Extraction d'informations en contexte de Petites et Moyennes Entreprises

Mylène Joséphine

Extraction d'informations en contexte de Petites et Moyennes Entreprises

Nom : JOSEPHINE
Prénom Mylène

UFR LANGAGE, LETTRES ET ARTS DU SPECTACLE, INFORMATION ET COMMUNICATION (LLASIC)

Mémoire de master 2 professionnel - 20 crédits – Sciences du Langage

Spécialité ou Parcours : Industrie de la Langue parcours TALEP

Sous la direction de M. Olivier KRAIF

Année universitaire 2011-2012

Extraction d'informations en contexte de Petites et Moyennes Entreprises

Nom : JOSEPHINE
Prénom : Mylène

UFR LANGAGE, LETTRES ET ARTS DU SPECTACLE, INFORMATION ET COMMUNICATION (LLASIC)

Mémoire de master 2 professionnel - 20 crédits – Mention Sciences du Langage

Spécialité ou Parcours : Industrie de la Langue parcours TALEP

Sous la direction de M. Olivier KRAIF

Année universitaire 2011-2012

Dédicace

A vous papa, maman que j'aime,

A vous Sandra et Murielle, qui chacune à votre tour m’avez largement encouragée,

Et à mon professeur, M Jacques COURSIL.

Avant-propos, Préface, Avertissement

En 1998, je découvrais la linguistique informatique, pas le Traitement Automatique du Langage ou TAL. En faisant cette distinction, il ne s'agit pas pour moi de relancer le vieux débat de la différence entre deux disciplines, qui pour certains sont synonymes l'une de l'autre, mais de déclarer l'extraordinaire ouverture que la linguistique informatique m'a offerte. Ouverture sur le potentiel de l'industrie de la langue, spécialement perceptible à dans l'orientation donnée aux trois années de formation du DEUG de Mathématiques Informatique et Statistiques Appliqués aux Science Humaines, mineure Science du Langage puis celle de licence, ayant suivi cette première expérience,. Je me souviens des déclarations de mon professeur de l'époque, M. Jacques COURSIL, pour qui tout enseignement devait toujours être dispensé en synergie avec les autres champs de la connaissance. L'enseignement devait se faire à l'image des modalités d'apprentissage de l'être humain celui-ci absorbant essentiellement par proximités conceptuelles et par confluences des savoirs. Aussi notre formation à l'époque s'était-elle efforcée de décliner les connaissances linguistiques sous les différentes approches disponibles dans le champ de la connaissance humaine, des mathématiques, des statistiques et de l'informatique.

Dispensée de cette manière, la parenté de la discipline avec l'intelligence artificielle nous apparut de manière tout à fait évidente. Plus évidente que dans l'approche pragmatique posée par le TAL, qui se préoccupe avant tout de la chaîne linguistique et des interventions mécaniques à lui appliquer pour obtenir une fonction.

La sémantique, tête avancée du traitement automatique, représente la plus grande chance de réconciliation du TAL avec l'intelligence artificielle. A ce jour, elle avance des solutions en ordre dispersé, du type de celles du Web sémantique. Des solutions qui dans leur stratégie n'engagent pas forcément de descriptions des couches structurelles du langage dégagées par le lexique, la grammaire, la syntaxe et la sémantique et qui permettent de former un édifice autonome qui s'auto-entretiendrait. Cet horizon fait parfois figure de chimère, poursuivie encore par quelques chercheurs et industriels.

Cet horizon qui à mon sens est riche de promesses, je pense l'avoir aperçu. Cependant en dehors du monde universitaire, comment l'atteindre ? Seule l'industrie est

sans doute assez dynamique pour offrir les opportunités voulues pour jeter les ponts nécessaires. Mais encore faut-il lui avoir donné les gages suffisants de compétences, avoir fait le tour de ses réalisations et avoir une idée plus ou moins exacte de leur potentiel.

Dans cette construction de compétences et de savoirs qui est celle de tout candidat à cette tâche/horizon, il convient de faire la preuve d'acquis communs avec sa main-d'œuvre principale, les informaticiens, être à même de discuter avec ces pairs de façon éclairée, pour pouvoir leur laisser sans restrictions inutiles ni freins intempestifs, le plein exercice de leurs prérogatives - mais également à contrario, pouvoir de façon éclairée, refuser leurs raccourcis ou solutions que l'on sait inadaptés à l'objet qu'ils ont la charge de décrire informatiquement.

Dans cette optique, un stage qui me familiariserait avec les outils et concepts de tout bon informaticien, avec l'environnement fonctionnel d'une société, qui me pousserait à mettre en place mes solutions, en deux mots, qui m'exposerait serait une chance. Quoi de mieux, en effet, qu'un environnement où tout ou presque est à faire, où les échanges entre spécialités seraient non seulement nécessaires mais encouragés afin de favoriser les échanges de savoirs. Au terme de mon entretien, mon stage au sein de la société N5 promettait tout cela. Si ce stage a été porté par une configuration spécifique de l'époque (mars/juillet 2011), il fut également le résultat de la démarche active de son gérant et de son conseiller informatique.

Remerciements

Je tiens à remercier M. Olivier KRAIF mon maître de stage qui, en plus de sa contribution à notre formation de « taliste », nous a appris que le développement informatique était affaire de passion.

J'adresse mes remerciements,

A mon tuteur en entreprise, M. NAVELLOU qui par l'ensemble de son parcours témoigne qu'outre la formation, la réussite professionnelle est affaire d'audace et affaire d'homme ;

A M. Frédéric LASNIER pour sa confiance qui a fait de moi un pair éprouvé et ressenti.

UNIVERSITE Stendhal GRENOBLE 3

Déclaration anti-plagiat
Document **à scanner** après signature
et **à intégrer** au mémoire électronique

DECLARATION

1. Ce travail est le fruit d'un travail personnel et constitue un document original.
2. Je sais que prétendre être l'auteur d'un travail écrit par une autre personne est une pratique sévèrement sanctionnée par la loi.
3. Personne d'autre que moi n'a le droit de faire valoir ce travail, en totalité ou en partie, comme le sien.
4. Les propos repris mot à mot à d'autres auteurs figurent entre guillemets (citations).
5. Les écrits sur lesquels je m'appuie dans ce mémoire sont systématiquement référencés selon un système de renvoi bibliographique clair et précis.

NOM : JOSEPHINE PRENOM : Mylène

DATE : 19/07/13 SIGNATURE :

Mise à jour mars 2013

Sommaire

Introduction

Durant les deux années qu'a duré le master Industrie De la Langue (IDL), la spécialité TALEP (Traitement Automatique de la langue Ecrit et Parole) s'est efforcée de dresser un instantané du traitement automatique de la langue à destination de ses étudiants, identifiant à leur intention, les grands domaines couverts par la discipline, à travers les deux grandes divisions de l'écrit et de l'oral, leur fournissant les outils conceptuels, et plus marginalement, les matériaux avec lesquels réaliser leurs innovations. Elle leur a ouvert l'horizon de la pluridisciplinarité, les formant à la programmation informatique et aux grandes théories linguistiques. Elle identifia pour eux, les lieux d'enjeux de la discipline, au plan technique, au plan de l'innovation commerciale, de la création de besoins nouveaux à satisfaire. Se faisant, elle a également contribué à faire émerger chez eux la conscience du cadre idéalement adapté à leurs compétences et par contraste celle plus probable de leur intervention : celui de la petite ou moyenne entreprise, non spécialisée, cadre d'intervention le plus fréquent au regard du tissu entrepreneurial.

Le TAL ne fait pas encore réellement partie de la culture de l'entreprise, au même titre que l'informatique pure et dure ou que la gestion des ressources humaines, par exemple. Cependant, il peut se targuer d'une vitrine brillante, une fois que celle-ci est présentée sous ses noms usuels à une audience grand public, largement utilisatrice de ces produits. En effet les utilisateurs de traitement de textes comprennent bien mieux ce que sont des correcteurs orthographiques, correcteurs grammaticaux, traducteurs automatiques, voix de synthèse que « Traitement Automatique de la Langue ». Ils ignorent souvent l'existence de la spécialité à l'origine de ces réalisations. La configuration est quelque peu la même, une fois plongé en milieu professionnel. Ce dernier quand il approche le TAL, ne le connaît qu'au travers de produits clefs-en-main, parfois surdimensionnés au point de vue financier et des services offerts, une fois ramenés à la puissance financière et aux besoins de la majorité des PME (Petites et Moyennes Entreprises).

Réussir à toucher cette énorme réserve d'emplois, cette masse de développement sur mesure revient à répondre à la question suivante : - comment faire sa place à la spécialité face aux produits emblématiques de la discipline et face à la méconnaissance des professionnels ? Existe-t-il une place pour « l'artisanat », en la matière ?

Les difficultés pour la discipline à s'imposer sont rapidement sensibles. Le déficit de connaissances ou d'informations a pour corollaire immédiat d'engendrer des réticences

à y investir, à y engager des sommes de la trésorerie jugées vitales. Par ailleurs, si les commanditaires appréhendent mal le bénéfice que peut introduire le TAL dans leur gestion de l'entreprise, ils évaluent encore plus difficilement le niveau de complexité introduit par la manipulation de la chaine linguistique

En intégrant la s.a.r.l N5, représentante par excellence de ce monde de la PME vierge de culture TAL, c'est bien le défi de l'intégration d'une spécialité dans le cadre de l'entreprise que je relevai : - le défi d'identifier les besoins spécifiques de ces petites structures susceptibles de justifier ma contribution professionnelle à ce milieu : - le défi de répondre sans faille à son besoin de connaissances sur la linguistique informatique : - le défi de dessiner les contours d'une discipline qui plus que la description de la langue doit permettre de répondre aux besoins de descriptions systémiques et informatiques de n'importe quel commanditaire.

N5 illustre parfaitement l'ensemble de ces conditions. Cette société, en coïncidence avec ses besoins, a fait le choix de s'ouvrir à une discipline montante dont cependant elle ne connait pas vraiment les attendus. C'est une s.a.r.l qui s'inscrit à la jonction de l'artisanat et de l'industrialisation. Car si elle vise la taille de l'industrie, ses caractéristiques artisanales sont marquées. Ses besoins informatiques somme toute classiques offrent un champ d'expérimentation intéressant à un débutant du TAL. Installation d'un moteur de recherche, contribution au développement d'un *crawler*, codage en java, reconfiguration et enrichissement d'une base de données sont autant de champs potentiels d'intervention. Quant à la part du TAL proprement dit, elle est entièrement à construire dans cet environnement.

Entre projections et réalisations, quelles seront les formes réelles de cette intervention ? Elles dépendront certainement de la configuration de la société, de son organisation, du niveau de priorité qu'elle accordera à la satisfaction de ses différents besoins.

Partie 1

N5, un statut juridique au service d’un projet

Chapitre 1 – N5 et le projet AssurGroup

La société N5 est avant tout l'histoire d'un projet et une affaire de famille. Elle est née en 1995 sous la forme sociétaire de s.a.r.l, avec pour gérante Mme L Navellou, mère de M. Jean-Luc Navellou, directeur de la société et véritable initiateur du projet. C'est l'activité professionnelle de consultant en stratégie de ce dernier qui soutient la vie de l'entreprise N5. Jean-Luc Navellou en est le bayeur de fond et le commanditaire.

1.1. Présentation de la société N5

1.1.1. Son activité

La s.a.r.l N5 se définit aujourd'hui comme une société d'édition de sites internet et de vente de services. La société est dans cette phase particulière de préparation et de construction de son produit. Celle-ci mobilise toutes les capacités d'investissements de son directeur.

1.1.2. N5 pour quel produit ?

N5 se propose de fournir tout un ensemble d'outils de décision au monde des assurances et à ses affiliés. Ces outils consistent essentiellement en des ressources informatives structurées dont les liens les unes aux autres auront été capturées, analysés et exploités au maximum pour un résultat que la société espère des plus fins. C'est là qu'elle entend réaliser sa plus value, d'autant meilleure rapportée à la concurrence entretenue par les réseaux de chacune des sociétés d'assurance de la place et à la coopération réduite existant dans le secteur.

L'ambition de N5 à long terme est de répondre plus largement aux besoins de cette sphère professionnelle en lui assurant aussi bien la couverture de ses emplois que celle de ses besoins logistiques.

La première étape de cette construction ambitieuse est tributaire de la capacité du projet à exploiter les informations foisonnantes d'une ressource que l'on sait prolixe et facilement accessible, celle du Web. Cette partie du projet repose déjà sur un gigantesque effort de renseignement et de collecte sur les sociétés faisant la dynamique du secteur des assurances, au plan national comme international, réalisé par J.-L. Navellou. Ces efforts, en effet, ont abouti à l'édification d'un annuaire de sites comptant à ce jour plus de 15 000 références.

La stratégie de développement privilégiée par le directeur de la société fixe à cette dernière comme objectif premier, la réalisation d'un prototype informatique, à même de révéler le potentiel du produit final visé.

Illustration 1 Fiche technique de la société N5

Chiffre d'affaires 2009 : 200 K€

Forme juridique : SARL

Sirene : RCS NANTERRE B 351 405 279

Siège social : 5 avenue du Marechal Juin – 92100 Boulogne

Lieu d'activité : 47 rue Legendre Paris 17è

Actionnariat : le capital est réparti majoritairement entre Jean-Luc NAVELLOU, et Jacques et Herve NAVELLOU, ses frères.

Gérant : L. NAVELLOU.

Directeur : J-L NAVELLOU

Propriété Intellectuelle (Marques exploitées) :

· AssurWebR, AssurWatchR, AssurWiki R, AssurActuR, AssurVideoR,

AssurJobR, AssurMarketR, le portail de l'assuranceR

Activités : activité de conseil et édition de sites Web

1.2. Les ressources humaines de N5

1.2.1 La direction de N5

La direction de N5 est assurée par M. Jean-Luc Navellou, principal investisseur dans N5. Son activité professionnelle de conseiller en stratégie, en finance les besoins en développement informatique. En tant que conseiller en stratégie, il est détenteur d'un MBA délivré par l'université de Columbia, Etats-Unis. Il a travaillé avec de nombreux cabinets de conseil dont le cabinet SILTEA. Il fut directeur d'une équipe d'une quinzaine de personnes pour une société américaine dont seul le rachat a empêché sa promotion en tant qu'associé.

Dans le cadre de son activité, il s'est consacré cinq ans à l'information d'affaires, il a accompli de nombreuses missions pour de grandes mutuelles françaises (la MGEN notamment) et d'autres compagnies d'assurances privées, se familiarisant ainsi avec le

monde spécifique des assurances qui va lui inspirer le projet AssurGroup poursuivi avec N5.

1.2.2 Le personnel de N5

M. Navellou fut longtemps la seule ressource permanente de la société, assurant à la fois le rôle de commanditaire et de chef de projet de l'aventure. Quelques deux mois avant mon intégration au sein de l'entreprise, de nouvelles évolutions s'initiaient dans la dite entreprise avec l'arrivée d'un nouvel homme, aux acquis d'expériences importants et apportant les compétences techniques à même de formaliser 15 années de réalisations conceptuelles et quatre ans de développement informatique épars. Deux mois durant, sa fréquence d'intervention au sein de notre équipe fut d'une à deux journées par semaine jusqu'à après cette époque, devenir permanente comme nouvel associé de N5.

N5 est une société dont l'effectif variable dépasse rarement 3 personnes et dont une partie des ressources humaines ces dernières années furent des stagiaires.

1.3. Histoire de N5

1.3.1. L'idée de départ

L'idée d'AssurGroup naquit à l'époque où le minitel paraissait être encore une technologie prometteuse. Son évolution marquée vers une organisation thématique de l'information laissa entrevoir à J.-L. Navellou un mouvement général de sectorisation de l'information qui gagnerait certainement lui aussi un domaine d'activité qu'il connaissait bien, les assurances. Des services étaient à proposer à ses usagers comme à ses acteurs.

