D'ALTERNANCE

[Alternance effectuée du 18/09/2017 au 06/07/2018]

Mondeca

35 boulevard de Strasbourg, 75010 Paris

Des bases de connaissances à la fouille de textes, lier des méthodes et domaines

Sous la direction de :

M. Arnaud CASSAIGNE

Soutenu le 06/07/2018 à l'UFR Phillia

Université Paris Ouest Nanterre La Défense

200 Avenue de la République 92001 Nanterre cedex

Année Universitaire 2017 - 2018

Je tiens à remercier en permier lieu Stéphane Senkowski et Christophe Prigent pour m'avoir acceuilli en stage dès la fin de ma licence et donner par la suite l'occasion de pouvoir continuer de travailler au sein de Mondeca en contrat de professionnalisation durant toutes mes études de master. Ce n'aurait pas été possible sans la recommandation d'Hacène Cherfi et sans mon tuteur, Arnaud Cassaige, à qui je dois aussi un encadrement exemplaire. J'ai pu avoir sous sa tutelle une solide formation sur la résolution de problèmes informatiques, la prise en main d'outils et le dévellopement de solutions. Je tiens aussi à remercier Benoit Carcenac qui a été comme un second tuteur et qui m'a accordé, très tôt, une très grande confiance et de belles opportunités qui n'ont cessé de me faire apprendre. Tout naturellement, je remercie aussi l'intégralité des équipes de Mondeca, qui, toujours avec délicatesse, continuent de me faire remarquer mes erreurs lorsque j'en fais et à me guider pour devenir un meilleur professionnel. Je tiens aussi à remercier Delphine Battistelli et Jean-Luc Minel pour m'avoir accompagné dès la licence sur la voie du traitement automatique des langues et, aujourd'hui, à la fin de mes études. Ces remerciements s'étendent au reste des équipes pédagogiques du master DEFI et TAL présentes à l'université Paris Nanterre, Sorbonne Nouvelle et à l'INALCO ainsi qu'à mes camarades. Sur une dernière note plus personnelle, je tiens finalement à remercier ma famille et mes amis proches qui se reconnaitront, sans qui rien de tout cela et plus encore ne se serait jamais passé.

Résumé :

Ce rapport de stage relate un projet d'entreprise mené pour un client afin de répondre à ses besoins en fouille de textes entre autres. Il a nécessité d'allier de nombreuses méthodes et domaines. Il reste un projet en cours à l'écriture de ce rapport. Ces domaines et méthodes incluent la catégorisation de paragraphes et d'éléments, les bases de connaissances, la sémantique inférentielle et la fouille de textes. Tous ces différents aspects ont été traités avec des solutions conçues et développés par Mondeca ainsi qu'avec GATE, un logiciel libre de droits de traitement de texte.

Mots-clés : Traitement Automatique du Langage, fouille de textes, Bases de connaissance, Sémantique, Inférences, Catégorisation, GATE

Abstract :

This internship report relates an enterprise project led for a client in order to meet his needs in text-mining among others. It needed to mix numerous methods and domains. It was still not completed when this report was written. The domains and methods used encompass subsections and elements categorization, knowledge bases, semantics and text-mining. All aspects have been treated with business solutions made and developped by Mondeca and also with GATE, an open source software for text engineering.

Keywords : Natural Language Processing, Knowledge bases, Semantics, Inferences, Categorization, GATE

Droits d’auteurs

Table des matières

Table des illustrations

Table des tableaux

Introduction

Il est primordial de commencer ce rapport en le contrastant sur un point important. Dès la fin de ma dernière année de licence en Sciences du langage – Ingénierie Linguistique, je me suis engagé dans un stage de 3 mois au sein de Mondeca à la suite duquel j'ai suivi deux années en contrat professionnel au cours de mes études en Master. Le travail réalisé et mon implication dans ce projet prend donc ses sources dès janvier 2017.

Réalisé pour le compte d'un client, voilà donc plus d'un an et demi que le projet se développe et continue à ce jour. La part de projet incombée à Mondeca se place autour d'un projet plus grand visant à relier les normes et réglementations ainsi que leurs contenus entre elles. Le tout serait ainsi stocké dans une base de données graphes pour permettre la mise en place d'un outil capable de lier toutes les normes indiquées par des références dans une norme donnée à d'autres. Plus loin encore, l'outil saurait catégoriser les normes en fonction de leur empreinte sémantique, les lier en fonction de leurs termes et définitions communes ainsi que de lister toutes les exigences présentes dans une norme. L'outil a même vocation à relier les réglementations françaises et européennes entre elles. C'est là que le travail explicité dans ce rapport prend place, il faut pouvoir extraire toutes les informations nécessaires à la mise en place de cet outil, non seulement dans le cas des normes, mais aussi dans les réglementations.

Derrière cet enjeu ambitieux se cache donc de nombreuses problématiques. Il existe de très nombreux organismes de normalisation nationaux et internationaux et autant de types de norme pour chaque organisme. Au sein d'un même organisme, les normes peuvent même être amenées à changer de structure, compliquant alors la tâche de fouille de textes. Les normes abordent d'ailleurs tous les thèmes qui puissent être réglementés, de l'agriculture à l'aérospatial. Uniformiser les techniques d'extraction pour la fouille de textes se complique alors graduellement au fur et à mesure que le nombre de types de norme augmente et que les différents éléments à extraire se complexifient. Tout ce développement doit également se faire pour permettre une certaine maniabilité à l'équipe finale en charge de l'outil chez le client.

C'est dans cette optique et avec ces problématiques que tout le travail de fouille de textes a pris place au sein de ce projet. Il a été le fruit de nombreuses réunions, retours, réflexions et développements. Encore en cours à ce jour, il concerne principalement quatre collections de normes : les normes françaises, européennes, internationales et étrangères. Nous aborderons dans ce rapport comment le projet a été divisé en différentes étapes cumulatives ainsi que les techniques et méthodes employées pour chacune d'elles.

1. Autour du stage

1.1. Entreprise et cadre de travail

1.1.1. La société et ses solutions

Mondeca est une société parisienne spécialisée dans les technologies sémantiques ainsi que dans la modélisation et la description de la connaissance. Présente au sein du W3C Consortium, la société est impliquée dans l'expansion du web sémantique et la normalisation des outils sémantiques pour internet. C'est une PME fondée en 2000 travaillant aussi bien avec des clients français qu'internationaux (le musée du quai Branly, National geographic, la Nasa etc...). Elle développe et distribue principalement deux logiciels - ITM et CAM – autour de sa solution SCF.

ITM ou *Intelligent Topic Manager* (Illustration 1) est un logiciel spécialisé dans la gestion de référentiels, de la taxonomie au dictionnaire en passant par le thésaurus et globalement tous les formats reconnus servant à encadrer et répertorier la connaissance . Il est fondé sur les standards du web sémantique ; c'est à dire qu'il est parfaitement adapté pour traiter les données au format XML, RDF, SKOS ou encore OWL.

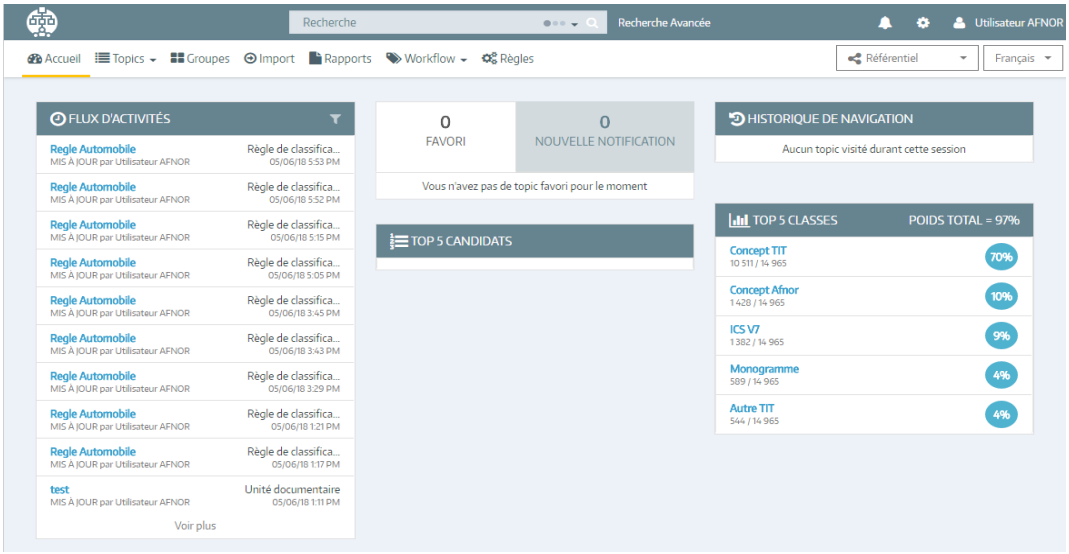


Illustration 1 : ITM – Interface utilisateur principale du logiciel

Au sein d'ITM, les objets sont définis sous des classes. Ces objets peuvent être liés entre eux par des associations de tous types et posséder autant d'attributs de types différents que la classe sous laquelle ils sont définis le permet. Il permet de modéliser n'importe quel type de relation imaginé par les standards des langages du web sémantique et possède une très forte adaptabilité en fonction des besoins de l'utilisateur.