Armé de cette certitude, il s'engagea dans le rachat d'une société détenue par France Télécom que celle-ci s'avéra incapable de faire fructifier. L'avènement du Web mondial et le succès technique mais aussi financier remporté par le moteur de recherche du nouvel acteur Google l'incita à opérer un transfert de technologie de son support d'informations, du minitel vers Internet. La marque AssurWeb fut ainsi déposée en 1996 et le premier annuaire de la société mis en ligne en 1997, au début en tant que simple liste de 500 références de site.

En 1998, le site devenu multilingue, prit alors en charge les références anglaises, puis, plus largement, européennes, allemandes, espagnoles, portugaises, italiennes et suédoises et enfin chinoises.

1.3.2. Une idée qui s'enrichit

AssurWeb ne fut pas le seul axe de développement de N5. L'idée de départ s'étoffa et devint pléthorique. Ainsi à ce jour, l'outil que nous nommerons de manière généraliste AssurGroup, s'enrichit de 10 pôles de services, liés, toujours, aux assurances.

- AssurActu dont l'objet est de concentrer et de structurer des flux d'informations consacrées à l'actualité journalière, en temps réel de la sphère d'activité et de tous les domaines dont elle dépend.
- AssurBooks qui centralise les parutions majeures proposant une formalisation du domaine. Elle s'associe à la branche d'activité consacrée à la formation des ressources humaines du métier, souhaitée par N5 et cristallisée par AssurLearning. Celle-ci envisage la création d'une plateforme d'apprentissage d'e-learning.
- AssurJob est la plateforme du projet AssurGroup consacrée au recrutement des personnels de l'assurance et de manière générale, à la mise en relation des demandeurs d'emplois et des chasseurs de tête des différentes directions des ressources humaines clientes du groupe.
- AssurMarket, volet du projet formalisé en 1997 et financé en 2000, est dédié à la commercialisation des produits du groupe.
- AssurWiki est pensé comme l'équivalent sectoriel d'un Wikipédia. Il doit s'ouvrir aux équipes de documentalistes des sociétés utilisatrices des services d'AssurGroup.
- AssurVide. Dans le flux d'informations dépouillées du Web, ce pôle identifie les flux vidéos en circulation et les classe. Ils doivent alimenter la réflexion des stratégies marketing des sociétés clientes du réseau AssurGroup
- AssurLex, défini comme un outil de veille sur le monde des assurances, doit permettre de surveiller la sortie des lois impactant les secteurs de l'assurance et de la mutualité. Il ciblera essentiellement les changements de réglementation, l'apparition de nouveaux concepts métier, l'apparition de nouvelles normes portées par la législation. Idéalement, il vise jusqu'à la détection des impacts législatifs induits à travers ses domaines connexes.
- AssurWords répertorie le lexique en usage dans le secteur des assurances et

parmi ses actuaires[1] afin de le mettre à disposition d'un public double, les assurés et les futurs assurés qu'il contribuera à guider et les compagnies d'assurance.

- AssurStocks correspond au volet d'AssurGroup qui s'occupe du suivi des marchés.

Si le Web est une ressource assumée du groupe d'activités, il n'en constitue pas la seule. Des partenariats stratégiques, pour certains déjà négociés, les autres appartenant à un horizon calculé, doivent faire la plus-value du futur portail.

Illustration 2 : Distribution des pôles d'activité d'AssurGroup en 2005

Chapitre 2 – AssurWeb vitrine du projet AssurGroup

Un large travail conceptuel définit l'ensemble des pôles d'activités d'AssurGroup. Cependant seul AssurWeb a fait jusqu'ici l'objet d'une réalisation informatique.

[1] Voir lexique page 128

2.1. Le projet AssurWeb

2.1.1. Présentation

Comme cela a été dit précédemment AssurWeb est cette partie du projet AssurGroup qui, exploitant les ressources du Web, propose à ses utilisateurs un retour d'informations spécifiques sur les activités et produits d'assurances : liste d'assureurs, de produits, informations presse, ressources multimédia accessibles, spécifiques d'une enseigne ou non, vocabulaires spécialisés, le tout à partir d'un site web. Assurweb propose à la fois des services de consultation multilingues, accessibles à des utilisateurs privilégiés ou anonymes, avec des différences évidentes de prestation, et un outil de type back office, un outil d'enrichissement dédié aux services de documentation de ses compagnies clientes ou partenaires.

Les informations réunies par le site vont des plus générales aux plus pointues: noms de compagnies d'assurance, adresses, produits proposés, jusqu'aux résultats d'exercices, répartitions de capital, bilans annuels, etc…

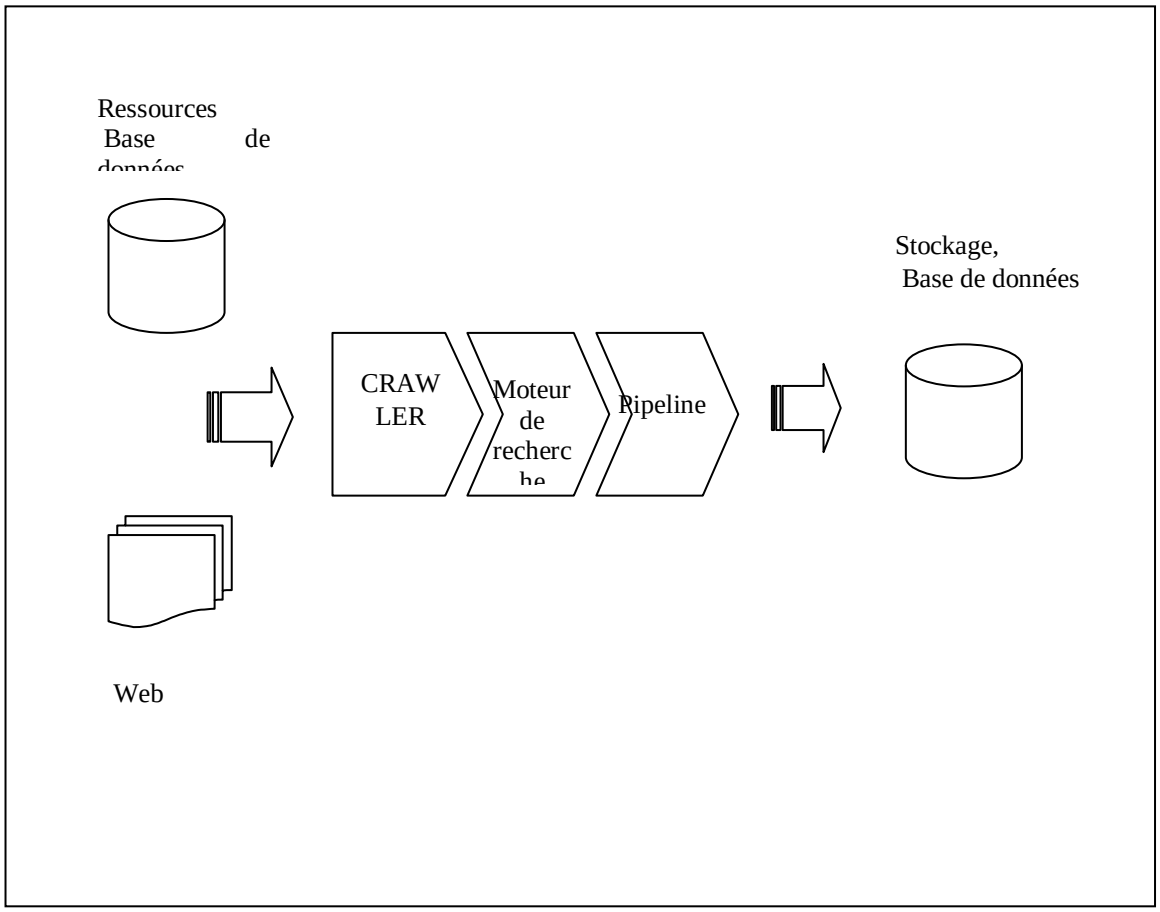
Exploitant le modèle de liens sponsorisés mis à l'honneur par Google, la proposition de service d'AssurWeb repose sur la rétribution des clics effectués par les utilisateurs du site.

2.1.2. Contours techniques d'AssurWeb

Le projet AssurWeb repose sur trois modules techniques :

- Un *crawler* entièrement paramétrable qui sillonne le Web à des fins d'extraction d'informations.
- Un moteur de recherche qui gère la restitution des informations pertinentes. Celles-ci sont, dans les grandes lignes, de deux types : des informations publiques, largement généralistes, et des informations privées et par conséquent réservées aux utilisateurs inscrits du site.
- Une chaîne de traitements ou pipeline qui regroupe l'ensemble des traitements applicables et appliqués aux pages fournies par le *crawler*. Cette chaîne consigne aussi bien les conditions d'acceptation et de rejet des pages ramenées, que l'ensemble des informations qui seront indexées par le moteur de recherche ou stockées dans la base de données du projet.

Illustration 3 : Diagramme métier du projet AssurWeb



2.2. Contexte immédiat et réalisations en cours

Trois tâches occupent le passé récent de l'entreprise, ainsi qu'une partie de son activité en cours, en matière de développement informatique:

- le développement inachevé du *crawler* réalisé en java par un ancien salarié,
- la refonte de l'interface web du site d'AssurWeb initiée par un étudiant stagiaire en informatique.

Coté interface publique, celle-ci implique :

- la refonte de l'ergonomie de l'interface ;
- l'ajout de nouvelles fonctionnalités à l'interface.

En back office, elle suppose la mise en place des outils participatifs prévus à l'intention des services spécialisés des sociétés adhérentes et autres centres de documentation.

L'outil sur lequel s'appuie cette mise à niveau est un framework PHP *open source*, dédié au développement de sites et d'applications web, *Symfony*[2]. Celle-ci devait apporter la garantie de la bonne structuration du développement html et des applications appelées à tourner en arrière plan des pages web.

2.3. Contributions attendues du TAL

Assurweb étant largement en re-construction ou parfois tout simplement en construction, mes interventions se devaient d'être en dépendance fonctionnelle réduite avec les autres phases du projet.

Un premier rôle dicté par ma spécialisation imposa d'emblée ma participation à la construction du *pipeline*, le module du projet travaillant à l'alimentation du système. Celui-ci ayant la charge de rapatrier les informations jugées pertinentes du Web, de les indexer pour les redistribuer à volonté vers les futurs utilisateurs du site selon leur type d'inscription. Les besoins de l'entreprise en développement informatique classique étant conséquents, mes compétences étaient requises aussi bien sur la rédaction du *crawler* que sur la mise en place de la base de données du projet. Enfin, un rôle prescriptif de conseiller en TAL était requis pour les développements ultérieurs du projet au regard de ses ambitions.

Chapitre 3 – Conditions matérielles et fonctionnelles du stage

3.1. Cadre matériel d'intervention

Le cadre matériel de mon intervention fut le suivant :

- Un emplacement dans un espace de travail ouvert et pouvant accueillir de 3 à 4 personnes.
- Un ordinateur portable d'occasion qui fut changé trois fois, les circonstances l'ayant décidé :
    - le premier fut mon portable personnel tombé en panne après moins d'un mois d'utilisation ;

[2] Voir Lexique page 131.

- le deuxième fournit par l'entreprise fut un portable de 13 pouces équipé du système d'exploitation Windows vista, utilisé avec un écran plat de 22 pouces ;
- le troisième, un portable professionnel Dell âgé de 5-6 ans et équipé d'un écran de 15,4 pouces.

- Trois systèmes d'exploitation de type Windows furent par conséquent utilisés, dans l'ordre: Windows 7, Windows Vista puis Windows XP.

Les ressources logicielles de développement consistèrent en :

- Un éditeur, Ultra-Edit ;
- deux kits de développement perl, successivement Strawberry Perl et Active Perl.

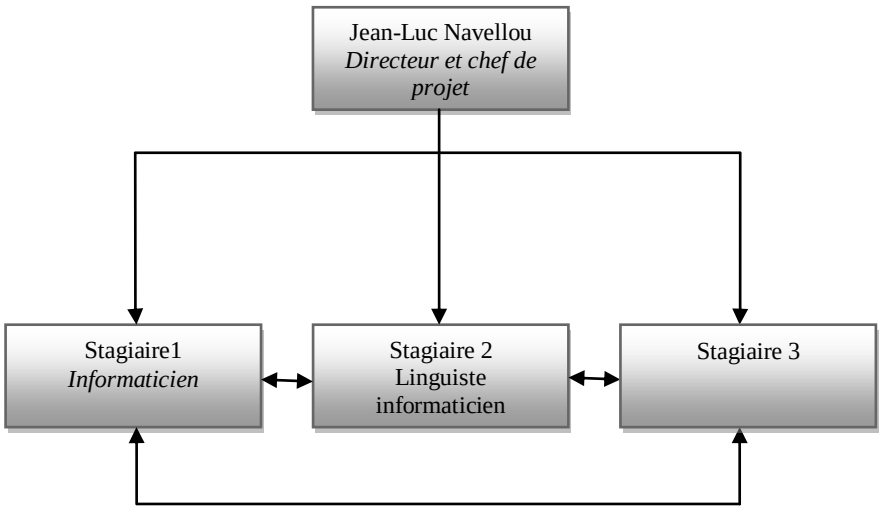
Aucun outil dédié à la linguistique informatique ne figurait au fond logiciel de l'entreprise, ni ne fut par la suite requis.

3.2. Organigramme et chaîne de décision de N5

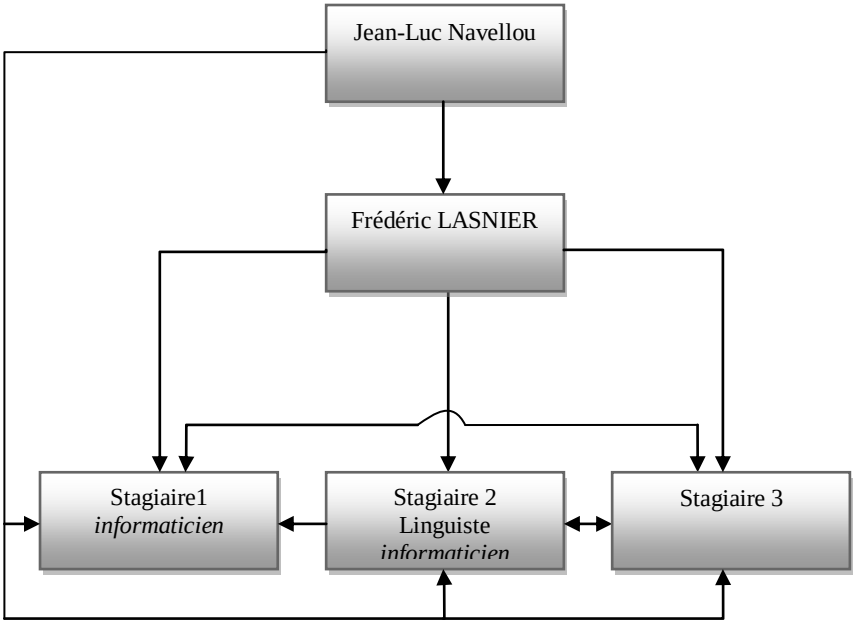
L'organisation fonctionnelle de N5 est celle d'une petite structure et par conséquent est très simple. A sa tête, le directeur de la société M. Jean-Luc Navellou tient à la fois le rôle de commanditaire et de chef de projet de fait, c'est lui qui distribue les tâches à accomplir et décide de leur niveau de priorité. Il garde la mémoire des documents produits au sein de l'entreprise. Il en est la mémoire technique puisque l'essentiel de la communication technique de l'entreprise était à mon arrivée essentiellement orale. Il est servi en cela par une mémoire formidable, desservi par ailleurs par son absence de formation informatique.

L'organisation que je viens de décrire subit des changements sensibles en cours d'exercice puisque Jean-Luc Navellou fit l'embauche d'un nouveau talent.