CAM ou *Content Augmentation Manager* (Illustration 2) est un logiciel annexe à ITM spécialisé dans l'extraction et l'enrichissement d'informations. Il permet d'analyser et d'enrichir des données structurées ou non allant du texte à l'image.

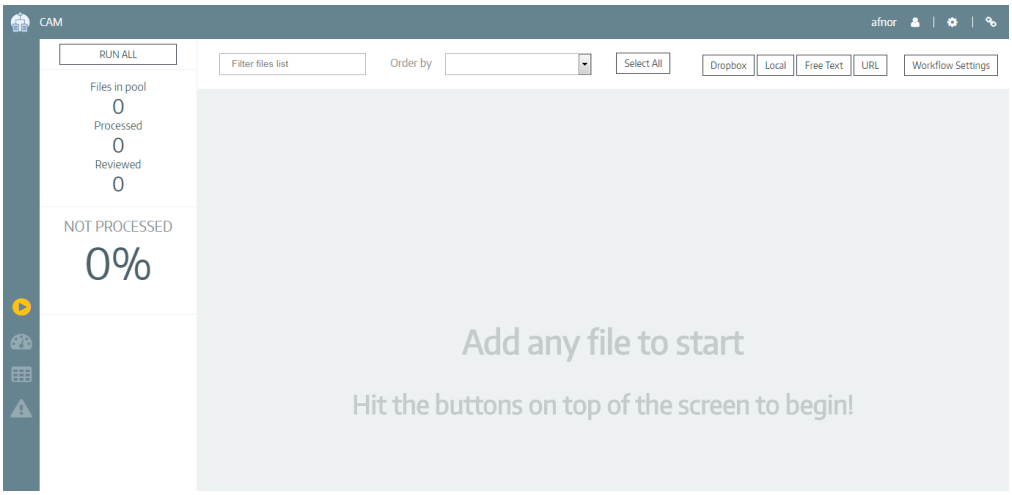


Illustration 2 : CAM – Interface utilisateur principale du logiciel

Fonctionnant sous un système de pipeline, CAM dispose de plusieurs modules à ordonner afin de permettre des traitements précis des documents comme de l'information. Il se connecte à une instance d'ITM de manière à utiliser les données modélisés et éventuellement les exploiter dans son traitement. Bien que CAM comme ITM soient accessibles par n'importe quel type d'utilisateur, les deux disposent d'API et de services Web de manière à automatiser des traitements et pour tout autre type d'opération.

L'alliance de ces logiciels et d'autres développements viennent ainsi former la solution SCF ou Smart Content Factory. De l'acquisition à la visualisation de la donnée, Mondeca propose ainsi à ses clients de multiples manières de répondre à leurs exigences sur différents projets.

Dans les points forts listés sur le site internet de l'entreprise, on cite notamment :

« L'intégration facile: utilisation des standards reconnus et API's simples et faciles à utiliser.
La disponibilité de la plupart des langues européennes et du chinois.
L'utilisation des meilleures technologies d'annotation d'image, de la voix et de la vidéo.
L'expérience des mise en application dans les contextes métier diverses. »

Source : *Mondeca, Making Sense of Content, http://fr.mondeca.com/*

1.1.2. Structuration de la société

On reconnaît 4 services au sein de la société : la direction, le commerce, le *consulting* et la recherche et le développement (sous le même service).

De par sa taille raisonnable d'une vingtaine d'employés, les différents services sont constamment en communication les uns avec les autres et les tâches viennent de ce fait souvent s'entremêler en fonction des compétences de chacun. Il existe cependant des tâches particulières attribuées majoritairement à chaque service (Illustration 3).

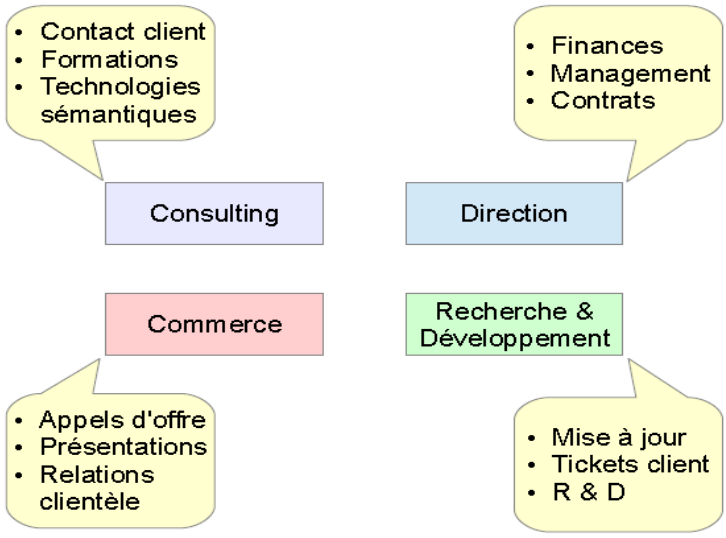


Illustration 3 : Services et taches principales de chacun des services

La direction est naturellement le service le plus important et le plus décisionnaire dans l'entreprise. Elle s'occupe ou délègue toutes les questions d'administration et comptabilité. Elle gère les finances, tout l'effectif de l'entreprise ainsi que les contrats et licences auprès des clients. Elle délègue aussi des responsabilités aux responsables des différents services en fonction des projets.

Le *consulting* ou service des consultants est en charge du contact client. Il contacte les clients, leur répond et reporte ainsi les problèmes rencontrés ou font avancer les projets. Ils sont aussi en charge des formations sur nos logiciels et sur tous domaines qui puissent toucher aux questions de modélisation et de technologies sémantiques. Ils se doivent de gérer les projets aussi bien au niveau des résultats que du temps consommé et d'apporter les solutions les plus adéquates aux clients. Les consultants viennent aussi accompagner le service commercial en cas de présentations des outils.

Le service commercial se charge de répondre aux appels d'offre et d'écrire les propositions aux clients en lien direct avec la direction. Il maintient la relation clientèle nationale comme internationale et prend contact avec entreprises et consultants afin de présenter les solutions proposées par Mondeca et d'effectuer des démonstrations des solutions.

La recherche et le développement sont regroupés sous le même service. Les deux services participent en effet activement à l'intégration de nouvelles technologies et aux développements. La partie recherche du service se concentre majoritairement sur la découverte de nouvelles technologies et à participer avec d'autres organisations au développement ou à la recherche autour du web sémantique et ses technologies. La partie développement de son cotée maintient et développe les fonctionnalités propres aux solutions proposées par Mondeca.

1.1.3. Cadre de travail

Ma place dans la société s'est placée directement sous la responsabilité de mon tuteur, Arnaud Cassaigne, en charge du logiciel CAM dans le service de recherche et développement. Elle a constitué à m'occuper en premier lieu au développement du module dédié à GATE (*General Architecture for Text Engineering*) au sein de CAM. Ce module reprend en effet des pipelines conçues dans GATE pour les intégrer à nos chaînes de traitement et extraction de l'information.

Selon son site officiel : « *Gate est un logiciel libre de droits capable de résoudre presque tous les problèmes associés au traitement de texte* » (Tiré du site officiel après traduction, Source : GATE, *General Architecture for Text Engineering* : https://gate.ac.uk/)

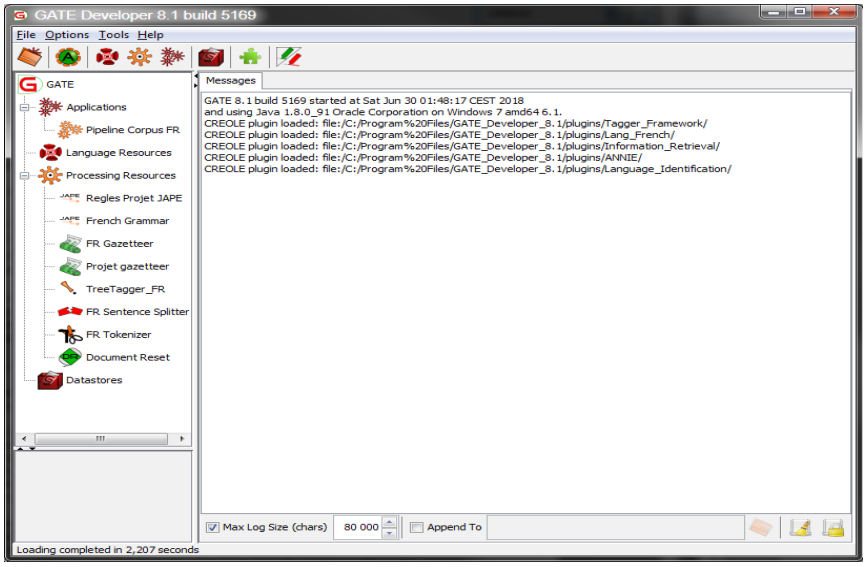


Illustration 4 : GATE – Interface Graphique Utilisateur

Mes premières tâches au sein de l'entreprise ont constitué à l'amélioration d'une pipeline GATE de manière à la rendre capable d'analyser et de reconnaître 12 langues différentes (français, anglais, espagnol, allemand, portugais, italien, polonais, finnois, hollandais, suédois, norvégien et danois). Ce travail a été réalisé au contact de plusieurs spécialistes en langue et a nécessité l'intégration de nombreuses technologies pour l'analyse morphologique de ces langues (plus particulièrement pour la lemmatisation).