Illustration 4 : Organigramme de N5



Configuration en début de stage



Configuration depuis le 01/09/2011

Partie 2

Temps forts du stage

Chapitre 4 – Missions et distribution des tâches

4.1. Vue d'ensemble des missions

Cinq temps forts articulent cette période de stage. Ils se répartissent entre période de découverte de l'entreprise, missions commandées par le directeur de la société et d'autres missions qui n'aboutirent pas mais dont le potentiel d'expérience fut loin d'être moindre.

Le premier temps fort se confondit avec la période de découverte de l'entreprise, période pendant laquelle les explications sur la naissance de celle-ci, son ambition générale, les projets qu'elle avait en cours ainsi que leur état d'avancement furent dispensés. Dans mon cas, cette phase de découverte fut largement accompagnée par mon tuteur. Elle s'avéra d'autant plus importante qu'elle déboucha sur ma première mission de stage, la réalisation d'un document de conseil sur les contributions possibles du TAL à l'ensemble du projet AssurGroup.

Mon premier ordre de mission ouvrit une autre phase du stage. Il commandait la réalisation d'un instantané aussi exhaustif que possible des spécialisations du TAL et de ses produits. La sélection et le test de certains de ces outils dans la perspective minimaliste de construire une chaîne de traitement des entrées linguistiques devait terminer cette phase de recherche.

Un autre de mes ordres de mission consista en la réalisation d'un prototype d'extraction sur des informations ciblées : les noms de personnes morales, leurs adresses physiques et ou postales, les informations définissant des articles de presse, le lexique spécialisé de ce domaine de connaissances, entrées lexicales et définitions. Cette dernière tâche fut le point d'orgue de cette période.

Entre ses principales missions, j'eus encore d'autres missions qui connurent divers degré d'achèvement. En effet, mes deux premiers mois de stage, j'eus un ordre de mission différent quasiment toutes les semaines et demies. Seule leur faisabilité ou leur-à propos permit aux missions citées précédemment de prévaloir. Deux d'entre elles méritent cependant d'être mises en lumière, l'achèvement du *crawler* d'AssurWeb et la création de sa base de données.

Quelles tâches furent accomplies pour qu'elle mission ? C'est la vision que se propose de brosser les lignes suivantes.

4.2. Tâches et missions majeures

4.2.1. Découverte de l'entreprise.

Pendant cette phase du stage, il s'agit avant tout d'écouter et de prendre des notes. Très tôt, je m'y préparai en me dotant d'un cahier de bord. Mon encadrant M. Navellou me fit appréhender le projet AssurGroup depuis son idée surgie 15 ans plus tôt jusqu'à la date du 4 juillet 2011, début du stage. Il en évoqua les coups d'arrêt, les mutations précipitées par la montée en force de la technologie nouvelle qui devait définitivement supplanter le minitel, Internet.

Ainsi les tâches de cette époque consistèrent à :

- Ecouter. Beaucoup.

- Prendre des notes sur l'histoire de l'entreprise, sur les différents chantiers qu'elle avait engagés informatiquement, et leur niveau d'achèvement. Cette dernière information fut plutôt difficile à obtenir et demeura trouble suivant les domaines abordés. Les données du contexte concurrentiel de l'entreprise, les données quant au public auquel elle entendait s'adresser, les concepts métiers du domaine avec lequel elle traitait, l'histoire des assurances et de la mutualité, les buts de l'entreprise, les réalisations à son actif m'ont été exposées de manière soutenue pendant les 15 premiers jours de mon embauche et en filigrane le reste du temps.

- Mettre à niveau mon environnement logiciel afin de l'adapter au mode communication en vigueur au sein de l'entreprise mais aussi pour exploiter l'énorme documentation formant les acquis de N5.

4.2.2. Mise à jour logicielle

Elle répond au besoin de s'adapter au modèle de communication en place au sein de l'entreprise. Elle signifie concrètement être capable d'en exploiter les ressources, et mieux encore de pouvoir les étendre. Dans mon stage, cette phase s'étala sur 2 jours et demi. Elle se déroula parallèlement à la découverte de l'entreprise et de l'ensemble de ses chantiers achevés et inachevés. Elle fut supervisée par J.-L. Navellou.

L'installation logicielle vit la mise en place :

- d'un système d'exploitation : Windows 7 ;
- d'outils de communication interne :

- Skype ;
- d'un espace collaboratif via le service web, Dropbox

- d'outils de formalisation de type bureautique
  - MindjetManager, pour une réflexion en cartes conceptuelles ou maps ;
  - Microsoft Office (Excel, Word) pour la bureautique traditionnelle ;
  - Microsoft Office Visio pro pour la création de documents techniques, la création de schémas ;
  - Dia qui rend le même type de services que l'outil précédent mais dans l'*open source* et à une échelle moins intéressante.

4.2.3. La réalisation du document de conseil TAL

Conformément aux prévisions de mon entretien avec M.M Navellou et Lasnier. M. Navellou me demanda d'identifier les contributions possibles du TAL à l'ensemble du projet AssurGroup, en identifiant les ressources informatiques, linguistiques et même humaines à mettre en place. Il fut construit comme un document-conseil, une forme très parlante pour M. Navellou puisque recourant à des concepts métiers qui lui étaient familiers. Les contours restés flous d'une discipline qu'il avait pourtant déjà rencontrée lui étaient désormais plus accessibles.

4.2.4. Inventaire des outils du TAL

Cette mission consista à lister les logiciels de TAL de l'écrit, en fonction de :

- leur couverture OS (Windows, systèmes Unix) ;
- de leur « gratuité » ou de leur caractère payant (*open source*, opportunités de test avant achat) ;
- de leur couverture linguistique ;
- des grands domaines de développement du traitement automatique mis en œuvre : traduction, résumé, éditeur de réseaux sémantique, réseaux sémantiques formalisés, analyseurs morphologies et/ou syntaxiques, simples étiqueteurs, etc…

Le but premier d'un tel document, outre de consigner une connaissance à un instant T, était de rendre cette connaissance facilement accessible pour toute ressource de l'entreprise. La forme informative adoptée à cet effet fut celle d'une « map » ou

représentation en disjonctions[3]. Très visuelle, elle propose des niveaux de détails progressifs, autorisant ainsi plusieurs niveaux de lecture et d'analyse, allant des plus généralistes aux plus pointus.

Des contraintes rarement évoquées pèsent sur les résultats offerts par ce type de travaux et les orientent grandement. Elles résident dans la possibilité effective que l'on a de tester les applications retenues. Le shareware n'est pas toujours accessible dans le domaine particulier du TAL ou de la linguistique informatique et moins encore pour les produits développés en France ou pour le français. S'il est pratique, le shareware revient en vérité à confier son code à un tiers sans garantie de retour financier sur celui-ci et plus grave encore, à voir grandir le risque de se le faire « voler ».

Quelques outils *open source* ont été rapatriés durant cette phase d'intervention. Des outils *open source*, essentiellement issus du monde anglophone. Reposant sur des modèles statistiques, ils sont naturellement optimisés pour l'anglais. Le monde universitaire français ou francophone propose également ses produits : des universitaires ayant fondé leur société afin de commercialiser leurs réalisations, à titre personnel ou au bénéfice de leur université de rattachement. Cependant, ils sont chers, de 1 000 euros à 3 000, voire 5 000 euros, selon qu'ils s'appliquent à des fins de recherche ou à des fins commerciales. Obtenir de les tester ne pouvait s'obtenir qu'au terme de négociations entamant largement la durée de ce stage.

4.2.5. Sélections et tests les logiciels

Les possibilités financières de l'entreprise, les choix de développement faits par le commanditaire (développement de certaines fonctions, privilégier les besoins pouvant être satisfaits dans l'immédiat et à moindre coût plutôt que d'autres plus onéreuses et plus coûteuses en temps) ont donc largement influencé cette sélection. Cependant concernant les outils retenus, le travail de test consista à :

- collecter des informations sur les procédures d'installation. : pré-requis, dépendances annoncées, dépendances réelles, dépendances logicielles ou en ressources de type dictionnaires par exemple). De fait, cette étape revient le plus souvent à consulter les fichiers INSTALL plus que sibyllins, à retrouver l'expérience d'autres utilisateurs ou les informations mises en ligne par les

[3] Voir représentation utilisée en annexe 2, 8 et 9

concepteurs de ces programmes. Ce travail devait être fait pour chacun des environnements Windows et linux quand l'outil existait sur les deux plateformes ;

- installer une machine virtuelle pour l'exécution des programmes disponibles uniquement sur systèmes Linux (version Fédora) ;

  installer les logiciels disponibles au test soit sur Windows soit sur la machine virtuelle sur Linux ;

- consigner la procédure suivie lors de l'installation dans un document à destination des futurs utilisateurs de ces ressources : décrire les difficultés rencontrées, la manière de les lever, le type de résultat à attendre.

4.2.6. Réalisation de modules d'extraction

4.2.6.1. Objectifs

Les trois derniers mois de travail consacrés à l'accomplissement de cette dernière tâche avait pour objectifs de :réaliser l'extraction des glossaires disponibles sur les pages crawlées, dans un premier temps en lien avec les 15 000 références évoquées précédemment, puis sur les pages ramenées du Web en général ;

- réaliser l'extraction de noms de produit présents sur ce même ensemble ;
- réaliser l'extraction d'informations sur les articles présents dans ces pages : les titres, la localisation des pleins contenus des articles, ainsi que leur url et domaine.
- réaliser l'extraction des contacts présents dans ces pages. Un contact étant une organisation, des coordonnées postales, web et mail éventuellement ou physique ;
- générer la documentation accompagnant le développement :
  - documentation *html* permettant la réutilisation des ressources créées par un concepteur de projet, un programmeur ou toute autre personne ;
  - la documentation technique nécessaire à l'intégration ultérieure du projet.

La langue traitée dans ces extractions était le français. Pourtant autant qu'il était possible, les autres langues que sont l'anglais, le portugais, l'espagnol, l'italien, l'allemand, le suédois et le chinois étaient visées.

4.2.6.2. Développement des modules d'extraction

La réalisation de cette mission a requis :

- de concevoir une grammaire capable de capturer l'information désirée ;
- d'installer les interpréteurs perl des deux distributions pour Windows, Strawberry Perl à l'origine puis Active Perl, en raison du plus grand nombre d'outils disponibles dans cette suite ;
- de réaliser l'implémentation du système en langage perl ;
- de commenter le code ;
- de créer des exécutables pour Windows à partir des codes perl ;
- de réaliser le synopsis du module d'extraction ;
- de réaliser un document détaillant les fonctions assurées par le module, leur couverture linguistique, au regard de l'objectif fixé par l'entreprise de couvrir les ressources documentaires disponibles en huit langues. Il faut également présenter cette couverture au niveau des fonctions élémentaires du code ;
- de réaliser la documentation technique des unités d'extraction (listage des fonctions, des bibliothèques, dépendances fonctionnelles et dépendances par bibliothèque) ;

4.2.6.3. Les outils de développement utilisés

Les outils de ce développement furent les interpréteurs et packages suivants :

- Active Perl ;
- StrawberryPerl ;
- des outils de génération automatique de documentation furent expérimentés tel Doxygen qui a la particularité d'autoriser la création automatique d'une documentation technique interactive à partir, notamment, des commentaires du code d'un certain nombre de langages et par l'utilisation d'un système spécifique d'annotations. Cet outil ne fonctionne pas nativement pour perl. Il suppose le recours à un filtre, Doxygen filter, pour parvenir à un certain niveau de compatibilité. Les résultats sont très aléatoires et incomplets, au regard de la puissance véritable de l'application. Utilisé en complément avec GraphViz,

Doxygen permet la traduction automatique du projet en organigramme, créant une documentation technique de projet hautement interactive. Par clic sur les représentations graphiques, il est possible d'accéder à un niveau de détail approfondi de la sélection.

Dans la création de cette documentation de projet, d'autres outils et d'autres formats, furent utilisés : le module POD, outil naturellement adossé à Perl pour cette tâche et celui très simple du RST.

Un des efforts pour formaliser l'environnement de travail des futurs recrues mais également pour organiser la communication sur le projet en construction supposait la mise en place d'un WIKI. Le format de ses commentaires étant le RST, une partie de la documentation créée fut reprise pour partie dans ce format.

4.2.7. Réalisation de documentations techniques

La documentation technique, incontournable des tâches de développement, fit partie des exercices que je refis souvent. Faite avant l'achèvement d'une version du développement, elle était encore à réécrire au fil de ses mutations. La seule explication étant que les habitudes passées de certains personnels engendrait en réaction, une vigilance accrue, un peu exagérée selon moi, sur ces phases de rédaction. La peur que le travail fourni reste sans commentaires était en effet bien réelle.

La documentation des modules consista dans sa forme définitive, en une feuille de route expliquant la stratégie globale suivie par les solutions d'extraction, d'un diagramme métier exposant entrées et sorties avancées par le module, d'un diagramme fonctionnel simple, d'un digramme fonctionnel détaillé et d'un diagramme physique.

4.2.7.1. Le diagramme métier

Entre dans la documentation technique quoiqu'il soit lui-même peu technique en vérité. C'est essentiellement un document de communication, largement orienté vers le commanditaire, le non spécialiste mais qui a pour valeur indispensable d'identifier les fonctionnalités portés par un développement.

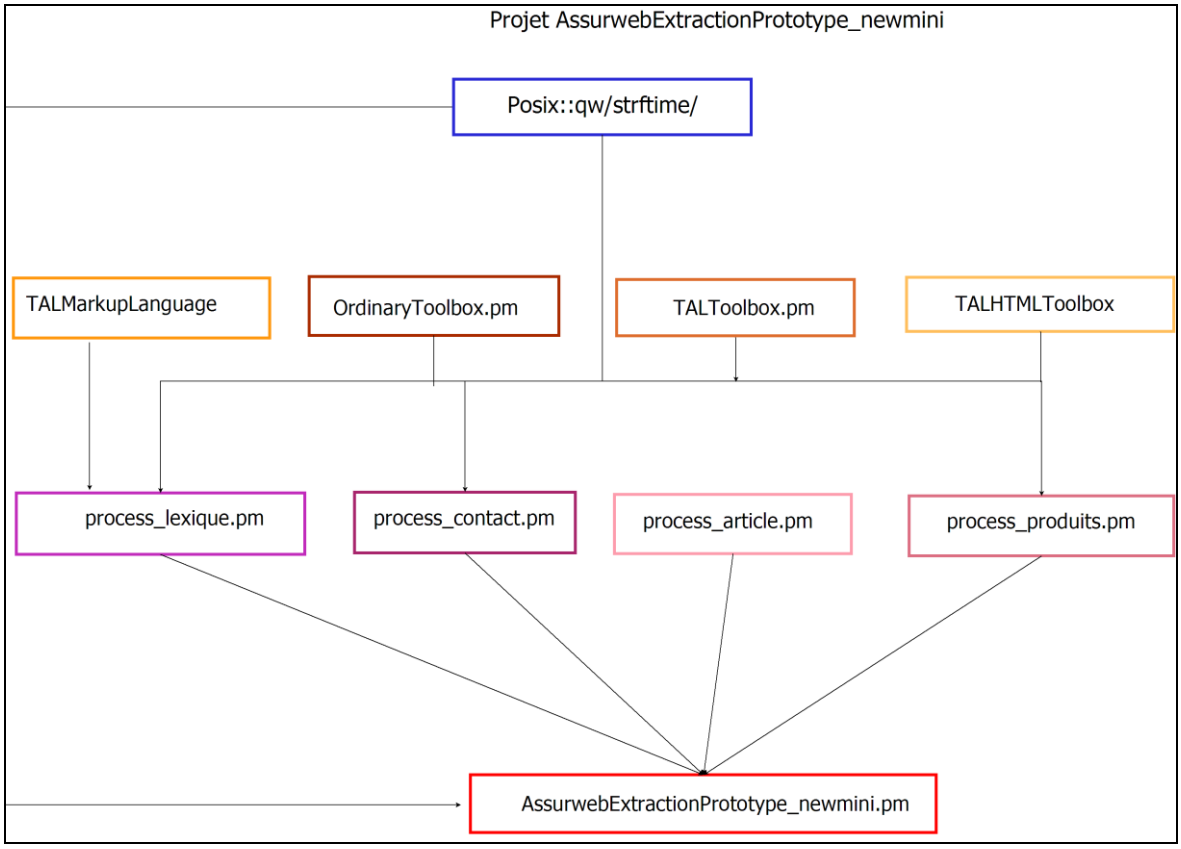
4.2.7.2. Le diagramme fonctionnel et le diagramme fonctionnel détaillé4

De manière générale, c'est un document qui rend compte visuellement et très précisément des données d'un développement, en termes d'entrées et sorties, d'enchaînements de processus et de systèmes dépendants. Dépendance du développement par rapport à d'autres sections du projet et avec lesquelles il a parties liées

La particularité du diagramme fonctionnel simple est de proposer cette description à une échelle relativement générale tandis que le diagramme détaillé la porte jusqu'à identifier exactement et en cascades les processus en œuvre. Ainsi à tout moment, le travail de développement peut-il se lire en termes de contraintes (allégeances supérieures) implications et imbrications. De tels documents rendent possible les opérations de maintenance, de modification ou de remaniement du code, sans oubli d'une variable, d'une fonction, d'une bibliothèque, d'un processus.