Ce projet et les suivants m'ont emmené à approfondir mes connaissances en langage Java et sur l'outil Maven (servant à la gestion des programmes). J'ai donc pu obtenir les compétences nécessaires pour commencer à participer aux développements directement liés à CAM qui a été conçu en Java, langage informatique majeur de l'entreprise. Selon le site officiel d'Oracle :

« Java [...] est conçu pour vous permettre de développer des applications sûres, très performantes et compatibles avec le plus large éventail de plateformes informatiques possible. Ces applications pouvant être déployées sur un ensemble d'environnements hétérogènes [...] et réduisent considérablement les coûts d'exploitation des applications destinées aux entreprises et de celles destinées aux particuliers. »

Source : *Oracle, Java, Technologies:* https://www.oracle.com/fr/java/technologies/index.html

Ainsi, lorsque le projet concerné par ce rapport s'est confirmé et qu'il a fallu s'occuper des développements aussi bien sur GATE que sur CAM, Arnaud m'a délégué la partie du projet lié à la mise en place de *workflows* ou flux de travail. C'est à partir de ce moment que j'ai commencé à activement travaillé avec Benoît Carcenac, en charge des produits au sein de la société, présent dans le service *consulting* et en charge du projet.

Le projet a nécessité de participer à de nombreux ateliers de travail avec le client, en contact direct avec leurs équipes, pour définir au mieux les besoins. Encadrés par des objectifs, les besoins étaient ainsi définis étape par étape de manière à valider les développements faits en conséquence au fur et à mesure.

Nous avons, au courant de cette année, été rejoint par Jing Kang, stagiaire et camarade de classe sur ce projet, ce qui a terminé d'encercler mon cadre de travail.

1.2. Mission réalisée

Le travail réalisé a principalement concerné les questions de fouille de textes, de liens entre connaissances, d'extraction et d'enrichissement des données. Ces questions se sont soulevées au fur et à mesure du projet en fonction des objectifs à accomplir.

Comme nous commencions à le préciser dans l'introduction (page 7), les données à traiter pour ce projet sont des normes et des réglementations françaises et européennes.Il existe de très nombreux organismes de normalisation. Il est primordial de bien comprendre les enjeux autour des normes pour bien encercler toutes les thématiques de ce rapport.

Il existe de nombreuses règles de rédaction entourant les normes sur des sujets très précis qui varient d'organisme à organisme mais qui finissent toujours le plus souvent à avoir des équivalences. Ces règles de rédaction sont accessibles à tous dans des ouvrages expliquant comment rédiger les normes comme le propose par exemple l'Organisation Internationale de Normalisation ou ISO (Source : ISO, Rédaction des normes, Comment rédiger une norme : https://www.iso.org/fr/drafting-standards.html). Elles mettent en avant plusieurs parties et notions présentes dans les normes qui concernent ce rapport.

Parties et notions présentes dans les normes	
Références normatives	Les références normatives sont des mentions à d'autres normes présentes au sein du paragraphe « Références normatives » ou dans les annexes de nature « normatives ».
Références citées	Les références citées sont tout le reste des mentions à d'autres normes qui ne soient pas des références normatives.
Références réglementaires	Les références réglementaires sont des mentions à des réglementations.
Exigences	Les exigences sont des phrases qui énumèrent des conditions sous un degré variable d'acceptabilité. On en reconnaît plusieurs types : exigences, recommandations, autorisations, possibilités et capacités.
Termes et définitions	Les termes et définitions sont présentes sous le paragraphe du même nom. Ils définissent et précisent le sens de certains termes dans le contexte du document.
Descripteurs	Les descripteurs sont des termes du Thésaurus International Technique qui

	sont posés au sein d'un document de manière à en définir les domaines principaux.
Codes ICS	Au même titre que les descripteurs, les codes ICS sont posés au sein d'un document pour en définir les domaines principaux mais sont tirés de la Classification Internationales des normes aussi appelée ICS (*International Classification for Standards*).

Tableau 1 : Présentation des parties et notions des normes inhérentes à ce projet

Toutes ces notions (Tableau 1), sur lesquelles nous reviendrons en détail, ont fait partie des différents objectifs à réaliser au cours de ce projet, certaines fois de manière différente. Il a fallu, en effet, développer deux flux de travail pour ce projet : un premier pour les fichiers au format XML, déjà structuré, et un second pour les fichiers PDF, non structuré. Ces deux flux ont donc dû différer par endroits sur leurs développements au vue de la nature différente des données en fonction des flux.

Chacun des objectifs a donc dû être accompli, avec ses enjeux, dans chacun de ces flux de travail. C'est ici tout l'essence du travail réalisé : développer ou intégrer les technologies nécessaires pour l'établissement de ces flux et le remplissage des objectifs. Bien que j'ai dû l'exploiter pour mes tâches en fouille de textes, la partie modélisation et connaissances n'a donc pas été réalisée par moi mais par Benoît Carcenac en plus de ses autres tâches en tant que responsable du projet dans la société. Les tests, les retours, les recettes et une partie du contact client ont quant à elles été prises en main par Jing Kang.

1.3. Intérêt pour l'entreprise

L'intérêt pour l'entreprise dans cette mission est multiple, aussi bien au niveau de la qualité des services rendus que dans la fidélisation et de la satisfaction du client.

Elle vise dans un premier temps à réaliser toutes les charges dans le cahier des charges du début à la fin du projet en temps et en heure. Notre intervention dans ce projet est avant tout l'intervention d'une société extérieure à celle du client, un certain nombre de jours et d'interventions ont été facturés à celui-ci dans ce projet. Il est donc non seulement nécessaire de livrer le client en temps et en heure sur les objectifs mais surtout de ne pas prendre de retard. En cas de retard sur notre travail, il est délicat de refacturer le client par rapport à un nombre de jours sur lequel nous nous étions engagé : la société peut perdre de l'argent sur le projet.

Bien que la satisfaction client se joue aussi sur la qualité des services rendus, elle s'obtient aussi par l'encadrement et l'accompagnement du client. Savoir prendre contact, répondre aux interrogations, être disponible pour des réunions, tenir à jour le client sont autant de tâches qui veillent à instaurer, non seulement, la confiance mais aussi un certain professionnalisme.

Fidéliser le client est essentiel dans la mesure où, si nous souhaitons continuer de travailler avec lui, cela peut potentiellement l'amener à renouveler son contrat et ses licences avec nous plutôt que de repartir chercher une nouvelle solution auprès de nos concurrents.

Un autre aspect important est l'apport et l'expérience technologique que le projet amène à la société. Elle amène à démontrer notre savoir-faire et à l'améliorer, permettant à la société d'agrandir son éventail de compétence et éventuellement séduire de nouveaux clients avec des solutions plus étendues.

1.4. Limites du stage et du rapport

A l'heure de l'écriture de ce rapport, le projet est en cours depuis plus d'un an et demi et n'a toujours pas été terminé. De plus, la part du projet confiée à Mondeca n'englobe pas l'intégralité du projet chez le client qui dispose de son propre service informatique en charge d'autres développements pour ce projet ainsi que d'autres entreprises qui y travaillent aussi. Ce rapport se concentrera donc avant tout sur les objectifs complétés et validés à ce jour et n'abordera donc pas la question des références réglementaires.

Ce rapport ne présentera pas non plus toutes les fonctions liées aux logiciels et solutions proposés au client puisqu'il a avant tout pour thème la fouille de textes, la catégorisation etc... Il ne s'attardera donc que sur les fonctions et éléments qui sont en lien avec ce projet.

De par la confidentialité imposée sur les données de notre client sur lesquelles le travail a été effectué pour les tests et la validation, les exemples de ce rapport, bien que correspondant dans l'esprit aux données traitées, ne seront que des exemples fictifs ou manipulés. Ils auront pour but d'illustrer le résultat voulu par les objectifs.

2. Objectifs et missions réalisées

2.1. Présentation des objectifs

Nous allons rappeler et préciser les objectifs dans cette première partie tout en recontextualisant le traitement des données en fonction de leur nature (structurées ou non).

Chacun des objectifs doit être complété dans deux flux différents, le premier pour les documents de type XML et le second pour les documents de type PDF. Ces documents peuvent être des normes, des réglementations voir d'autres types de documents comme des ouvrages ou des textes divers. L'avantage de proposer ainsi une solution aux objectifs cumulatifs est que, même si certains éléments spécifiques aux normes ne sont pas repérés (termes et définitions ou références normatives par exemple), le résultat final d'analyse permettra toujours un éventuel enrichissement (avec l'empreinte sémantique ou les références citées par exemple).

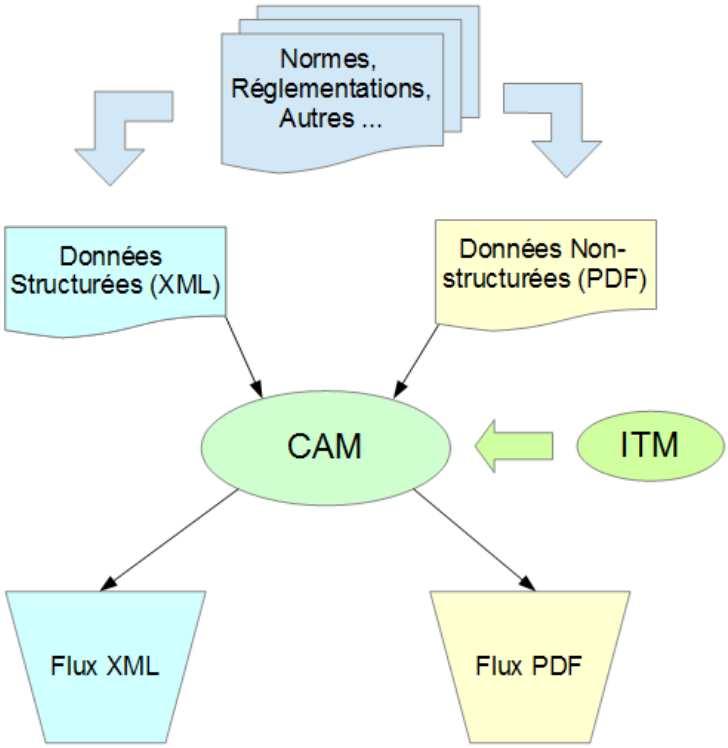


Illustration 5 : Schéma de traitement

Du côté utilisateur, les documents sont envoyés dans CAM dans un premier temps où le flux de traitement utilisé sera ensuite spécifié : l'un pour les XML et l'autre pour les PDF (Illustration 5). Comme CAM est directement lié à ITM, lorsque les terminologies ou classifications sont mises à jour il suffit de recharger les données dans CAM et redémarrer le logiciel pour tout actualiser (Annexe 1 & 2). Pouvoir actualiser facilement les terminologies est essentiel puisque cela participe à définir non seulement les inférences mais aussi les termes à détecter ou qui participent à l'aide dans la détection de certains éléments.

Le tableau suivant récapitule tous les objectifs qui seront abordés dans ce rapport plus en détails dans les parties suivantes (Tableau 2).

Tableau des objectifs	
Objectif 1 – Références citées et normatives	Repérer, extraire et catégoriser toutes les références normatives et citées
Objectif 2 – Exigences et pondération	Repérer, extraire et pondérer les exigences en fonction de leur appartenance (exigence, recommandation, possibilités, etc...)
Objectif 3 – Termes et définitions	Repérer, extraire et associer chaque terme dans le paragraphe approprié avec sa ou ses définitions
Objectif 4 – Empreinte sémantique	Repérer, extraire, catégoriser et développer un score en fonction des termes repérés et de leur appartenance

Tableau 2 : Tableau des objectifs

2.2. Fonctionnement de CAM et GATE

Avant de s'attaquer à la description des objectifs, il est important de prendre un moment pour bien comprendre comment CAM et GATE fonctionnent. Cette compréhension nous permettra de comprendre l'intérêt d'établir deux flux de travail en fonction du type de document.

En plus de son interface d'utilisateur (Illustration 2), CAM dispose d'une interface d'administration appelée AdminUI (Illustration 6). Cette interface d'administration permet d'éditer facilement le logiciel, ses propriétés et ses composants principaux. Elle propose 3 onglets : un onglet *Overview* pour le statut du logiciel (OK ou KO), CAM Configuration pour éditer les composants principaux du logiciel ainsi que ses propriétés ainsi qu'un onglet GATE Configuration pour éditer les composants propres à GATE sans passer obligatoirement par l'interface utilisateur (Illustration 4).

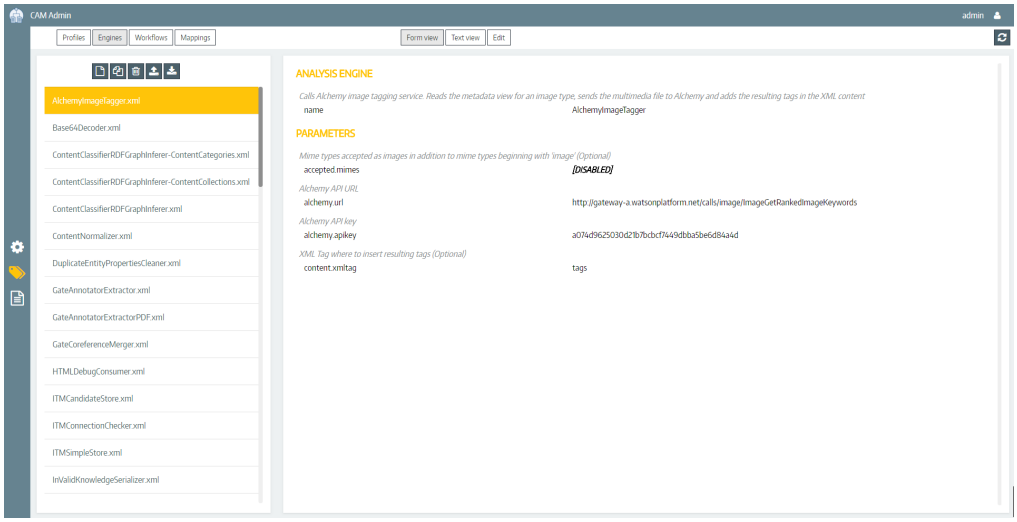


Illustration 6 : Une vue de l'interface administrateur aussi appelée AdminUI

On retrouve 4 catégories au sein de l'onglet CAM configuration (Tableau 3).

Profiles	Contient les fichiers de propriétés pour définir les utilisateurs, les terminologies, classifications, scripts et connexions à d'autres services.
Engines	Liste et permet la création, la copie, l'édition et suppression de modules.
Workflows	Liste les différents flux de travail disponibles et permet leur édition.
Mappings	Les mappings sont des fichiers XML permettant la transformation des annotations dans le résultat de traitement du module GATE pour être listé sur CAM.

Tableau 3 : CAM configuration

Les modules sont donc individuels alors que les flux de travail viennent les organiser (Annexe 2.1). On peut alors créer autant de flux de travail souhaités que de modules.

On retrouve 4 autres catégories au sein de l'onglet GATE configuration (Tableau 4).

Gate-init	Fichier de propriété de GATE
Gate	Liste les différentes pipelines GATE et permet leur édition.
Gate-resources	Liste et permet la création, la copie, l'édition et suppression de ressources.
Gate-jape	Liste et permet la création, la copie, l'édition et suppression de règles JAPE.

Tableau 4 : GATE configuration

CAM dispose d'un module dédié à GATE essentiel à la tâche d'extraction d'informations et de fouille de textes. C'est pourquoi la configuration de GATE pour l'application est directement intégré à AdminUI : permettre des modifications plus aisément.

Les interfaces pour modifier les composants et configurations du logiciel sont un véritable gain de temps puisque CAM comme ITM sont déployés comme des applications web sur des serveurs et non en local. Cela évite les modifications à distance pour modifier les logiciels.

2.3. Enjeux des différentes collections de normes

Ce rapport se concentre sur le traitement de trois collections. On entend par collection les collections des normes ; qui regroupent les normes par organisme de normalisation. Ces trois collections sont celles des normes françaises, européennes et internationales soit NF (Norme Française), ISO (*International Standard Organization*) et CEI (Commission Electrotechnique Internationale).

Si toutes ces normes finissent aujourd'hui par être disponibles au format XML, ce n'a pas toujours été le cas. Certains organismes de normalisation comme la Commission Electrotechnique Internationale ou l'AFNOR ont été créées en 1906 (Source : Genève internationale, CEI) et 1926 (Source : L'internaute, Création de l'AFNOR) respectivement. Le seul format accessible pour certaines normes devient donc le PDF en raison de l'âge de certains documents, d'où l'intérêt du second flux.

Chacun de ces flux doit par ailleurs être capable de traiter l'anglais comme le français, les deux langues dans lesquelles les normes de ces collections peuvent être avec l'allemand. Il a cependant été choisi à ce niveau de ne pas encore traiter l'allemand.

2.4. Objectif 1 : Références citées et normatives

2.4.1. Repérer toutes les références

L'une des premières missions en fouille de textes a été de repérer et récupérer toutes les références citées ou normatives mentionnant donc des normes. Pour une distinction entre ces deux types de références, nous vous invitons à consulter le Tableau 1 dans la partie 1.2 Mission réalisée.

Afin de pouvoir déterminer ces références, le client nous a communiqué dans un premier temps un tableur regroupant trois informations essentielles : monogrammes, *patterns* et exemples de références de normes.

Les monogrammes sont les éléments constants au sein des références alors que les *patterns* sont les éléments variables. Considérons le tableau suivant (Tableau 5).

Monogramme	Pattern (texte)	Pattern (règle)
NF EN	18060	XXXXX
NF EN	18093:14:00	XXXXX:AAAA
NF EN	3456-546-1	XXX-XXX-X
ISO/CEI	45601-1	XXXXX
ISO/CEI	45601-1:2000	XXXXX-X:AAAA
CEI	4561-A	XXXX-A

Tableau 5 : Exemples de références

Les monogrammes font référence à l'origine de la norme au niveau de l'organisation, on peut parler d'éléments constants puisqu'ils sont déjà déterminés en fonction de leur organisme de provenance. On peut d'ailleurs remarquer que certains organismes collaborent ensemble sur des normes sous le monogramme ISO/CEI par exemple.