4.2.7.3. Le diagramme en dépendances

Illustration 5 : Diagramme en dépendances du prototype d'extraction à la fin novembre 2011



Commentaire 1 : Le diagramme en dépendances rend compte de l'architecture sur laquelle repose le code créé. En prolongeant le digramme fonctionnel en amont, il fait l'inventaire des différentes bibliothèques, des

[4] Voir annexes 3 et 4.

liens qu'elles ont les une aux autres et leur participation aux principales fonctions. Il répertorie bien les créations/extensions d'outils que les réutilisations de bibliothèques. Celles des packs, par exemple.

4.2.7.4. Le diagramme physique

Se présente comme un inventaire exhaustif des contenus des différentes bibliothèques créées par le développeur. Le diagramme liste les fonctions, les procédures, les constantes, les variables globales, les variables locales en terme de types clairement identifiés et exposés, de qualité d'entrées ou de sorties. Sa lecture prolonge le diagramme fonctionnel détaillé.

4.3. Tâches récurrentes

Si les lignes précédentes font état de mes principales missions, elles ne les résument pas toutes cependant. Ces dernières se répartissent encore entre tâches récurrentes croisant l'ensemble des missions évoquées précédemment et des tâches achevées prématurément.

4.3.1. Création de documents de communication

Une de mes tâches fut de procéder à la communication orale de ces différents travaux, lors de réunions-bilan avec le commanditaire et le chef de projet. Dans leurs phases intermédiaires et terminales: recherches en phase intermédiaire, collectes de données, cours accélérés sur le TAL. Les informations obtenues étaient communiquées non seulement aux responsables mais à l'ensemble du personnel qu'elles impactaient. Elles furent plus d'une fois l'occasion de revoir la pertinence de certaines demandes communiquées par le donneur d'ordres, M. Navellou. Ces réunions permirent par l'exposé vulgarisé qu'elles proposaient, d'appréhender le niveau de complexité des opérations engagées par le traitement d'une chaîne linguistique par exemple. Et par conséquent, d'évaluer le poids réel de l'investissement du projet en TAL. Mieux encore, de penser son échelonnement. Ce genre de résultats s'obtient notamment par l'inventaire chronologique des fonctionnalités[5] et services requis par la réalisation d'une tâche. Une présentation succincte pose les contours de chacune de ces étapes accompagnées par la présentation de leur coût global, matériel et temporel. Ces documents ont pour effet de créer une conscience partagée suffisante entre des personnes venant d'horizons divers.

Un des enseignements reçus à l'université me servit particulièrement de fil rouge durant tout le stage effectué à N5. Il y existait en effet une forte tendance à entretenir les

[5] Exemples annexes 5 et 6

mêmes discussions. Discussions au terme desquelles, croyais-je, chacun de nous nous étions entendus. Ces discussions ne m'impliquaient pas particulièrement face au donneur d'ordre. Elles avaient toujours les mêmes résultats. Ces décisions n'étant jamais consignées nulle part, celui dont les choix n'avaient pas prévalu y revenait toujours. Une de mes tâches consista donc à consigner ses décisions dans des documents de planification, ou bien des dans des dossiers de suivi : collection de mail, ou de documents proposant un début de formalisation qui prouvaient que la question avait été tranchée. L'écrit, en effet, n'ayant pas pour propriété de s'envoler. Ces discussions devinrent de moins en moins nombreuses. Elles ne se situèrent plus qu'au niveau de la direction, témoins qu'une entente sur les réalisations du projet Assurweb était encore à atteindre.

4.4. Missions avortées ou suspendues:

Les missions qui furent entamées et pour différentes raisons qui ne furent pas menées à terme ne furent pas moins consommatrices de temps que les autres. Elles furent elles aussi initiatrices d'enseignements.

4.4.1. Conception et développement java : le *crawler* AssurWeb

Une de ces missions programmées, entamées et restées inachevées consistait à terminer la conception et l'implémentation du *crawler* situé au cœur du projet AssurWeb en langage java. Celle-ci commença par des tâches de relecture et de documentation du code du *crawler*. Cette phase dura peu de temps. Ce code n'avait connu aucun entretien. Aucune documentation n'avait été tenue jusque là, ni dans le fil du code, ni dans un quelconque document directeur ou documentation technique, même incomplète. Le code était entièrement à se réapproprier.

Les séances de relecture du code s'organisèrent de la sorte :

- A partir de la fonction « main », repérer les bibliothèques appelées, en faisant la distinction entre les bibliothèques natives et celles créées par le salarié auteur du code.
- Identifier l'enchaînement des processus en fonction de leurs entrées et sorties, leurs objets d'appel, interfaces de rattachement, leurs actions.

Le but à terme étant :

1 - de créer la documentation technique pour le code existant

2 - d'exhumer la stratégie de résolution sous-tendue par le développement

Le développement mis en œuvre était loin d'être inintéressant mais son degré d'achèvement et l'investissement en temps (relecture, développement futur) fut jugé inadéquat par le chef de projet au regard des outils qui existaient dans le domaine. Mener les tâches mentionnées plus haut représentait un surcoût inacceptable dans la configuration de l'époque. Elles furent finalement laissées de côté après quelques séances de travail en commun.

4.4.2. Mise en place et configuration du serveur de recherche *SOLR*[6]

La mise en place du serveur *SOLR* que s'était choisi l'entreprise et sur laquelle s'était conclue, en partie, mon embauche me fut enlevée quant à elle après deux semaines à me familiariser avec le serveur *SOLR* et son index et moteur de recherche Lucene.

SOLR est un serveur de recherche. Il peut fonctionner de manière tout à fait autonome puisque disposant de son propre moteur de recherche. Il accepte aussi le couplage avec d'autres moteurs que son moteur natif. C'est là qu'intervient Lucene[7] dans le choix de l'entreprise de la gestion de cette étape.

Les premières tâches de découverte consistèrent à réunir les informations pratiques nécessaires à l'installation, la configuration et au travail de l'application. En bref cela signifia par exemple découvrir des serveurs comme Jetty qui permettent le déploiement des applications et servlets java sans recours à un serveur physique dédié (ce qui s'avérait nécessaire dans mon cas puisqu'aucun accès au serveur ne m'était aménagé et surtout n'était possible) et qui contiennent des exemples de configuration pour *SOLR*.

SOLR est un objet java qui connait deux modes possibles d'utilisation : une utilisation classique, du type clef en main qui suppose la réutilisation pure et simple de la bibliothèque *SOLR* et une autre qui permet de personnaliser son code, en en redéfinissant les propriétés, les comportements, en fonctions des buts de la configuration particulière d'un projet.

Son potentiel est rapidement accessible à travers son architecture. Celle-ci s'apprivoise au travers de neuf répertoires majeurs *Etc, SOLR, Multicore, Home, Lib, Src, Bin, Conf, Data*. Ceux qui m'intéressèrent plus particulièrement furent *Multicore* et *SOLR*

[6] Voir Lexique page 130.
[7] Voir Lexique page 129.

puisqu'ils définissent le contexte de déploiement de *SOLR*, selon qu'il soit multi-cœurs ou non, en même temps que son répertoire d'installation.

Chaque instance de *SOLR* est chargée à partir d'un fichier de configuration de type xml. Exemple, solar.xml ou multicore.xml pour les versions de *SOLR 1.3* et antérieures. Ils sont tous stockés dans *SOLR/Conf* ou *multicore/Conf*. Ces répertoires accueillent d'autres types de documents .xml et .txt définissant eux-même d'autres types d'actions, à l'exemple des documents « schema.xml » qui consignent la structure de l'index mobilisable dans l'instance lancée de *SOLR* et organisent les étapes d'analyse textuelle (l'action d'nalyseurs) devant forger le contenu de l'index.

/Data accueille le contenu de l'index. Celui-ci est obtenu grâce au travail d'un autre programme dont *SOLR* dépend, *Lucene*.

/Lib quant à lui accueille les API à même de faire des programmes extérieurs communiquer avec *SOLR*. *Lucene* est un de ces programmes, comme STAX en est un autre qui s'occupe de la gestion des procédés xml. Lib accueille encore toutes les librairies java requises à un moment ou un autre du processus faisant tourner *SOLR*.

Le fonctionnement de *SOLR*, nous l'avons vu, implique également le fonctionnement d'un système d'indexation de la ressource. Celui-ci suppose la maîtrise d'un ensemble de composantes, actives dans le fameux document schema.xml, composantes et langages qu'il faut apprendre à mobiliser. C'est au milieu de cette étape que mon auto-formation s'est interrompue.

4.4.3. Refonte de la base de données d'AssurWeb

La troisième mission à avoir été annulée en cours de route fut l'analyse et la refonte du modèle de la base de données qui sous-tendait le fonctionnement du site d'AssurWeb. Cette refonte, outre la vérification de la validité du travail réalisé par le salarié précédent, prévoyait l'intégration des données apportées par les mises en œuvre du TAL. Elle devint caduque avec l'arrivée d'un ingénieur informaticien, nouvelle ressource permanente et désormais chef de projet de l'équipe. Elle resta sa chasse gardée.

La base de données initialement présente dans le fond de l'entreprise était conceptuellement achevée mais ingérable dans la pratique, tant elle était pléthorique. L'essentiel du travail que j'accomplis à ce niveau fut donc de constater des caractères de complexité et de redondance inutiles et répétés. Dans la préparation de la tâche de

conception, j'eus à me familiariser avec Microsoft Office Visio Pro 2007, version plus évoluée d'un autre outil du même type dont j'avais eu à faire l'expérience en formation, DIA.

Chapitre 5 – Les modules d'extraction: données générales

Dans la trame de ce stage, la création des modules d'extraction d'informations ciblées par la société N5 mêle les activités d'analyse systémique et celles de développement informatique, piliers du traitement automatique. Montrer en quoi les solutions mises en place dans leur cadre a pu satisfaire aux enjeux de la discipline semble un incontournable.

5.1. Description physique du projet : les modules

5.1.1. Les processus d'extraction

Le développement du projet d'extraction fut donc réalisé en perl et s'appuya exclusivement sur les cours en ligne et les tutoriaux disponibles sur différents sites de développeurs. Quatre modules permettent de réaliser les extractions présentées plus tôt. Leurs processus sont lancés sur des pages qualifiées au préalable par une autre section du projet. Cependant pour les besoins de test et en l'absence de développement de ces fonctionnalités, mon projet s'équipa de son propre système de filtre. Ainsi les pages candidates à l'extraction devaient elles valider l'une des quatre catégories suivantes:

- des pages contenant des définitions lexicales, typiques du domaine des assurances ou non ;
- des pages contenant des listes de contacts ;
- des pages de presse où l'information cible était la structure « article ». Ce qui caractérise un article étant pour nous :
    - une date (facultative),
    - un titre,
    - une accroche (facultative), c'est-à-dire les premières lignes de l'article,
    - et un contenu, rarement sur la même page mais atteignable via une url.
- Des pages "contacts" qui présentent les coordonnées d'acteurs du secteur : coordonnées postales et parfois téléphoniques et web (email, site).

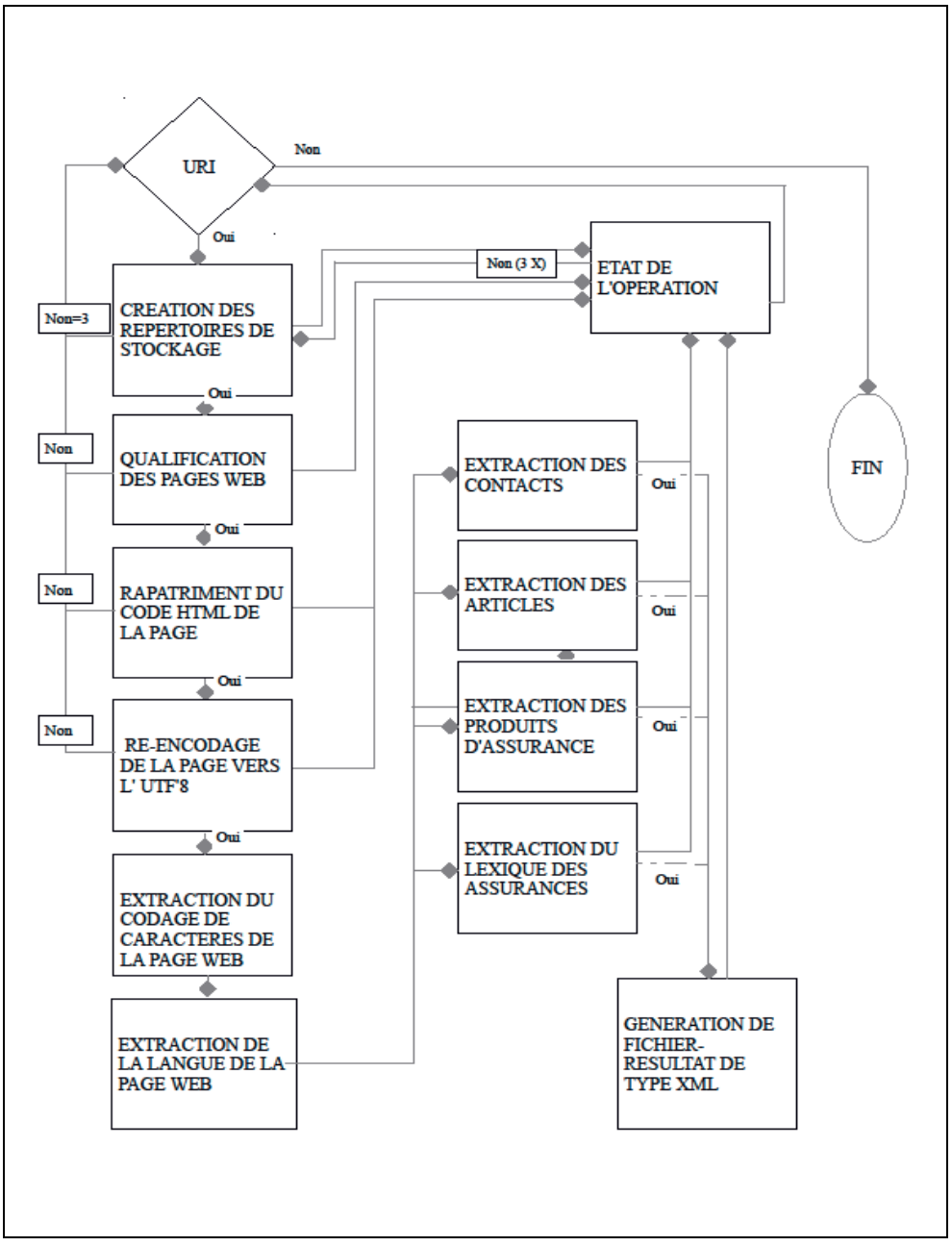
Les résultats de l'extraction devaient être consignés dans des fichiers xml, la génération de ces fichiers revenant à des modules dédiés.