Les patterns sont des successions de chiffres et de symboles représentant, en plus du monogramme, le numéro d'identification de la norme. En fonction des changements autour des normes, des références à des parties précises de normes, des rééditions ou des évolutions dans la réglementation, ces patterns ont tendance à évoluer. C'est pourquoi l'on parle d'éléments variables : contrairement aux monogrammes, il est difficile de faire une liste

exhaustive. Regrouper toutes les références existantes dans une même terminologie n'étant pas non plus le but de cet objectif, puisqu'il s'agit d'aussi potentiellement repérer des références à de nouvelles normes.

En premier lieu, le client nous a fourni un tableur nous faisant part des monogrammes et *patterns* à suivre. La méthode choisie a été de procéder par des règles JAPE sur GATE. JAPE est un « Java Annotation Patterns Engine » (Source : GATE, Chapitre 8, JAPE), un langage basé sur Java pour manipuler, supprimer et créer des annotations avec des expressions régulières. Les fichiers JAPE se constituent autour de règles dotées de deux parties : *Right-Hand-Size* et *Left-Hand-Size.* Ils peuvent ensuite être enclenchés par un autre fichier JAPE qui peut lister l'ordre d'exécution d'autres fichiers JAPE présents dans le même répertoire. Cela permet donc de créer des flux à l'intérieur d'un module JAPE sous GATE (Illustration 7).

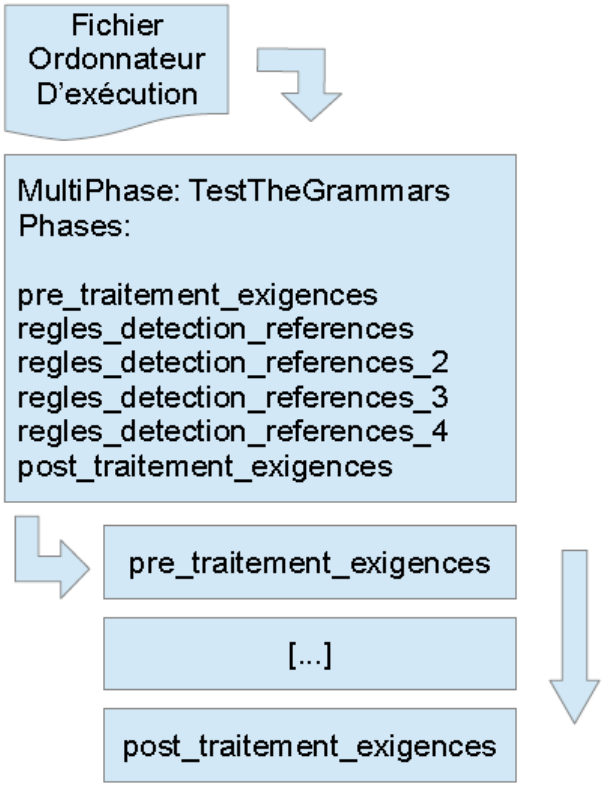


Illustration 7 : Fonctionnement de l'ordre d'exécution des fichiers

La *Right-Hand-Side* (RHS) d'une règle JAPE se constitue d'une grammaire très précise spécifique au langage. En début de fichier, le nom des annotations est inscrit pour être pris en compte suite à quoi diverses opérations peuvent être effectuées sur les différents attributs de l'annotation pour la détection (chaîne de caractères, longueur, type, etc...).

La Left-Hand-Size (LHS) d'une règle JAPE permet de sélectionner l'intégralité des éléments détectés en RHS et les transformer en de nouvelles annotations. La LHS est d'ailleurs

propice à recevoir directement du code JAVA pour augmenter le nombre d'opérations possibles sur les annotations et leurs attributs.

Un exemple de règle JAPE détaillé est disponible en annexe 3.1.

Une fois les règles établies et testées sur un document, nous nous sommes cependant rendu compte que le nombre de références récupérées était bien en deçà des performances espérées avec seulement 5 occurrences contre 46 au total à repérer. Il a donc fallu proposer de nouvelles méthodes pour la récupération la première n'ayant pas convenu.

Nous avons décidé de proposer deux méthodes basées sur les exemples qui nous ont été fournis en même temps que les règles de la première méthode (plus de 71000 exemples de références). La première de ces méthodes reprenait la liste des monogrammes mais diversifiait ses *patterns* de manière à repérer l'intégralité des exemples. La seconde méthode ne se basait que sur des expressions régulières pour repérer l'intégralité des exemples (il ne se basait donc pas sur la liste des monogrammes).

La première méthode a bien pu récupérer les 46 références à retrouver alors que la seconde a récupéré 98 références, soit 52 de plus. Les résultats des trois méthodes peuvent se retrouver en annexe 3.2.

Loin de ne ramener que du bruit (des annotations non-souhaitées), la méthode 3 a permis de mettre en évidence de nouvelles vraies références qui n'avaient pas été repérées dans le document par les équipes du client, notamment de nouveaux monogrammes. En définitive, c'est la méthode 2 qui a été retenue (liste des monogrammes et patterns), de manière à pouvoir créer sur ITM une terminologie des monogrammes qui pourrait être importé depuis CAM en cas de nouveaux monogrammes. Cet import mettrait ainsi à jour les *gazetteers* de GATE. Les *gazetteers* sont des listes d'expressions dans GATE qu'on peut utiliser au travers de modules pour créer des annotations « Lookup » quand une des expressions des listes est retrouvée dans le document. Avec ces annotations créées, on peut ainsi inclure les « Lookup » au sein de règles JAPE pour repérer toutes les références du document.

2.4.1. Distinguer les références normatives des citées

Une fois la question du repérage des références résolue, il a fallu s'attaquer à celle de la distinction entre les références normatives et citées. Les références normatives ne se retrouvent que dans deux cas : en cas de présence dans le paragraphe « Références Normatives » ou dans des annexes normatifs. Il fallait donc catégoriser certaines parties des documents pour pouvoir identifier les références à l'intérieur de ces paragraphes et requalifier leurs annotations en références normatives.

La question n'a posé aucun problème ou que très peu pour les XML : une fois les balises délimitant le paragraphe « Références Normatives » ainsi que les annexes normatifs, ils ont pu être immédiatement catégorisés et les références transformées en références normatives (Annexe 3.3).

Plus fastidieuse, la tâche pour délimiter le paragraphe « Références Normatives » sur les fichiers PDF a été plus compliquée et s'est révélée trop hasardeuse dans le cadre des annexes normatifs. En fonction des organismes, l'emplacement du paragraphe varie de document à document ainsi que l'orthographe du titre introduisant le paragraphe (Annexe 3.4). Non seulement le titre introduisant le paragraphe mais aussi celui du titre suivant, introduisant le paragraphe suivant ; de manière à délimiter la zone représentant les références normatives. Heureusement, l'outil que nous utilisons pour traiter les fichiers PDF les fait passer par une transformation HTML que nous retransformons en XML. C'est à dire que les passages en gras ou italiques sont indiqués par des balises (b pour le gras, i pour l'italique). L'intégralité des titres étant en gras, cela nous a permis une première distinction.

Dans le cadre des annexes normatifs, même s'il était possible de commencer à catégoriser les paragraphes correspondants, il était très hasardeux de délimiter le point d’arrêt dans la catégorisation. Les annexes ne sont pas tous en effet de nature normative, certains peuvent être informatifs voir tout simplement non catégorisés. Cela a soulevé de nombreuses problématiques : comment prévoir si l'annexe normatif est le dernier du document ou suivi d'un autre annexe ? Comment différencier à coup sûr l'index normatif de l'informatif ? La question de reconnaître les annexes normatifs a donc été abandonnée pour le PDF.

2.5. Objectif 2 : Exigences et pondération

Les exigences sont des phrases qui citent des spécifications particulières qui doivent être ou non respectés en fonction de leur nature. On reconnaît plusieurs natures d'exigences qui sont reconnaissables de par leurs formes verbales.

1	Exigence	doit, ne doit pas	shall, shall not
2	Recommandation	il convient de, il convient de ne pas	should, should not
3	Permission et autorisation	peut, ne peut pas	can, cannot
4	Possibilité et capacité	peut, ne peut pas	can, cannot

Tableau 6 : formes verbales de base

On remarque déjà un premier problème pour n'importe quelle catégorisation à effectuer : comment distinguer l'autorisation de la possibilité et capacité si les deux utilisent les mêmes formes verbales ? Il a donc été décidé dans un premier temps de catégoriser les permissions ainsi que les possibilités sous la même catégorie.

Une question primordiale a été posée en réunion qui a alors complexifié l'objectif : étions-nous sûr que les exigences ne puissent pas être exprimés par d'autres formes verbales ? La question est alors remontée chez notre client pour découvrir que, non-seulement, ces formes verbales n'étaient pas les seules mais que les exigences de même nature ne disposent apparemment pas du même degré d'importance selon les formes utilisées. Il fallait donc repérer les exigences mais aussi les classifier en fonction de leur nature et de leur degré d'importance. Nous avons donc décidé d'exprimer la nature des exigences par un attribut catégorie et le degré d'importance par une intercatégorie, d'où la pondération des exigences. La liste complète de cette pondération est disponible en Annexe 4.1

GATE dispose en anglais comme en français d'un découpeur de phrases (*Sentence Splitter*) qui effectue globalement un bon travail à quelques exceptions - que nous aborderons par la suite - et créé des annotations « Sentence » pour chacune d'elles. La méthode choisie a donc été de repérer les formes verbales depuis des règles JAPE de manière à créer une annotation « TempRequirement » contenant à la fois la catégorie de la forme verbale mais aussi son intercatégorie en attributs (Annexe 4.2). Il suffisait donc par la suite de vérifier si une annotation Sentence contenait une forme verbale pour créer une annotation « RequirementSentence » qui reprenait tous les attributs de la forme verbale pour conserver la pondération (Annexe 4.3). Si jamais une phrase comprend de multiples formes verbales, autant d'annotations « RequirementSentence » sont créées en retour.