5.1.2. Entrées et sorties des processus d'extraction

En entrée, ces processus acceptent deux chaînes de caractères : l'une contenant l'Url de la page à traiter et l'autre, la chaîne de caractères correspondant à son code html.

En sortie, ils produisent un tableau d'enregistrements consignant pour chaque entité extraite, la valeur trouvée pour ses différents champs, tous des chaînes de caractères.

Illustration 6 : Fonctionnement général du prototype d'extraction



5.1.3. Les modules de structuration de l'information au format XML

Les modules du projet dédiés à la mise en forme des informations extraites, répertorient spécifiquement les informations suivantes :

- generer_xml_glodico,
  - l'id de l'entrée lexicale
  - l'entrée lexicale
  - nombre de mots composant l'entrée lexicale
  - la définition associée à l'entrée lexicale
- generer_xml_produit
  - l'id du produit
  - le nom du produit
- generer_xml_article sauvegarde dans un fichier xml:
  - la date de l'article (facultative)
  - le titre de l'article
  - l'adresse où trouver l'article complet
- generer_xml_contacts stocke en lien avec la définition du contact
  - le nom de l'organisation ou personne,
  - le numéro de porte (13, 1 bis),
  - le type de voie (avenue, ruelle, impasse etc...),
  - le nom de la voie,
  - le code postal,
  - la ville,
  - l'adresse mail,
  - l'adresse du site,
  - en lien avec le traitement de la page, le nombre de contacts extraits.

Un certain nombre d'informations sont partagées communément par les différents xmls:

- L'url de la page traitée,
- le codage de la source html,
- le domaine de la source html,
- la langue de la page,

- le codage cible, celui utilisé sur l'ensemble du projet par les développeurs.

Ces différents modules fonctionnent à partir des tableaux d'enregistrement produits par leur pendant en extraction. Ils reçoivent donc des tableaux d'enregistrements et produisent des fichiers xml, en retour.

5.1.4. Le prototype extractionassurwebprototype

Le prototype extractionassurwebprototype déploie les 4 processus majeurs, présentés plus haut. Dans le détail son action est la suivante :

- il qualifie ces adresses en 5 grands groupes coïncidant avec les 4 types de pages définis plus haut, le cinquième étant celui des pages non reconnues.
- il créé les répertoires correspondant aux 5 types de pages traités.
- il récupère les contenus HTML de la liste d'adresses qualifiées.
- il stocke chaque contenu dans son répertoire dédié.
- il lance l'extraction d'informations différentielle sur les 4 répertoires.
- il génère les xmls consignant le résultat du traitement de chacune des pages.

Depuis la description figurant ci-dessus, certaines spécifications ont évolué, obéissant en cela, aux spécifications générales du projet qui n'étant pas fixées préalablement à mon intervention, changeaient de forme encore pendant celle-ci. Ainsi, une part importante de mon travail a consisté à modifier régulièrement mes options et choix de développement, à changer mes structures de données, modifier la nature des entrées sorties, à modifier les actions spécifiques des modules. A la fin de mon stage par exemple, la création d'*XML*, une disposition déclarée pourtant incontournable du projet devenait inutile. Dans ce processus, j'ai donc, de fait, expérimenté quelque chose d'assez déplaisant : reprendre en permanence mon code pour satisfaire des conditions contextuelles en constante évolution ; le tout en maintenant une documentation technique correspondant à chaque évolution souvent jamais achevée. Il en a résulté, chose positive, une compréhension approfondie du rôle idéal de la documentation technique et des habitudes de développement à avoir.

De nombreuses fonctions devraient encore évoluer d'ici la fin de la phase d'intégration du module, selon ce que sera la forme finale souhaitée par le chef de projet.

Initialement le prototype prenait en entrée une structure simple, une pile d'*Urls* (une pile de chaîne de caractère) et produisait en sortie une pile d'enregistrements eux-mêmes structurés autour de l'*url* de la ressource, de la catégorisation de cette dernière en fonction de 5 types (les 4 types de page recherchés et tout ce qui n'en était pas), d'une pile informative sur le déroulement des différents processus engagés par le prototype, du statut du processus d'extraction (tableau d'entiers), le nom de la ressource xml produite (type : chaîne de caractères).

5.2. Stratégie de résolution des modules :

5.2.1. Exemple : Extraction des contacts

Il s'agit d'extraire les champs constitutifs d'un contact c'est-à-dire un nom, une adresse, un ou des numéros de téléphone ou de fax. Dans l'ordre, on aura réalisé :

- l'étiquetage des occurrences de numéros ;
- l'étiquetage des types de voie (rue, avenue, zone industrielle, rond-point, etc et leurs variantes d'écriture) ;
- l'étiquetage des entités nommées, sur le critère simple de la forme de surface (la présence d'une majuscule en début de mot) ;
- la requalification et l'étiquetage des entités nommées de type pays sur la base d'une reconnaissance d'une entrée du dictionnaire « Pays » ;
- la requalification et l'étiquetage des entités nommées de type région sur la base d'une reconnaissance d'une entrée du dictionnaire des Régions ;
- la requalification des occurrences de numéros en codes postaux français ;
- la reconnaissance de plusieurs consécutions possibles des informations identifiées en fonction du caractère facultatif de leur apparition.

5.2.2. Exemple : Extraction des références d'articles mis en ligne.

- Comme cela a été évoqué plus haut, le but d'AssurWeb est de fournir un accès privilégié à ses clients à une ressource triée et organisée. Une de ses ressources étant les articles publiés par les éditeurs de sites d'assurance et la presse spécialisée, la première de ses tâches était donc de proposer une localisation de ces ressources. Le traitement qu'elles subiraient dans le cadre de la valorisation de services de N5

restant encore à définir. Me fixant sur les informations essentielles à ce stade du projet, j'identifiai les informations à capturer, leurs caractéristiques générales et leurs modalités d'apparition dans le fil d'un énoncé: la date de mise en ligne ou d'écriture de l'article, son titre, l'adresse où le trouver. Ces différentes informations étant là aussi étiquetées. L'extraction s'effectue bien évidemment sur une chaîne de caractères « nettoyée ».

5.2.3. Exemple : Extraction des glossaires

Le principe de l'extraction du lexique est à peu de chose près, le même que dans l'extraction d'articles. Les glossaires sont extrêmement normés et répondent à des contraintes de mise en valeur.

Pour ce qui est de ces caractères observables, les voici :

- un ou plusieurs mots, rarement une phrase, soit 1 à 6 mots, limite là encore fixée arbitrairement.
- un paragraphe. Par paragraphe je n'entends pas la mise en valeur par un type de balise précis <p> par exemple mais j'entends une mise en contraste du texte par rapport au titre, avec lequel il ne formera jamais un bloc indistinct.

Certaines contraintes pèsent sur le texte définitoire, comme l'apparition en plein exposé d'une information de type <a>je_suis_une_défintion_de_concept_dans une_définition_principale. Je me suis efforcée d'identifier ces cas de figure, en analysant le texte avant et après l'apparition de ces inserts. Si le texte trouvé correspondait à la définition que nous nous étions donné d'une définition lexicale, il était retenu et marqué comme tel.

L'avant dernière phase de l'extraction fut donc de qualifier les blocs pertinents, sachant qu'une définition n'apparaît jamais avant son entrée lexicale. Cette dernière règle présida donc l'ultime phase d'extraction.

Dans cette résolution, un parti pris a été choisi, celui de ne pas définir au préalable le lexique recherché. L'information rapatriée aurait été moins bruitée certes mais elle aurait été aussi, d'emblée, limitée. Elle n'aurait pas autoriser l'apprentissage. Or je le rappelle, le projet AssurWeb particulièrement et AssurGroupest dans une phase de collecte et dans une large mesure prévoit le concours d'une importante main d'œuvre de documentalistes, à même de nettoyer les résultats obtenus.

5.2.4. Exemple : Extraction des noms de produit.

L'extraction des noms de produit est à mon avis l'élément le plus faible de la trame d'extraction. L'information que l'on tente ici de saisir relève d'avantage de l'observation linguistique. Force est de reconnaître que les noms des produits de consommation obéissent à de véritables règles, qui elles même changent avec les époques. Un phénomène assez singulier existe par ailleurs, dans les règles de nommage des produits d'assurance. Ceux-ci sont évoqués par phrases nominales ou phrases verbales, autonomes ou juxtaposées

Il est très difficile d'identifier « l'étiquette » qui dans le langage réfère sans ambiguïté au produit. Entre phrases de présentation, phrases chocs ou même, références généralistes par domaines historiques des assurances, les occasions de se tromper sont nombreuses à la base, pour un intervenant humain.

Parmi les formulations suivantes, laquelle désigne un produit, ou même le contrat après tout ultime étape d'identification des ces produits ?

« L assurance auto au km près » ?

« L assurance auto au km » ?

« L'assurance auto Pay as you drive » ?

« L'assurance auto classique » ?

Les trois dernières uniquement.

Les procédures en matière de nommage sur ce marché consistent, en français, à utiliser :

- des formules maison :
  - HABITATION EVOLUTIS[8],
  - AUTO RISQUES SPECIFIQUES,
  - CONDUIRE,
  - MOBIL'HOME - quoique parfois généraliste.
- une nomination par catégorisation par domaine historique d'assurance :
  - L'ASSURANCE AUTO
- une sous catégorisation de la formulation précédente :
  - AUTO STANDARD,
  - AUTO RISQUES SPECIFIQUES,

[8] Source : April.fr

- AUTO JEUNE CONDUCTEUR,

- Un des domaines de l'assurance (personnalisé ou non), accompagné du nom de l'assureur postposé ou antéposé à celui-ci. Exemple, « Groupama Santé Active ». Là dessus, toutes les déclinaisons d' « option », de « formules », de « bonus » viennent enrichir/compléter/complexifier la dénomination, en apparitions explicites ou non.
    - Amaguiz santé double bonus[9]
    - Amaguiz santé option éco
    - Groupama Santé Active JEUNES budget légers
    - Groupama Santé Active Famille protection renforcée
    - Groupama Santé Active 55 ans et couverture adaptée +
    - CONDUIRE[10] (formule) assistance au véhicule
    - Défense pénale/recours suite à accident
    - attentat et actes de terrorisme,
    - événements climatiques,
    - dommage tout accident,
    - Assistance 0 km,
    - Auto Presto
    - Auto Presto Privilège
    - Vol,
    - Incendie,
    - Catastrophes naturelles

Le nom du produit d'assurance se construit littéralement au fil des liens cliqués, sans jamais apparaître de but en blanc, ou rarement, dans une seule et même page de présentation. Nul doute que l'extraction des noms de produit dans ce domaine d'activité, vaille à lui seul une étude.

Tous ces exemples viennent véritablement interroger la notion de nom. La définition la plus adaptée que j'aie trouvé dans le cadre de mon intervention, est celle de référent permettant à des acteurs humains d'identifier les référés de manière relativement rapide et surtout univoque.

Ne disposant pas d'analyseurs de chaînes, pas plus que du temps nécessaire à l'analyse approfondie des schémas de construction de ces expressions, le module construit

[9] Amaguiz.fr
[10] Goupama.fr

pour les produits va au plus simple, cherchant dans l'univers restreint des pages consacrées aux produits, des occurrences de mots distingués par une forme de surface en majuscules sur l'attaque du mot (ou de la suite de mots) et offrant une redirection de page.

Afin de restreindre la proportion d'extractions inadéquates, un ensemble de mots sont écartés par opération de filtrage. Les expressions récurrentes « A Lire », « La suite », « Ici », « La », « Nous contactez », « Demandez votre devis », « Faire une simulation », « devis », les occurrences de dates, tous les accidents « prévisibles » dans la configuration qui nous intéresse. Le lexique ainsi couvert est, pour partie la somme des lexiques particuliers mobilisés par les divers modules d'extraction vus jusqu'ici, le lexique des mots outils de la langue (prépositions, adverbes notamment), le lexique récurrent des liens des sites web en général et ceux des assurances, en particulier.

Les entrées du module d'extraction des produits sont celles mentionnées précédemment, à savoir une chaîne de caractères contenant le code de la page à traiter et son URL. Sa sortie consiste en un fichier xml répertoriant les diverses valeurs extraites et quelques informations contextuelles à propos de la page, telles sa langue, son codage de caractères original, celui posé par le traitement en extraction.

5.3. Les objectifs à moyen et long terme

Bien entendu, ces différents modules supporteront aisément d'être améliorés. La première initiative à engager sera l'extension des motifs d'extraction afin d'étendre les possibilités de reconnaissance, et l'extension des lexiques.

L'effort à faire doit se concentrer d'abord sur la langue centrale du français. Celui-ci rencontre des habitudes culturelles et parfois institutionnelles et administratives différentes suivant le pays francophone qui le mobilise. Ainsi la France, mobilise ses normes autrement que la Belgique ou le Canada, pour ne citer qu'eux.

Un effort particulier est à faire également avec l'anglais dont les adresses physiques, quoique très logiques varient des pratiques françaises. Ainsi l'anglais ne connaît pas de code postal tel qu'en France (suite maximale de 5 chiffres) mais un code mêlant points cardinaux et initiales de division administrative/régionale.

Des modèles d'extraction sont entièrement à construire pour des langues comme le chinois.

Partie 3

Extraction d’articles et extraction de contacts

Chapitre 6 – L'extraction des contacts

6.1. Analyse contextuelle

L'ambition de ce module d'extraction, on le répète était de réussir la capture des noms et libellés d'adresses d'entités nommées, personnes morales et personnes physiques, en français prioritairement mais aussi et autant que possible, dans les autres langues du projet. L'autre demande du commanditaire étant d'identifier les différents champs constitutifs de ces adresses, il fallait donc élaborer une typologie des champs d'informations qui y apparaissent, s'appuyant sur le degré de précision qu'ils introduisent et sur leurs lieux potentiels d'apparition, dans un libellé idéalement complet ou non.

En France, libeller une adresse se fait selon des règles éditées par La Poste. Celles-ci reprennent largement des traits organisationnels de ses services de distribution (bureaux), ceux de l'Etat et des ses divisions administratives. L'une de ses règles bien connues veut que le code postal entièrement numérique ne compte que 5 chiffres. Il doit être suivi du nom de la ville. La mention du pays intervient en dernière position dans le libellé et reste facultative.

A l'intérieur de ces règles, une certaine latitude s'exprime dans l'agencement des informations. Elle est liée aux individus ou à des pratiques ou normes antérieures non oubliées. Sur le Web, on peut s'attendre à ce que la standardisation des informations permette de laisser de côté ces infra-langages, tout au moins sur les sites à caractères officie, mais ceci est loin de constituer une garantie.

Pour rester sur le plan des standards, il existe d'autres règles pour libeller les adresses en langue française, que celles en vigueur en France : celles de la Belgique[11], de la Suisse et du Canada, pour ne citer que ces pays-là. Car le français, c'est aussi et largement la francophonie.

Ces derniers standards ne nécessitent, très souvent, qu'un aménagement des règles de cette norme. Le code postal suisse qui lui n'est pas exclusivement numérique, est le plus souvent précédé de 2 lettres, celles de la subdivision administrative. La suite numérique terminant le code postal, n'excède pas 4 chiffres quant à elle. Le code postal canadien tient également du code alphanumérique où une partie du code désigne, l'état auquel appartient

[11] Voir exemples d'adresses libellées en français Annexe 10 et 11.

l’adresse. Le code postal belge est comme en France, entièrement numérique mais composé uniquement sur quatre chiffres.

L’usage des séparateurs à l’intérieur des signatures alphanumériques, séparant code alphabétique et code chiffre, intéresse lui aussi la reconnaissance. Il n’est pas fixe dans la pratique des utilisateurs et n’est pas toujours utilisés, même quand il est prescrit.

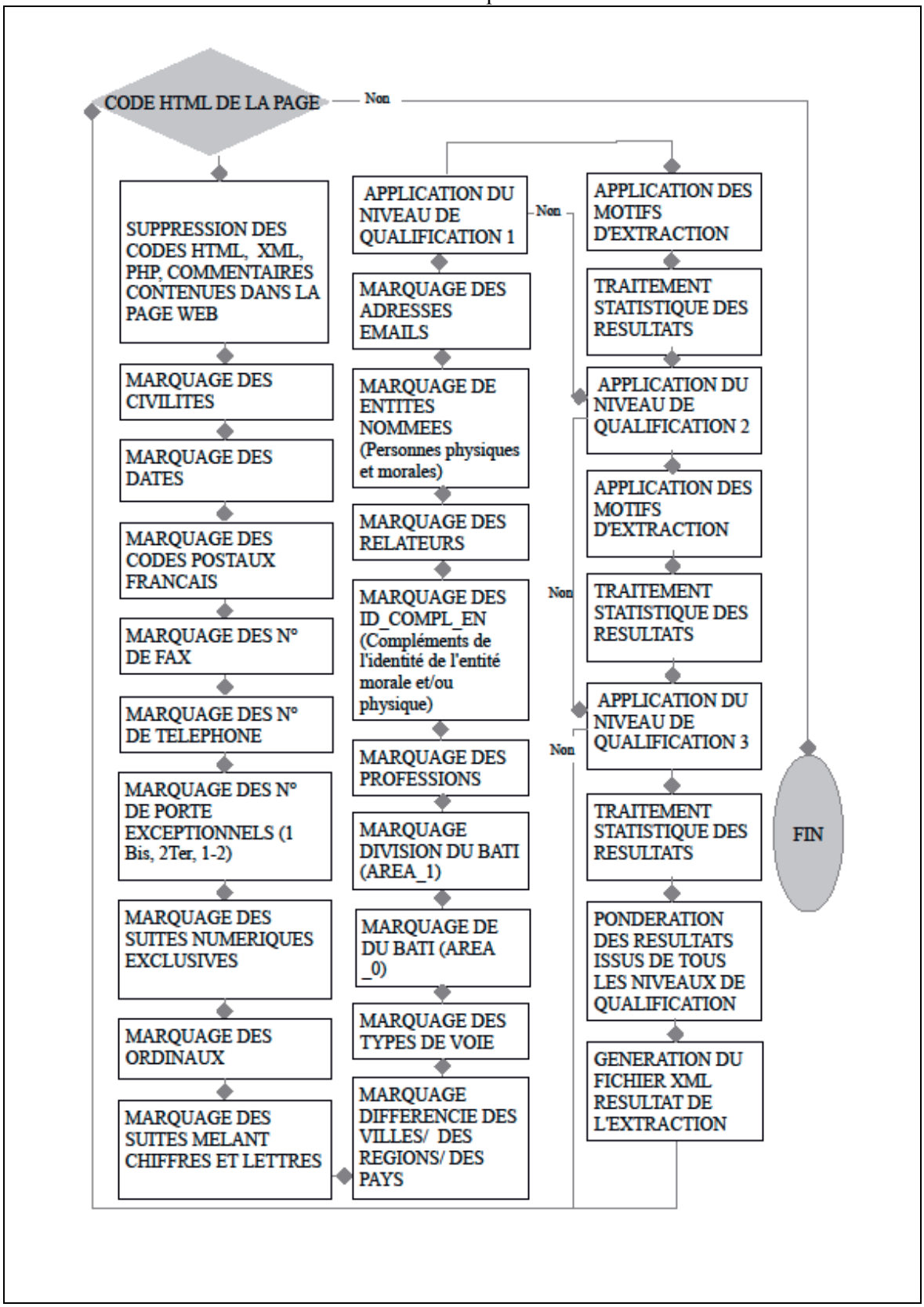
Autre particularité liée au modèle belge celle-là, le numéro du lieu de résidence apparait systématiquement après la mention du nom de la rue.

Pour poursuivre avec les particularités, au Canada français et anglophone, la règle de composition veut que le code postal apparaisse toujours en fin de ligne, précédé uniquement par le nom de la ville, pour ce qui est de la pratique française, du nom de la ville et du nom de l’Etat, pour la pratique anglaise. Dans tous les cas de figure, le nom du pays est la dernière information apparaissant dans l’adresse.

Toutes les dispositions évoquées jusqu’ici font qu’une gestion particulière des suites numériques et alphanumérique doit permettre la prise en charge des différents types de codes postaux. Un effort d’analyse supplémentaire doit surveiller les ordres d’apparition des informations du type « localité » et « état », permettant ainsi de rendre compte d’une large part des réalités anglophones.

La pratique des adresses belge quant à elle, parce qu’elle passe par la maîtrise de l’ambiguïté entre le numéro de pas de porte et code postal, doit définir un élément de contrainte dans la reconnaissance, ignoré longtemps dans le mode de résolution que je m’étais chois, le « retour à la ligne ».

Illustration 7 : Schématisation du processus d'extraction des contacts



6.2. Grammaire d'extraction : Définitions formelles

6.2.1. Les lexiques

L'entrée soumise à l'extraction est du texte simple, soit le code html de la page renvoyée par le *crawler*. Ce texte subit alors une phase préparatoire.

Les balises non pertinentes, à l'exemple des images, sont supprimées, les balises restantes sont simplifiée ;

Le texte est régularisé : les blancs multiples sont réduits à des blancs uniques. Cette définition de « blancs » comprend les tabulations, les espaces, simples ou multiples, les composés des deux. Ils sont systématiquement remplacés par un unique espace simple. Seuls les caractères de retours à la ligne sont préservés, en une occurrence unique toutefois ;

Les suites exclusivement numériques sont formatées. Les espaces entre les chiffres formant une seule entité, en effet, sont supprimés.

Les lexiques sont les suivants :

LEX_pays : répertorie les noms des pays majeurs du globe, en français et en anglais ;

LEX_ville : répertorie les noms des villes majeurs du globe, dans les mêmes langues ;

LEX_etat : consigne les noms en extension et les sigles des Etats américains et canadiens, toujours en anglais et français ;

LEX_région : Fait la même chose, cette fois-ci avec le nom des principales régions du globe. L'inventaire reste centré sur le continent américain pour l'instant.

LEX_ann_pays : lexique introduisant l'annonce du nom d'un pays ;

LEX_ann_ville : lexique introduisant l'annonce du nom d'une ville ;

LEX_ann_etat : lexique faisant l'annonce du nom d'un état ;

LEX_ann_region : lexique faisant l'annonce du nom d'une région ;

LEX_bp = (PO|P.O|BP|B.P|p.o|PO BOX|PO Box|po box|b.p|Boîte Postale|Boite postale|BOITE POSTALE, Bte| bte| Bte postale| bte postale).

LEX_cedex= (Cedex[Cédx|CEDEX|CEDX|cédex|cedex|CIDEX|Cidex|cidex).

LEX_tel= (TEL|TÉL|Tel|Tél|TELEPHONE|TÉLÉPHONE|Téléphone|Telephone).

LEX_area0 : Lexique en langue française et anglaise décrivant le niveau de localisation spatiale des adresses, en termes d'artères, de systèmes structurant reconnaissables. Ce sont aussi bien les rues, avenues, allées, voies, (étages, floor), galeries, que les jonctions d'artères, ronds-points, places ou plus discriminant, les pas porte avec les entrées comme « étage », « pallier », « escalier », « cage d'escalier », « porte », « appartement ». Une partie du contenu de cette sélection comporte notamment des suites formées autour de ces mots. Ainsi, rue, allée, impasse, s'étendent dans des expressions comme: grand rue, grand'rue, grande allée, allée principale, etc …

Ce répertoire rassemble également une partie du vocabulaire topographique, celui autour duquel nombre de lieux-dits, localités ont formé ou dorment encore leur nom (en anglais mais pas seulement): les criques, anses, mare etc… Ce niveau de description s'avère on ne peut plus nécessaire quand le seul élément de localisation est du type lieu-dit, sans référence à une voie[12].

LEX_area1 : cet ensemble regroupe le lexique en langue française et anglaise décrivant un niveau de description spatiale supplémentaire qui repose sur la qualification de la construction urbaine. Relativement autonome, il autorise la localisation, plus difficile certes, grâce à la connaissance sociale d'un environnement. Ce sont les unités du type quartiers, hameaux, zones artisanales, industrielles, résidences, immeubles, bâtiment, édifices, complexes, ville nouvelle.

Ce sont tous les éléments urbains, ceux notamment construits autour de termes comme *hôtel, cité, complexe, centre. immeuble, centre des congrès, centre commercial, centre artisanal, village artisanal, zone artisanale, zone d'activités, zone commerciale, palais des congrès, tour, ensemble, etc.* Leur énoncé ne coïncide avec aucun système de localisation pourtant ils sont significatifs par rapport à une aire géographique, comme construction sociale partagée.

LEX_civilites :

La reconnaissance des civilités regroupe les civilités usuelles, M, Mme, Melle (formes longues et contractées), les civilités professionnelles, honorifiques, titre de noblesse, en anglais, espagnol et autres langues latines, en plus du français. Soit toutes les

[12] Voir l'exemple n°3 de l'annexe 12 : adresse anglaise.

formules du type « Maître », « Professeur », « Docteur », « Premier Ministre », « Chancelier » etc..

LEX_ssen: « *service de recouvrement », « service technique », « direction départementale », « direction financière », etc.* Des termes comme « bureau », « département » sont ambigus car ils tombent aussi bien dans les catégories « service » (LEX_ssen), que celles d'entités nommées, ou de qualification de la construction (Area_1)

LEX_profession : *« directeur », « responsable », « trésorier », « consultant », « avocat »*, etc sont le plus souvent les têtes lexicales d'expressions plus longues précisant l'activité sectorielle des individus.

rel ={à la, à, aux, au, de le, de la, de l', des, du}

Exemple 1 :

Mr DUPONT Eric

Résidence Les Terrasses

Immeuble Sérénité.

Porte 311

Exemple 2

Mr Peter SMITH

Immeuble TF1

Place François Mitterrand

Services des Ressources humaines

Porte 821.

Ce sont des organisations de l'espace qui s'apparentent à des lieux dits, c'est-à-dire, qu'ils sont identifiés dans vie sociale, culturelle de la localité, de la ville, voire du pays.

A noter qu'en anglais il recoupe toutes les dénominations en « Centre », « Center », « Plaza », « Gallery » etc.

La grammaire du langage créé dans le cadre de cette extraction suppose l'usage d'un vocabulaire.

Tableau 1 : Vocabulaire de la grammaire des contacts

Libellés	Codification
Civilités	<CIVILITES></CIVILITES>
Balises marquant l'occurrence d'une entité nommée	<EN></EN>
Balises marquant l'occurrence d'une « sous entité ». Celles-ci étant construite sur une tête lexicale du type « service », « direction », « secrétariat » dans des expressions complètes à l'exemple de celles-ci : « Service Client », « Direction financière », « Service de Prospection », etc.	<SSEN></SSEN>
Balises marquant la présence d'un métier	<PROFESSION></PROFESSION>
Balises marquant la présence d'un numéro de téléphone.	<TEL></TEL
Balises marquant la présence d'un numéro de fax.	<FAX></FAX>
Balises marquant la présence d'un numéro cedex (norme française)	<CEDEX></CEDEX>
Balises marquant l'occurrence d'une boîte postale	<BP></BP>
Balises marquant l'emplacement d'un numéro, compris comme une suite	<NUM></NUM>

unique de chiffres.	
Balise marquant l'occurrence de numéros à 1 ou 2 chiffres uniquement	<NUM2></NUM2>
Balise marquant l'existence d'un code, soit un mélange de lettres et de chiffres	<CODE></CODE>
Balise marquant la présence d'un ordinal	<ORDINAL></ORDINAL>
Balises marquant la reconnaissance d'un code postal, français ou anglais.	<CODE_POSTAL></CODE_POSTAL>
Balises pointant les marqueurs lexicaux appartenant à LEX_area0 (typologie 0).	<AREA_0></AREA_0>
Balises pointant les marqueurs lexicaux appartenant à LEX_area1 (typologie 1).	< AREA_1></AREA_1>

6.2.2. Les règles de la reconnaissance

Complétant ce vocabulaire, on a encore :

Ø : mot vide (début et fin de chaîne), espace simple.

r : caractère de retour à la ligne ou de fin de ligne.

Les règles du langage sont :[13]

R1 : rel → de| du|des|de la| de l'|les|le|la|l'(rien que des minuscules!)

R2 : rel2 → De|Du|Des|De La|De L'|Les|Le|La|L'|DE| DU|DES|DE LA| DE L'|LES|LE|LA (des majuscules en attaque et le reste de la casse en minuscules (uniquement) ou majuscules (uniquement).

R3 : article→ les|le|la|l'|La|Les|Le|La|L'|LES|LE|LA (articles définis)

R4 : PUNCT→ [- ?./ !:;,]

(Nettoyage de la ponctuation dans les espaces inter-balises)

[13] Les espaces dans ces notations n'ont pas d'autre pertinence que de rendre la lecture plus aisée. Ils ne correspondent pas à un élément de reconnaissance.

R5 : </(EN|BP|NUM|A|…|CODE)> ((Ø|w|W|NUM) (PUNCT (Ø|w|W|NUM)$^{+}$)$^{+}$ < ┐ /→ </(EN|BP|NUM|A|…|CODE)> (Ø|w|W|NUM) (Ø|w|W|NUM)$^{+}$)$^{+}$ < ┐ /

R6 : A →[a-zA-Z]

R7 : W→mot commençant par une majuscule ;

R8 : w→mots en minuscules ;

R9 : NUM →[0-9]

R10 : NUM2→[0-9]

R11 : NUM →NUM NUM

R12 : ┐ NUM →┐ [0-9]

R13 : C2→NUM2 NUM2|NUM2

R14 : ((Ø|PUNCT Ø)(W|w|r) |r|Debutdechaîne) NUM^{+} ((Ø|PUNCT Ø) (W|w|r)|r|Findechaîne)→ ((Ø|PUNCT Ø)(W|w|r)|r|Debutdechaîne) <NUM>NUM^{+}</NUM> ((Ø|PUNCT Ø) (W|w|r)|r|Findechaîne)

R15 : ORDINAL→ NUM^{+} Ø (e|è|ème|ième|ier|er|st|rd|th)

R16 : <NUM>NUM+</NUM> Ø (e|è|ème|ième|ier|er|st|rd|th) (Ø|r) → <ORDINAL>NUM^{+} Ø (e|è|ème|ième|ier|er|st|rd|th)</ORDINAL> (Ø|r)

R17 : ORDL→premier|première|second|seconde|troisième|first|second|third|...)