Nous avons cependant observé quelques problèmes avec la récupération des exigences dans certains cas. Notre outil pour transformer les PDF en XML afin de faciliter le traitement a des difficultés à gérer les PDF en double colonnes ainsi que les tableaux où certaines phrases sont parfois présentes. Cette difficulté entraîne des phrases soudainement coupées qui nuisent à la bonne reconnaissance des annotations les représentant. Au sein des fichiers XML, ce sont les listes à balises ainsi que les tableaux qui nuisent au bon découpage des phrases et donc à la détection des exigences.

2.6. Objectif 3 : Termes et définitions

Le paragraphe des termes et définitions spécifie certaines expressions présentes dans le document. En fonction du type de norme, un terme peut être amené à changer complètement de définition ou même en posséder plusieurs. Dans le cadre de notre objectif, il s'agit d'associer chaque terme de ce paragraphe à sa définition.

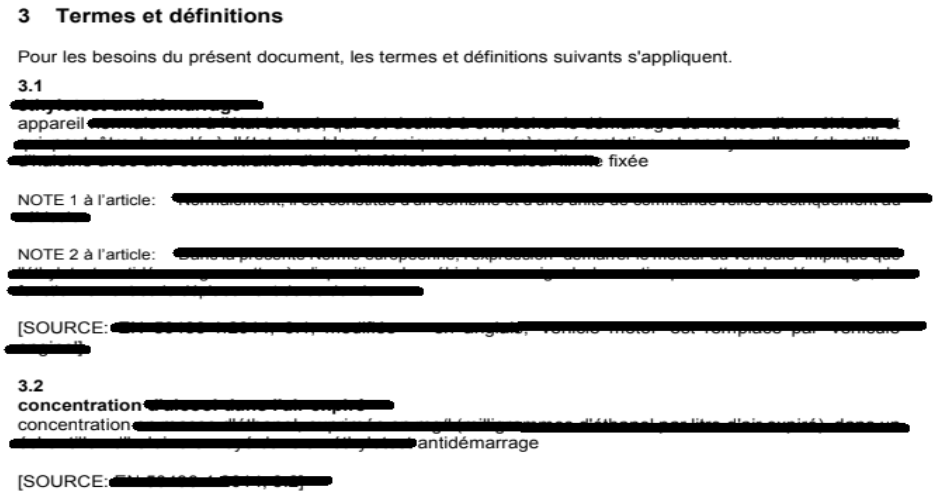
Dans le cadre des fichiers XML, au sein du paragraphe, les termes associés à leurs définitions se présentaient globalement ainsi au niveau de la structure :

```
<concept1 gtext="3.1">
        <multi_termdef>
                <termdef type="principale" lang="fr">
                        <hp format="bold">contenu du titre</hp>
                </termdef>
        </multi_termdef>
        <multi_termdes>
                <termdes lang="fr">
                        <p>contenu de la définition</p>
                </termdes>
        </multi_termdes>
</concept1>
```

Exemple d'un terme associé à sa définition

Afin de bien garder uniquement le terme associé à la bonne définition, il suffisait de se placer au sein de chaque balise concept1 pour récupérer et associer les deux grâce aux balises.

Dans le cadre des fichiers PDF, la tâche a été étonnamment plus complexe de par la nature de base non structurée du document. Sans balises sur lesquelles s'appuyer aussi explicite que pour les fichiers XML, il a fallu recréer un contexte pour pouvoir associer chaque terme à sa définition.



3 Termes et définitions

Pour les besoins du présent document, les termes et définitions suivants s'appliquent.

3.1

appareil

fixée

NOTE 1 à l'article:

NOTE 2 à l'article:

[SOURCE:

3.2

concentration

concentration

antidémarrage

[SOURCE:

Illustration 8 : Termes et définitions

Une fois le paragraphe des Termes et définitions repéré de la même manière pour celui sur les Références Normatives, il a fallu s'appuyer sur des éléments comme la numérotation, la police en gras de certains termes et sur le découpage effectuée par la conversion du PDF vers le XML.

On remarque ainsi plusieurs choses (Illustration 8) :

- La numérotation (3.1, 3.2) permet de séparer chaque terme et définition.
- Les termes sont toujours en gras alors que les définitions non.
- La définition suit toujours directement son terme associé.

Il a donc fallu délimiter dans un premier temps les zones entre chaque terme et définition. Une fois cette délimitation effectuée avec la numérotation, le terme était reconnu par sa police de caractères en gras et annoté ensuite sous l'étiquette « Terme ». La définition par le fait qu'elle soit le premier ensemble de texte après le terme qu'on a annoté comme « Def »(découpé naturellement en balise « p » par la conversion PDF vers XML).

2.7. Objectif 4 : Empreinte sémantique

L'empreinte sémantique symbolise l'intégralité des thèmes et termes abordés dans un document. Ces thèmes et termes sont décidés en fonction de deux terminologies particulières :

- Le Thésaurus International Technique (TIT), créé par l'organisme de normalisation français, l'AFNOR, pour attribuer des thèmes principaux à ses normes.
- La Classification Internationale pour les normes (ICS) qui a été conçu pour servir à l'élaboration de catalogues de normes à toutes les échelles (nationale et internationale).

Ces deux terminologies ont été choisies par le client de par leur contact direct avec les normes. De plus, l'ICS est une terminologie utilisée dans toutes les collections traités actuellement, ce qui prouve son rayonnement et son intérêt.

On appelle aussi les termes issus du TIT des descripteurs, dans le sens qu'ils décrivent effectivement le contenu des normes françaises dans lequel ils sont mentionnés. On les retrouve à la fois dans les fichiers XML ainsi que dans les fichiers PDF sous la balises « descripteurs » pour le XML et sous le paragraphe du même nom en page de garde des PDF pour les normes françaises (Illustration 9).

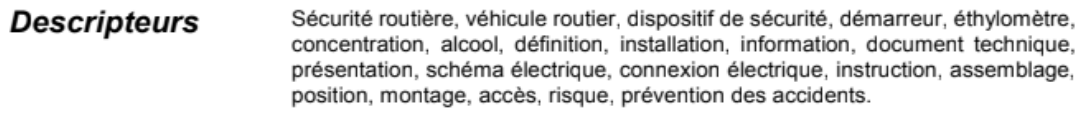


Illustration 9 : Descripteurs dans le PDF

De la même manière, on retrouve aussi des sections dédiées pour ICS dans le XML comme le PDF avec un contraste ; on ne fait pas directement mention d'un terme mais plutôt d'un code pour ICS. Dans la classification, ces codes sont associés à des thème et démontrent une certaine hiérarchie (Illustration 10).

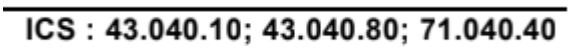


Illustration 10 : Codes ICS liés à la norme dans le PDF

Au sein d'ICS, 43.040.10 fait référence aux équipements électriques et électroniques et 43.040.80 aux protections contre les chocs et systèmes de retenue. Ces deux codes découlent de leur père dans la hiérarchie 43.040 qui représente les systèmes automobiles. 43.040 qui découle lui-même de 43, le sommet de la hiérarchie ; les véhicules routiers.

De la même manière mais sans code pour représenter les termes, TIT établit une relation entre ses termes (Annexe 5.1).

Une fois les termes de ces deux terminologies détectés et annotés dans le document, il fallait pondérer l'importance de l'empreinte sémantique en fonction des termes. Si un code ICS ou un descripteur avait été posé dans le paragraphe approprié, il n'avait pas la même importance qu'un terme d'une de ces deux classifications retrouvé autre part dans le document. Nous avons donc créé 4 classes différentes : les Meta TIT et les Meta ICS qui sont les termes dans les paragraphes prévus pour qu'ils y soient mentionnés ainsi que les TIT et ICS qui sont les autres.

Nous avons par ailleurs utilisé deux fonctions liées directement à CAM pour affiner l'empreinte sémantique : les inférences et le score.

Les inférences sémantiques sont des liaisons de parenté d'un ou plusieurs termes à un autre, c'est à dire qu'un terme inféré n'est pas présent en tant que tel forcément dans le document mais qu'il est du moins lié à un terme qui lui, est présent. A titre d'exemple, 43.040 n'est pas présent dans les codes de l'illustration 10 mais est le père de deux termes qui y sont présents, il pourrait donc être inféré. Nous avons d'ailleurs choisi de n'inférer que les ancêtres, c'est à dire les pères pour éviter une avalanche d'inférences.

Le score est calculé en fonction du nombre d'occurrences du terme dans le texte et en fonction de son indice de pertinence calculé en fonction des modules intégrés dans le flux de travail de CAM. CAM dispose de modules comme le « RelationScore » qui permet de calculer en fonction du nombre et du type de relation d'une entité à sa taxonomie. Le score a principalement été utilisé pour ne garder que les termes les plus importants dans le résultat de l'extraction. Les éléments peuvent ainsi être filtrés s'ils n'ont pas un score assez haut (Illustration 11).

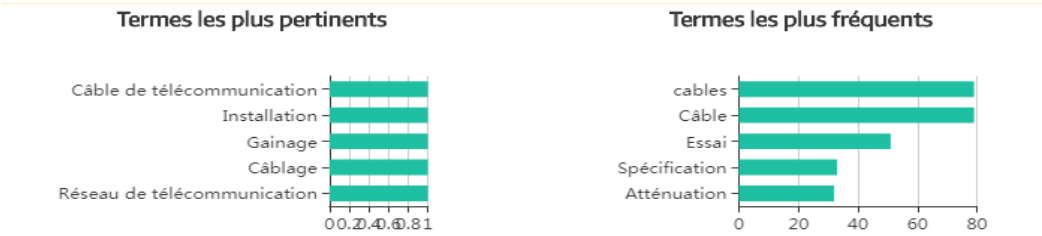


Illustration 11 : Score sur CAM

On peut voir sur l'illustration dci-dessus que de tous les termes pertinents sont à 1 de score : c'est parce que ce sont des Meta TIT ou Meta ICS et leurs inférences, leur score a été directement monté au maximum. Dans les termes fréquents, on peut voir les termes les plus présents dans le document issus des taxonomies.

Conclusion

Au cours de ce rapport nous avons abordé quatre des objectifs du projet et mis volontairement de côté certains dans la mesure où nous n'étions pas assez avancé dans le temps lors de la rédaction de ce rapport pour les aborder. Il reste donc plusieurs étapes à venir pour le projet : des références réglementaires à d'autres collections de normes comme les normes étrangères (qui feront peut-être évoluer les flux de travail pour traiter plus de langues). Nous avons tout de même traité toutes les thématiques présentées dans notre introduction : de la fouille de textes aux bases de connaissances par les terminologies au travers de tous les objectifs présentés.

Notre problématique était de lier tous ces domaines entre eux de manière à répondre aux besoins du client avec les outils et logiciels dont nous disposions. Nous avons ainsi utilisé la catégorisation de paragraphes pour discerner les annotations récupérées dans l'objectif sur les références citées, utiliser des techniques de fouille de textes comme les expressions régulières au travers des règles JAPE dans l'objectif sur les exigences et la pondération, utiliser des terminologies et en manipuler les termes repérés dans nos données pour dégager les plus sémantiquement importants ainsi qu'inférer de nouveaux termes. Le tout pour permettre un traitement automatique de données en langue écrite à notre client, ce qui se marie dans la thématique du traitement automatique de la langue.

Il reste cependant de nombreuses problématiques en suspens à la fin de ce rapport. Comment les données finiront-elles par être exploiter par le client ? Quand est-ce que le projet sortira de son développement pour devenir un véritable outil à utiliser ? Quels seront les enjeux posés par les nouvelles collections sur la chaîne de traitement déjà actuellement mise en place et les divers flux de travail ? Aujourd'hui, le client développe sa propre solution pour transformer les PDF en véritables documents structurés ; permettra-t-elle de résoudre le problème de détection sur les annexes normatifs ?

Ce projet a été jusqu'à aujourd'hui et certainement encore demain, une chance incroyable de prendre contact avec le client, comprendre ses intérêts et problèmes, certaines fois formuler et répondre à ses besoins. En plus d'une expérience technique, cela aura aussi été une véritable expérience humaine où le travail de développement a toujours été motivé par des problématiques exprimées directement par des équipes humaines au cours d'ateliers et réunions.

Références bibliographiques

Mondeca, Making Sense of Content : http://fr.mondeca.com/

GATE, *General Architecture for Text Engineering* : https://gate.ac.uk/

Oracle, Java, Technologies : https://www.oracle.com/fr/java/technologies/index.html

ISO, Rédaction des normes, Comment rédiger une norme : https://www.iso.org/fr/drafting-standards.html

Genève internationale, CEI : http://www.geneve-int.ch/fr/commission-electrotechnique-internationale-cei-0

L'internaute, Création de l'AFNOR: http://www.linternaute.com/histoire/jour/evenement/22/6/1/a/60662/creation_de_l_afnor_association_francaise_de_normalisation.shtml

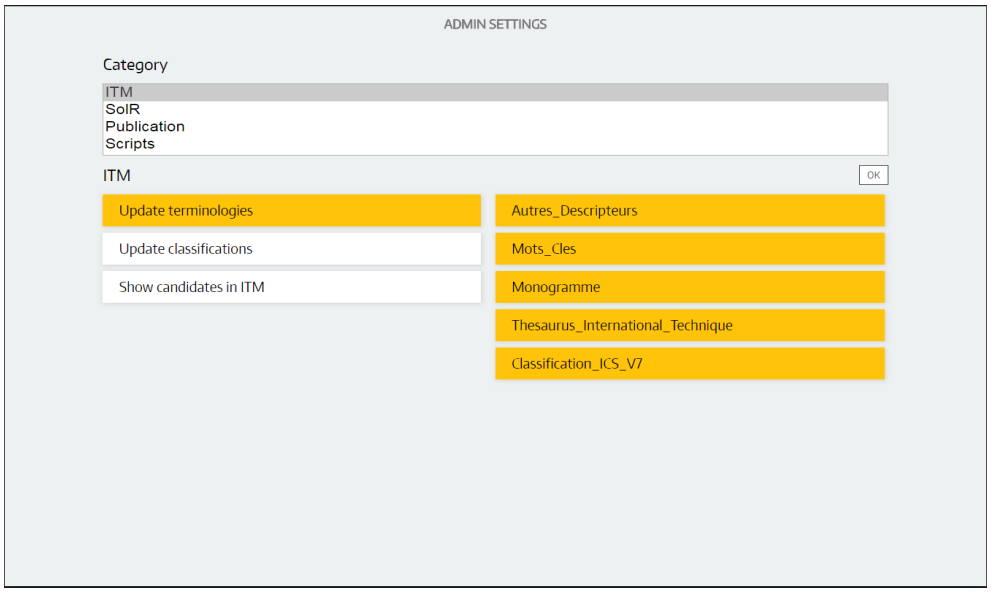
GATE, Chaptire 8, JAPE : https://gate.ac.uk/sale/tao/splitch8.html

Table des annexes

Table des annexes

Annexe 1. Mission réalisée

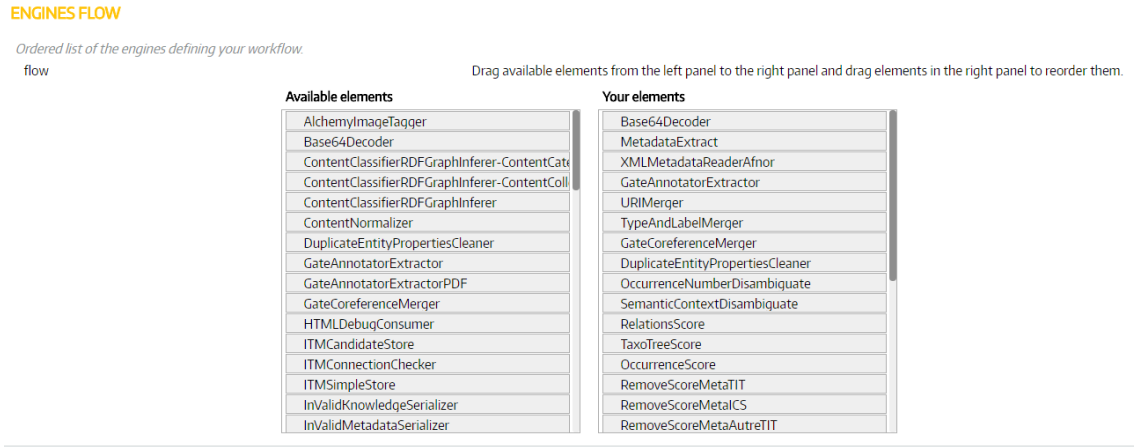
Annexe 1.1. Scripts ITM vers CAM



Ici, on propose à l'utilisateur l'intégralité des terminologies renseignées pour lui permettre de choisir lesquelles recharger.

Annexe 2. Objectifs et missions réalisées

Annexe 2.1. Organisation flux de travail



Les modules peuvent être déplacés et enlevés à volonté.