R18 : (┐ NUM Ø|r|debut de chaine) (ORDINAL|ORDL)(Ø|PUNCT) → (┐ NUM Ø|r|debut de chaine) <ORDINAL>(ORDINAL|ORDL)</ORDINAL> (Ø|r|PUNCT)

R19 : (Ø|r) LEX_cedex ((Ø|r) <NUM>C2</NUM>)$^{(0,1)}$ → (Ø|r)LEX_cedex (Ø|r) <CEDEX>Ø C2$^{(0,1)}$</CEDEX>

R20 : T→ LEX_tel

R21 : F→ LEX_fax

R22 : T (-|+|.|/|Ø|NUM2|)|(|,| ;)$^{+}$ (Ø (w|W))→ T <TEL>(-|+|.|/|Ø|NUM2|)|()$^{+}$</TEL> (Ø (w|W))

R23 : T ((-|Ø|.| :|/|+|,| ;)$^{+}$ <NUM>NUM</NUM>)$^{+}$→ T (<TEL>(-|Ø|.| :|/ |,| ;)$^{+}$NUM</TEL>)$^{+}$

R24 : (Ø$^{+}$ ((-|.| :|/) Ø)$^{(0,1)}$ (+ Ø$^{+}$)$^{(0,1)}$ (\((<NUM>NUM$^{(10,15)}$</NUM>|NUM2$^{(10,15)}$)\) |(<NUM>NUM</NUM>|NUM2))$^{+}$ Ø*,$^{(0,1)}$ (([-./]$^{(0,1)}$ Ø <NUM>NUM</NUM>)$^{+}$ |[-./]$^{(0,1)}$ NUM2^{+})* (Ø ┐ [NUM|NUM2|<]*)→ <TEL> (Ø$^{+}$ ((-|.| :|/) Ø)$^{(0,1)}$ (+ Ø$^{+}$)$^{(0,1)}$ (\((<NUM>NUM$^{(10,15)}$</NUM>|NUM2$^{(10,15)}$)\) |(<NUM>NUM</NUM>|NUM2))$^{+}$ Ø$^{(0,1)}$,$^{(0,1)}$ (([-./]$^{(0,1)}$ Ø <NUM>NUM</NUM>)$^{+}$ |[-./]$^{(0,1)}$ NUM2^{+})* (Ø$^{(0,1)}$ ┐ [NUM|NUM2|<]*)< /TEL>

R25 : F (-|+|.|/|Ø|NUM2|\)|\(|,| ;)$^{+}$ (Ø [w|W])→ F <FAX>(-|+|.|/|Ø|NUM2|\)|\()$^{+}$</FAX> (Ø [w|W])

R26 : F ((-|Ø|.| :|/|+|,| ;)$^{+}$ <NUM>NUM</NUM>)$^{+}$→ F (<FAX>(-|Ø|.| :|/ |,| ;)$^{+}$NUM</FAX>)$^{+}$

Règles gérant les personnes physiques et morales :

R27 : (┐W Ø|debut de chaine|r) W^{+} (┐W Ø|Findechaine|r)→ (┐W Ø|debut de chaine|r)<EN>W^{+}</EN> (┐W Ø|Findechaine|r)

R28 : ┐(</EN> & PUNCT) Ø (<EN>W^{+}</EN>)$^{+}$ ┐(</EN>|PUNCT)→ ┐(</EN>|PUNCT) <EN>(W^{+})$^{+}$</EN> ┐(</EN>|PUNCT)

R29 : (Ø|r) LEX_ssen (Ø|r|PUNCT)→ (Ø|r) <SSEN>LEX_ssen</SSEN>(Ø|r|PUNCT)

R30 : <SSEN>LEX_ssen</SSEN> ((Ø<EN>W^{+}</EN>)*(Ø rel (Ø w^{5}| Ø <EN>W^{+}</EN>))*→

<SSEN>LEX_ssen (Ø<EN>W^{n}</EN>)*(Ø rel (Ø w^{5}| Ø <EN>W^{+}</EN>))*<SSEN>

R31: (Ø|r) LEX_profession (Ø|r|PUNCT)→ (Ø|r) <PROFESSION>LEX_profession </PROFESSION>(Ø|r|PUNCT)

R32 : <PROFESSION>LEX_profession</PROFESSION> (Ø <EN>W^{+}</EN>)*(Ø rel (Ø (w^{5}| <EN>W^{+}</EN>)))*→ <PROFESSION> LEX_profession (Ø <EN>W^{+}</EN>)* (Ø rel (Ø (w^{5}|<EN>W^{+}</EN>)))* </PROFESSION>

Règles sur les civilités :

R33 : (Ø|r)Lex_civilites (Ø rel ØLEX_civilites)* (Ø|r)→ <CIVILITES>LEX_civilites (Ø rel Ø)* Lex_civilites</CIVILITES> (Ø|r)

Règles gérant les codes :

R34 : C→((A NUM| NUM A) (A|NUM)*((-|.|/|Ø) (A|NUM)) *| A)

R35 : (A$^{(2,n)+}$|NUM^{+})$^{(0,1)}$ (Ø|r)(C Ø)$^{+}$ C (Ø|r) (([A-Z]$^{(2,n)}$|[a-z]$^{(2,n)}$|[A-Z] [a-z]$^{(2,n)}$|NUM^{+})$^{(0,1)}$ (Ø|r))→ A$^{(2,n)+}$|NUM^{+})$^{(0,1)}$ (Ø|r) <CODE>(C Ø)$^{+}$C</CODE> (Ø|r) (([A-Z]$^{(2,n)}$|[a-z]$^{(2,n)}$|[A-Z] [a-z]$^{(2,n)}$|NUM^{+})$^{(0,1)}$ (Ø|r))

Dans cette définition, un code est une lettre isolée ou la succession autonome d'au moins deux caractères, commençant par un chiffre ou une lettre, mêlant dans son corps, chiffres et lettres, d'éventuels caractères de séparation, mais ne finissant jamais sur l'un d'eux.

Règles gérant les codes postaux :

R36 : CP→ NUM2 NUM2 NUM2 NUM2 NUM2

R37 : CP→ NUM2 NUM2 NUM2 NUM2

R38 : (Ø|r) (LEX_cp (Ø|-|.| /|:)*) (CP) (Ø|r)→ (Ø|r) (LEX_cp Ø (-|.| /)*) <CODE_POSTAL>CP</CODE_POSTAL>

Règles gérant les boîtes postales :

R39 : Ø LEX_bp (Ø|-|.| :|/)$^{+}$ Ø NUM^{+}→ Ø LEX_bp (Ø|-|.| :|/)$^{+}$ <BP>NUM^{+}</BP>

R40 : Ø NUM^{+} Ø LEX_bp (Ø|-|.| :|/)$^{+}$ → Ø <BP>NUM^{+}</BP> Ø LEX_bp Ø

R41: (Ø|r) LEX_bp (Ø|-|.| /| :)$^{+}$ (<CODE>C</CODE>|<NUM>NUM</NUM>|<CODE_POSTAL>CP<CODE_POSTAL>) → (Ø|r) LEX_bp (Ø|-|.| /| :)$^{+}$<BP>(C|NUM|CP)</BP>

R42 - [ANGLAIS] :

(Ø|r)(<CODE>C</CODE>|<NUM>NUM</NUM>|<CODE_POSTAL>CP<CODE_POSTAL>) (Ø|.|-) LEX_bp (Ø|r|PUNCT)→ (Ø|r) <BP>(C|NUM|CP)</BP> (Ø|.|-) LEX_bp (Ø|r|PUNCT)

Règles gérant les données du bâti :

R43 : SPB→ LEX_area0

R44 : (Ø|r)SPB (Ø|r|PUNCT)→ (Ø|r)<AREA_0>SPB</AREA_0>(Ø|r|PUNCT)

R45 :

(Ø(<ORDINAL>ORDINAL</ORDINAL>|<NUM>NUM</NUM>|<CODE>C</CODE>|<CODE_POSTAL>CP</CODE_POSTAL>))* <AREA_0>SPB</AREA_0> ((Ø*rel$^{(1,1)}$Ø<EN>W^{+}</EN>)$^{+}$| Ø<EN>rel2$^{(0,1)}$W^{+}</EN>| Ø<NUM>NUM</NUM>|Ø<CODE>C</CODE> | Ø<CODE_POSTAL>CP</CODE_POSTAL> | Ø<ORDINAL>ORDINAL</ORDINAL>)$^{(1,2)}$→ Ø <AREA_0>(<ORDINAL>ORDINAL</ORDINAL>|<NUM>NUM</NUM>|<CODE>C</CODE>|<CODE_POSTAL>CP</CODE_POSTAL>)* SPB ((Ø*rel$^{(1,1)}$Ø<EN>W^{+}</EN>)$^{+}$|Ø<EN>rel2$^{(0,1)}$W^{+}</EN>|Ø:<NUM>NUM</NUM>|Ø<CODE>C</CODE>|Ø<CODE_POSTAL>CP</CODE_POSTAL>|Ø<ORDINAL>ORDINAL</ORDINAL>)$^{(1,2)}$</AREA_0>

La règle permet de rendre compte des extractions suivantes :

« 10 rue/avenue/faubourg Caumartin/des Angelots »

« Porte/Escalier/appartement 25 »

Elle prend pareillement en charge les variations typographiques :

10 rue de la république / 10 rue DE LA REPUBLIQUE/ 10 rue De La République

Elle capture l'exception belge avec l'inversion du pas de porte :

« **Bâtiment administratif de la Pontaise**

Avenue des Casernes 2

10 14 Lausanne »

R46 - [anglais] :

(<CODE>C</CODE>|<CODE_POSTAL>CP</CODE_POSTAL>|<NUM>NUM</NUM>)*(((<ORDINAL>ORDINAL</ORDINAL>) ((rel$^{(1,1)}$<EN>W^{+}</EN>)$^{+}$ | <EN>rel2$^{(0,1)}$W^{+}</EN>)) | ((rel$^{(1,1)}$<EN>W^{+}</EN>)$^{+}$ | <EN>rel2$^{(0,1)}$W^{+}</EN>) (<ORDINAL>ORDINAL</ORDINAL>) | ((rel$^{(1,1)}$<EN>W^{+}</EN>)$^{+}$ |<EN>rel2$^{(0,1)}$W^{+}</EN>) | <ORDINAL>ORDINAL</ORDINAL>) Ø <AREA_0>SPB < /AREA_0> (Ø (<NUM>NUM</NUM>|<CODE>C</CODE>|<CODE_POSTAL>CP</CODE_POSTAL>|<ORDINAL>ORDINAL</ORDINAL>)) → Ø <AREA_0> (<CODE>C</CODE>|<CODE_POSTAL>CP</CODE_POSTAL>|<NUM>NUM</NUM>)*(((<ORDINAL>ORDINAL</ORDINAL>) ((rel$^{(1,1)}$<EN>W^{+}</EN>)$^{+}$ | <EN>rel2$^{(0,1)}$W^{+}</EN>)) | ((rel$^{(1,1)}$<EN>W^{+}</EN>)$^{+}$ | <EN>rel2$^{(0,1)}$W^{+}</EN>) (<ORDINAL>ORDINAL</ORDINAL>) | ((rel$^{(1,1)}$<EN>W^{+}</EN>)$^{+}$|<EN>rel2$^{(0,1)}$W^{+}</EN>) |<ORDINAL>ORDINAL</ORDINAL>) SPB (Ø (<NUM>NUM</NUM>|<CODE>C</CODE>|<CODE_POSTAL>CP</CODE_POSTAL>|<ORDINAL>ORDINAL</ORDINAL>))</AREA_0>

R47 : TPLDTU→ LEX_area1

R48 : (Ø|r) TPLDTU (Ø|r|PUNCT)→ (Ø|r) <AREA_1>TPLDTU</AREA_1>(Ø|r|PUNCT)

R49 : <AREA_1>TPLDTU</AREA_1> ((Ø rel$^{(1,1)}$Ø<EN>W^{+}</EN>)$^{+}$ | Ø <EN>rel2$^{(0,1)}$W^{+}</EN> | Ø <NUM>NUM</NUM> | Ø <CODE>C</CODE> | Ø <CODE_POSTAL>CP</CODE_POSTAL> | Ø <ORDINAL>ORDINAL</ORDINAL>)$^{(1,2)}$→ <AREA_1> TPLDTU ((Ørel$^{(1,1)}$Ø<EN>W^{+}</EN>)$^{+}$ | Ø <EN>rel2$^{(0,1)}$W^{+}</EN> | Ø <NUM>NUM</NUM> |<CODE>C</CODE> | Ø <CODE_POSTAL>CP</CODE_POSTAL> | Ø <ORDINAL>ORDINAL</ORDINAL>)$^{(1,2)}$</AREA_1>

R50 - [anglais] :

((Ørel$^{(1,1)}$Ø<EN>W^{+}</EN>)$^{+}$ | Ø<EN>rel2$^{(0,1)}$W^{+}</EN> | Ø<NUM>NUM</NUM> |Ø<CODE>C</CODE>|Ø<CODE_POSTAL>CP</CODE_POSTAL>|Ø<ORDINAL>ORDINAL</ORDINAL>)$^{(0,1)}$<AREA_1>TPLDTU</AREA_1> (Ø(<NUM>NUM</NUM>|<CODE>C</CODE>|<CODE_POSTAL>CP</CODE_POSTAL>|<ORDINAL>ORDINAL</ORDINAL>))$^{(0,1)}$→ <AREA_1>((Ø rel$^{(1,1)}$Ø<EN>W^{+}</EN>)$^{+}$| Ø <EN>rel2$^{(0,1)}$W^{+}</EN>|Ø<NUM>NUM</NUM>|Ø<CODE>C</CODE>|Ø<CODE_POST

AL>CP</CODE_POSTAL>|Ø<ORDINAL>ORDINAL</ORDINAL>)$^{(0,1)}$ TPLDTU ((Ø rel$^{(1,1)}$<EN>W^{+}</EN>)$^{+}$| Ø <EN>rel2$^{(0,1)}$W^{+}</EN>)*</AREA_1>

R51 : (<EN>LEX_etat</EN>|(Ø|r) LEX_etat (Ø|r|PUNCT)) → Ø <EN_etat> LEX_etat</EN_etat> Ø

R52 : (<EN>LEX_region</EN>| Ø|r) LEX_region (Ø|r|PUNCT))→ Ø EN_region>LEX_region</EN_region> Ø

R53 : (<EN>LEX_ville</EN>|(Ø|r) LEX_ville (Ø|r|PUNCT)) → Ø <EN_ville>LEX_ville</EN_ville> Ø

R54 : (<EN>LEX_pays</EN>| (Ø|r) LEX_pays(Ø|r|PUNCT)) → Ø <EN_pays>LEX_pays</EN_pays> Ø

Les motifs d'extraction

Après l'application des règles de grammaire, la chaîne textuelle créée est de nouveau traitée afin d'en supprimer tous les espaces nés des manipulations des phases précédentes. Ils sont réduits, comme en début de traitement.

Motifs : les sous unités

Dans leur énoncé, les motifs définis ci-dessous font l'impasse de la gestion des espaces, dans le but d'offrir une lisibilité accrue. Il n'en reste pas moins qu'ils sont à prendre en compte lors de la mise en œuvre de l'extraction.

(<CIVILITES> Lex_civilites (rel LEX_civilites)*</CIVILITES>)* <EN>W^{+}</EN> (rel <EN>W^{+}</EN>)* (<PROFESSION>Lex_profession (rel* <EN>W^{+}</EN>)$^{+}$ (rel (w^{5}| <EN>W^{+}</EN>))*</PROFESSION>)*.