Annexe 3. Références normatives et citées

Annexe 3.1. Règle JAPE détaillée

```
Phase: test     #Nom de la phase de la règle
Input: Token    #Annotations prises en compte, ici juste Token
Options: control = all  #Option de repérage, on recherche ici toutes les formes possibles

Rule: CleanEggs       #Nom de la règle et début le RHS
(
 ( {Token.string == "dying"} |
  {Token.string == "Dying"} ):egg       #On catégorise ce groupe d'annotations sous egg
 ( {Token} )[0,8]
 ( {Token.string == "egg"}  |
  {Token.string == "eggs"} |
  {Token.string == "Egg"}  |
  {Token.string == "Eggs"} )
)         #Fin de la RHS
→         #Début de la LHS
{
 gate.AnnotationSet egg = (gate.AnnotationSet)bindings.get("egg");
 outputAS.removeAll(egg);
}         #Fin de la LHS
```

Annexe 3.2. Tableau récapitulatif des méthodes utilisées

Résultats des méthodes		
Méthode n°1	Méthode n°2	Méthode n°3
ISO 24978:2009	ISO 24978:2009	ISO 24978:2009
CEN/TS 16454:2013	CEN/TS 16454:2013	CEN/TS 16454:2013
CEN/TS 16454	CEN/TS 16454	CEN/TS 16454
NF EN 16062	NF EN 16062	NF EN 16062
ISO 24978	ISO 24978	ISO 24978
	ETSI/TS 122·101	ETSI/TS 122·101
	ETSI/TS 124·008	ETSI/TS 124·008
	ETSI/TS 126·268	ETSI/TS 126·268
	ETSI/TS 126·267	ETSI/TS 126·267
	ETSI/TS 122·011	ETSI/TS 122·011
	EN 15722	EN 15722
	EN 16072	EN 16072
	ETSI/TS 122·071	ETSI/TS 122·071
	EN 15722:2011	EN 15722:2011

	EN 16062:2011	EN 16062:2011
	EN·16062:2015	EN·16062:2015
	EN·16062:2011	EN·16062:2011
	EN 16062:2015	EN 16062:2015
	EN 15722:2015	EN 15722:2015
	EN 16102:2011	EN 16102:2011
	EN 16072:2011	EN 16072:2011
	CEN/TS 16454	CEN/TS 16454
	EN 16062	EN 16062
	EN 16102	EN 16102
	ETSI/TS 122 003	ETSI/TS 122 003
	ETSI/TS 126 269	ETSI/TS 126 269
	ETSI/TS 124·008	ETSI/TS 124·008
	ETSI/TS 127·007	ETSI/TS 127·007
	ETSI/TS 122·003	ETSI/TS 122·003
	ETSI/TS 122 071	ETSI/TS 122 071
	ETSI/TS 121 133	ETSI/TS 121 133
	ETSI/TS 151·010-1	ETSI/TS 151·010-1
	ETSI/TS 102 164	ETSI/TS 102 164
	ETSI/TS 127 007	ETSI/TS 127 007
	ETSI/TS 123·018	ETSI/TS 123·018
	ETSI/TS 122·001	ETSI/TS 122·001
	ETSI/TS 122·004	ETSI/TS 122·004
	ETSI/TS 122·002	ETSI/TS 122·002
	ETSI/TS 124·008	ETSI/TS 124·008
	ETSI/TS 126·269	ETSI/TS 126·269
	EN ISO 24978	EN ISO 24978
	ETSI/TS 134·123	ETSI/TS 134·123
	EN·ISO 24978	EN·ISO 24978
	ETSI/TS 124·123	ETSI/TS 124·123
	ETSI/TS 151·010	ETSI/TS 151·010

	ETSI/TS 102·164	ETSI/TS 102·164
		3GPP TS 23.078
		3GPP TS 23.079
		3GPP TS 24.008
		3GPP TS 25.413
		3GPP TS 48.008
		3GPP TS 29.002
		TS 101
		TS 22.101
		3GPP TS 23.07
		3GPP TS 23.08
		3GPP TS 23.09
		3GPP TS 27.005
		TIA 617
		3GPP TS 23.227
		3GPP TS 22.004
		3GPP TS 23.011
		3GPP TS 26.267
		TS 26.267
		TS 26.268
		TIA IS-99
		TIA IS-135
		3GPP TS 25.301
		3GPP TS 24.007
		3GPP TR 22.967
		3GPP TS 26.268
		3GPP TS 26.269
		3GPP TR 26.969
		3GPP TS 24.002
		3GPP TS 23.002
		3GPP TS 44.003

		WGS 84
		WGS 72
		WGS 66
		WGS 60
		GSM·850
		GSM·900
		3GPP·TS·34.123-1
		ETSI 121·133
		3GPP·TS·21.133
		GSM·450
		GSM·480
		GSM·710
		GSM·750
		GSM·810
		ISO/IEC 9646
		3GPP·TS·51.010-1
		ETSI 126·267
		3GPP·TS·22.101
		CN·13
		CEN/TC 278
		CE·91/263/CE
		P·99-258

Annexe 3.3. Catégoriser des références en références normatives

```
Rule: References_Normatives
Priority: 100
(
 ( {Candidate.type == "Reference_Cite", Candidate within References_Normatives} |
  {Candidate.type == "Reference_Cite", Candidate within ReferencesNormativesASTM} )
):ref_norm
-->
{
 gate.AnnotationSet norme = (gate.AnnotationSet) bindings.get("ref_norm");
 gate.Annotation normeAnn = (gate.Annotation) norme.iterator().next();
 FeatureMap normeFeatures = normeAnn.getFeatures();
 gate.FeatureMap features = Factory.newFeatureMap();
 features.putAll(normeFeatures);
 features.remove("type");
 features.put("type", "Reference_Normative");
 try{
  outputAS.add( norme.firstNode().getOffset(), norme.lastNode().getOffset(), "Candidate",
features);
 }catch(InvalidOffsetException e){
  throw new LuckyException(e);
 }
 inputAS.remove(normeAnn);
}
```

Annexe 3.4. Titres pour les références normatives

2. Références Normatives
3. Références Normatives
2. Normative References
3. Normative References
Normative References
3. References

Annexe 4. Exigences et pondération

Annexe 4.1. Pondération des formes verbales

Annexe 4.2. TempRequirement

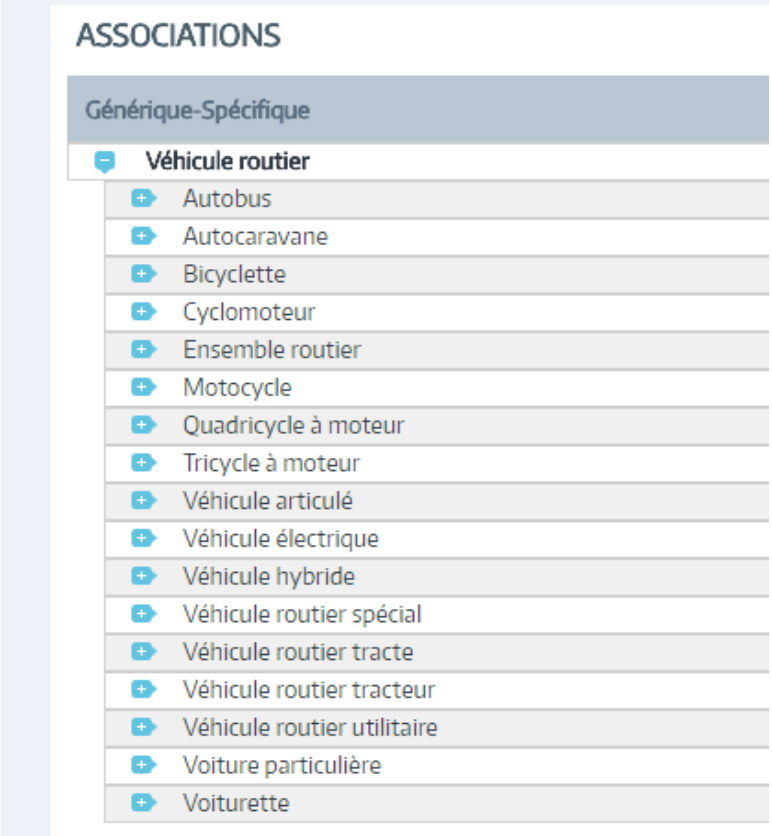
```
Rule: RequirementsClue1
(
 ( {Token.string ==~ "[Dd]oit"} |
  {Token.string == "doivent"} |
  ({Token.string == "ils"} {Token.string == "doivent"}) |
  {Token.string == "dois"} |
  ({Token.string == "il"} {Token.string == "doive"}) |
  ({Token.string == "ne"} {Token.string == "doive"} {Token.string == "pas"} ) |
  ({Token.string == "ne"} {Token.string == "doivent"} {Token.string == "pas"}))
):requirement
-->
 :requirement.TempRequirement = {type="Requirement", rule="RequirementsClue",
category="1", intercategory="1"}
```

Annexe 4.3. Requirement

```
Rule: Requirement
Priority:105
(
 {Sentence contains TempRequirement}
):isRequ
-->
{
 gate.AnnotationSet Requ = (gate.AnnotationSet) bindings.get("isRequ");
 gate.AnnotationSet tempRequ = inputAS.get("TempRequirement",
Requ.firstNode().getOffset(), Requ.lastNode().getOffset());
 gate.Annotation tempRequAnn = (gate.Annotation) tempRequ.iterator().next();
 FeatureMap tempRequFeatures = tempRequAnn.getFeatures();
 gate.FeatureMap features = Factory.newFeatureMap();
 features.putAll(tempRequFeatures);
 features.remove("rule");
 features.put("rule", "Requirement");
 try{
  outputAS.add( Requ.firstNode().getOffset(),
Requ.lastNode().getOffset(), "RequirementSentence", features);
 }catch(InvalidOffsetException e){
  throw new LuckyException(e);
 }
}
```

Annexe 5. Empreinte sémantique

Annexe 4.1. TIT et ses relations



On peut ici voir le terme Véhicule routier est parent d'une multitude d'autres termes au sein du thésaurus.