1) (<CIVILITES>LEX_civilites</CIVILITES>)* ((article défini|rel) (<CIVILITES>LEX_civilites</CIVILITES> | (<PROFESSION>LEX_profession (<EN>W^{+}</EN>)*(rel (w^{5}|<EN>W^{+}</EN>))*</PROFESSION>)))$^{+}$ (<EN>W^{+}</EN>)*

a) <AREA_0> (<ORDINAL>ORDINAL</ORDINAL>|<NUM>NUM</NUM>|<CODE>C</CODE>|<CODE_POSTAL>CP</CODE_POSTAL>)* SPB ((rel$^{(1,1)}$<EN>W^{+}</EN>)$^{+}$|<EN>rel2$^{(0,1)}$W^{+}</EN>)(<NUM>NUM</NUM>|<CODE>C</CODE>|<CODE_POSTAL>CP</CODE_POSTAL>|<ORDINAL>ORDINAL</ORDINAL>)

(rel$^{(1,1)}$<EN>W^{+}</EN>)$^{+}$|<EN>rel2$^{(0,1)}$W^{+}</EN>|<NUM>NUM</NUM>|<CODE>C</CO

DE>|<CODE_POSTAL>CP</CODE_POSTAL>|<ORDINAL>ORDINAL</ORDINAL>)
</AREA_0>

Le motif rend compte d'adresses du type :« 10 avenue du Grésivaudan », « 1bis rue des Alloyes ».

a') [anglais]
<AREA_0>(<CODE>C</CODE>|<CODE_POSTAL>CP</CODE_POSTAL>|<NUM>N
UM</NUM>)* (((<ORDINAL>ORDINAL</ORDINAL>) ((rel$^{(1,1)}$<EN>W^{+}</EN>)$^{+}$ |
<EN>rel2$^{(0,1)}$W^{+}</EN>))|((rel$^{(1,1)}$<EN>W^{+}</EN>)$^{+}$ | <EN>rel2$^{(0,1)}$W^{+}</EN>)
(<ORDINAL>ORDINAL</ORDINAL>)|((rel$^{(1,1)}$<EN>W^{+}</EN>)$^{+}$
|<EN>rel2$^{(0,1)}$W^{+}</EN>)|<ORDINAL>ORDINAL</ORDINAL>) SPB
(<NUM>NUM</NUM>|<CODE>C</CODE>|<CODE_POSTAL>CP</CODE_POSTAL
>|<ORDINAL>ORDINAL</ORDINAL>)$^{(0,1)}$</AREA_0>

Le motif permet de rendre compte de chaîne du type :
« 10 Wellington Avenue »,
« 1220 Jameson Street »,
« 1220 of Jameson Street »,
« 1220 North 45th Street »,
« 1220 45th North Street »,
« 1220 45th of North Street »,
Ou encore de :
« 4th floor/Door »,

c) <AREA_1> TPLDTU
(((rel$^{(1,1)}$<EN>W^{+}</EN>)$^{+}$|<EN>rel2$^{(0,1)}$W^{+}</EN>)(<NUM>NUM</NUM>|<CODE>C</C
ODE>|<CODE_POSTAL>CP</CODE_POSTAL>|<ORDINAL>ORDINAL</ORDINAL
>) | (
(rel$^{(1,1)}$<EN>W^{+}</EN>)$^{+}$|<EN>rel2$^{(0,1)}$W^{+}</EN>|<NUM>NUM</NUM>|<CODE>C</CO
DE>|<CODE_POSTAL>CP</CODE_POSTAL>|<ORDINAL>ORDINAL</ORDINAL>)
$^{(0,1)}$)</AREA_1>
« Complexe sportif des Eyrolles »

c') - [anglais]:
<AREA_1>((<NUM>NUM</NUM>|<CODE>C</CODE>|<CODE_POSTAL>CP</COD
E_POSTAL>|<ORDINAL>ORDINAL</ORDINAL>)

((rel$^{(1,1)}$<EN>W^{+}</EN>)$^{+}$|<EN>rel2$^{(0,1)}$W^{+}</EN>)|
((rel$^{(1,1)}$<EN>W^{+}</EN>)$^{+}$|<EN>rel2$^{(0,1)}$W^{+}</EN>|<NUM>NUM</NUM>|<CODE>C</CODE>|<CODE_POSTAL>CP</CODE_POSTAL>|<ORDINAL>ORDINAL</ORDINAL>)$^{(0,1)}$) TPLDTU ((rel$^{(1,1)}$<EN>W^{+}</EN>)$^{+}$|<EN>rel2$^{(0,1)}$W^{+}</EN>)*</AREA_1>

« 10 Rockfeller Center », « Mobito Plaza », « Centre Morhiarty »

d)
((LEX_bp* <BP>(NUM|CP|C)</BP>)* (LEX_cedex <CEDEX>(C2|Ø)</CEDEX>)* LEX_cp* (<CODE_POSTAL>(CP|C)</CODE_POSTAL>|<CODE>C</CODE>) LEX_ann_ville* (<EN_ville>LEX_ville</EN_ville>|<EN>W^{+}</EN>) (LEX_cedex <CEDEX>(C2|Ø)</CEDEX>)* ((LEX_ann_etat|LEX_ann_region)* (<EN_etat>LEX_etat</EN_etat>|<EN_region>LEX_region</EN_region>|<EN>W^{+}</EN>)) (LEX_ann_pays* (<EN_pays>LEX_pays</EN_pays>|<EN>W^{+}</EN>))*
|
(LEX_bp* <BP>(NUM|CP|C)</BP>)* LEX_ann_ville* (<EN_ville>LEX_ville</EN_ville>|<EN>W^{+}</EN>) (LEX_cp* <CODE_POSTAL>(CP|C)</CODE_POSTAL>|<CODE>C</CODE>)$^{(0,1)}$ ((LEX_ann_etat|LEX_ann_region)* (<EN_etat>LEX_etat</EN_etat>|<EN_region>LEX_region</EN_region>|<EN>W^{+}</EN>)) (LEX_cp* <CODE_POSTAL>(CP|C)</CODE_POSTAL>|<CODE>C</CODE>))$^{(0,1)}$ (LEX_ann_pays* (<EN_pays>LEX_pays</EN_pays>|<EN>W^{+}</EN>))*)
d') [anglais]
(LEX_bp* <BP>(NUM|CP|C)</BP>)* LEX_ann_ville* (<EN_ville>LEX_ville</EN_ville>|<EN>W^{+}</EN>) (LEX_cp* <CODE_POSTAL>(CP|C)</CODE_POSTAL>|<CODE>C</CODE>)$^{(0,1)}$ ((LEX_ann_etat|LEX_ann_region)* (<EN_etat>LEX_etat</EN_etat>|<EN_region>LEX_region</EN_region>|<EN>W^{+}</EN>)) (LEX_cp* <CODE_POSTAL>(CP|C)</CODE_POSTAL>|<CODE>C</CODE>))$^{(0,1)}$ (LEX_ann_pays* (<EN_pays>LEX_pays</EN_pays>|<EN>W^{+}</EN>))*

e) <SSEN>LEX_ssen (EN>W^{n}</EN>)* (rel (w^{7}|<EN>W^{n}</EN>))*</SSEN>

f) <TEL>+$^{(0,1)}$ NUM</TEL>

g) <FAX>+$^{(0,1)}$ NUM</FAX>

6.2.3. L'automate de reconnaissance

L'automate de reconnaissance s'applique à une chaîne écrite non pas telle que l'écrirait un humain avec toutes ses variations mais à une chaîne prétraitée où les « blancs » du type « espace », quand ils sont multiples ou combinés à des tabulations, sont réduits à un seul espace, et où les chiffres sont formatés de manière à former un bloc continu !

Tableau 2 : Chaîne de règles construisant la reconnaissance

1	2	3	4	5	6	7
ORDINAL	CODE	CODE_POSTAL	NUM	CEDEX	BP	FAX
R15→R18→	R34→ R35	R36→R38→R37→ R38→	R9→R11^{+} →R14→	R19→	R41 →	R21→ R25→ R26
[14]Idem	Idem	Idem	Idem	Idem	R42	Idem

8	9	10	11	12	13
TEL	CIVILITES	PROFESSION	SSEN	AREA_0	AREA_1
R20→ R22→ R23→R24→	R33 →	(R31→)$^{+}$	(R29→)$^{+}$	(R43→)R44 →	(R47→R48)
Idem	Idem	Idem	Idem	Idem	Idem

14	15	16	17	18	19
EN	Fin ORDINAL	Fin PROFESSION	Fin SSEN	Fin AREA_0	Fin AREA_1
R27→R28 →	Facultatif R16→R20	R32	R30→	R45→	R49→
Idem	Idem	Idem	Idem	R46	R50→

20
Nettoyage de la chaîne textuelle

[14] Ligne des règles traitant l'anglais

R5
Idem

Une résolution alternative existe. Elle consiste à effectuer une reconnaissance sur les entités nommées et à identifier celles qui entrent dans la catégorie pays, ville, région et état. On insère, pour cela, la succession suivante, après la position 14 :

Tableau 3 : Résolution alternative

14A	14B	14C	14D
EN_VILLE	EN_Region	EN_etat	EN_pays
R53→	R52→	R51→	R54

6.2.4. Les motifs d'extraction

6.2.4.1. Du français

(0|1) a+ e* c* $d^{(0,1)}$ f* g*

(0|1) e+ a* c* $d^{(0,1)}$ f* g*

(0|1) c^+ (a + a e^+ + e + e^+a^+) $d^{(0,1)}$ f* g*

(0|1) c^+ d f* g*

(0|1) a c e $d^{(0,1)}$ f* g*

Sachant que chacun des motifs précédents admet au moins une permutation supplémentaire dans laquelle les elements f et g apparaissent en tête de motif, on a donc aussi:

f^+ g* (0|1) a+ e* c* $d^{(0,1)}$

f^+ g* (0|1) e+ a* c* $d^{(0,1)}$

f^+ g* (0|1) c^+ (a + a e^+ + e + e^+a^+) $d^{(0,1)}$

f^+ g* (0|1) c^+ d*

f^+ g* (0|1) a c e $d^{(0,1)}$

Sachant que certaines entités nommées sont formées autour des têtes lexicales répertoriées dans LEX_ssen, LEX_area1, on a donc encore les motifs 2, 3, 4 et 7, 8, 9 qui deviennent:

e^+ a* c* $d^{(0,1)}$ f* g*

c^+ (a + a e^+ + e + e^+a^+) $d^{(0,1)}$ f* g*

c^+ d f* g*

et

f^+ g* (0|1)* e+ a* c* $d^{(0,1)}$ f* g*

f^+ g* (0|1)* c^+ (a + a e^+ + e + e^+a^+) $d^{(0,1)}$ f*

f^+ g* (0|1)* c^+ d*

6.2.4.2. La reconnaissance de l'anglais

La reconnaissance des contacts en langue anglaise pose des problèmes majeurs de syntaxe qui ne permettent pas de réutiliser purement et simplement le modèle réalisé pour le français, sans entraîner des conflits majeurs: la qualification en anglais rejette la reconnaissance en avant-garde du nom qualifié. Exemple par excellence, le nom d'une rue apparait avant la mention de sa qualité de rue. Même chose pour la mention des boîtes postales. Les résultats seraient alors absolument aléatoires, même s'ils peuvent inférer un bon niveau de reconnaissance, puisque l'ordre de la langue française est également représenté en langue anglaise.

Plusieurs pistes existent à cette étape du problème : développer une chaîne de reconnaissance précise spécifique à l'anglais et qui pré-supposera que l'on aura identifié le texte analysé comme étant de l'anglais, ou mettre en place une dérivation dans la reconnaissance en fonction de la langue de la liste activée dans l'utilisation des mots clefs et têtes lexicales (du type LEX_x). Suivant cette liste, les règles spécifiques de la langue qui ne constituent pas un tronc commun seront mises en œuvre de façon privilégiée. Cette méthode a l'avantage certain d'augmenter les chances d'une reconnaissance correcte et véritablement en contexte. En effet, une adresse qui figure dans une page web n'est pas forcément une adresse aux normes de la langue de la page[15].

La combinatoire des sous motifs d'extraction pour l'anglais (extraction partielle de l'évènement extraction de contacts) est identique au français à ceci près que a devient a', c devient c', d devient d'.

(0|1) a'+ e* c'* $d'^{(0,1)}$ f* g*

(0|1) e+ a'* c'* $d'^{(0,1)}$ f* g*

(0|1) c^+ (a' + a e^+ + e + e^+ a'^+) $d'^{(0,1)}$ f* g*

(0|1) c'^+ d' f* g*

[15]Voir Annexe 12, exemple 1, commentaire.

$(0|1)$ $a'\ c'\ e$ $d'^{(0,1)}$ f^* g^*

e^+ a'^* c'^* $d'^{(0,1)}$ f^* g^*

c'^+ $(a' + a'\ e^+ + e + e^+a'^+)$ $d'^{(0,1)}$ f^* g^*

c'^+ $d'\ f^*\ g^*$

Les permutations de f et g induisent en plus:

$f^+\ g^*$ $a+$ e^* c^* $d'^{(0,1)}$ $f^+\ g^*$

$(0|1)$ $e+a'^*$ c'^* $d'^{(0,1)}$

$f^+\ g^*\ (0|1)$ $c'^+\ (a' + a'\ e^+ + e + e^+a'^+)\ d'^{(0,1)}$

$f^+\ g^*\ (0|1)$ $c'^+\ d'^*$

$f^+\ g^*\ (0|1)$ $a'\ c'\ e\ d^{(0,1)}$

$f^+\ g^*\ (0|1)^*$ $e + a'^* c'^* d'^{(0,1)} f^* g^*$

$f^+\ g^*\ (0|1)^*$ $c'^+\ (a' + a'\ e^+ + e' + e^+a'^+)\ d'^{(0,1)} f^* g^*$

$f^+\ g^*\ (0|1)^*$ $c'^+\ d'$

Illustration 8 : Combinatoire des sous-motifs traitant le Français majoritairement.

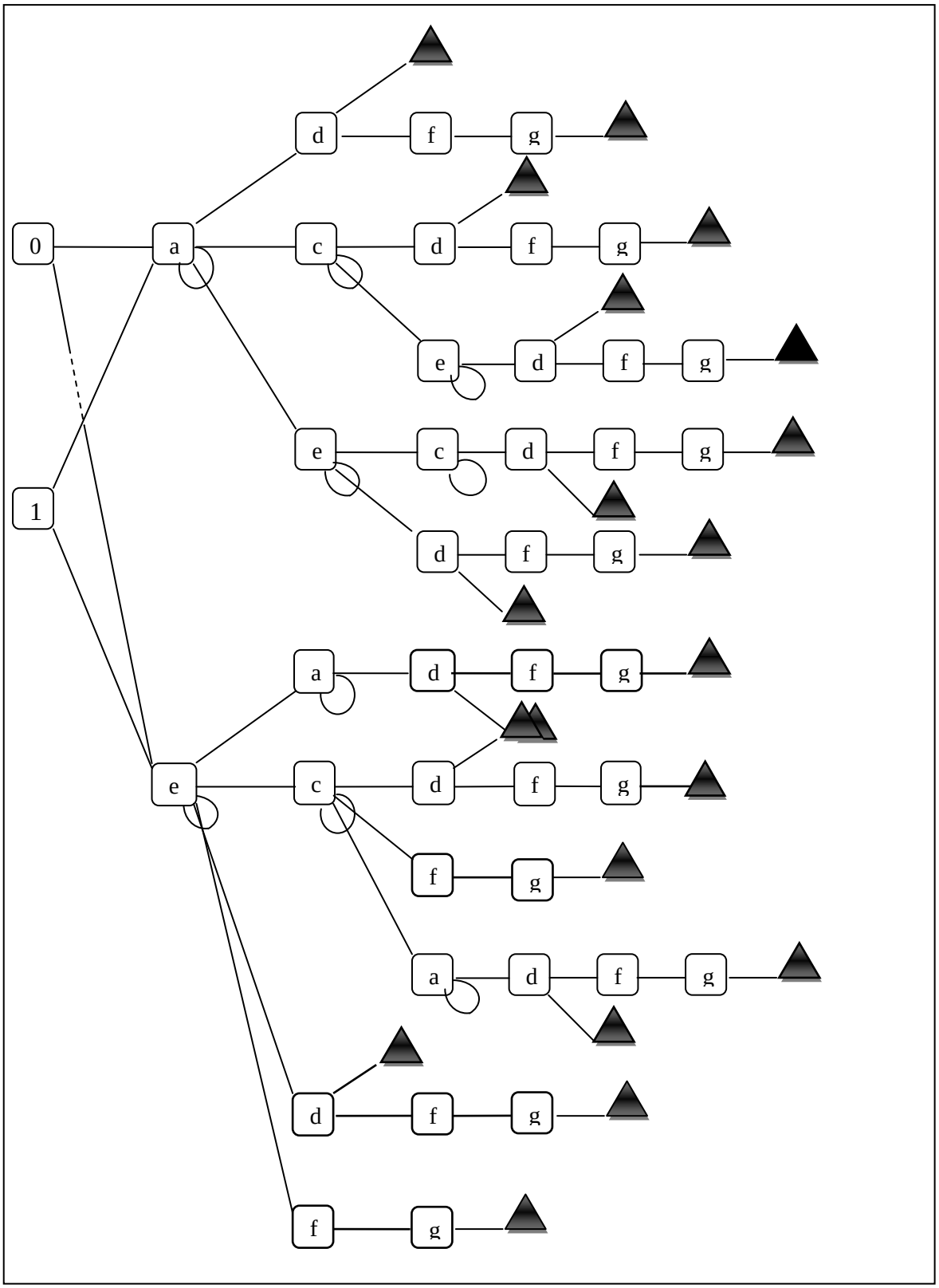


Illustration 8 : Combinatoire des sous-motifs traitant le Français majoritairement (suite).

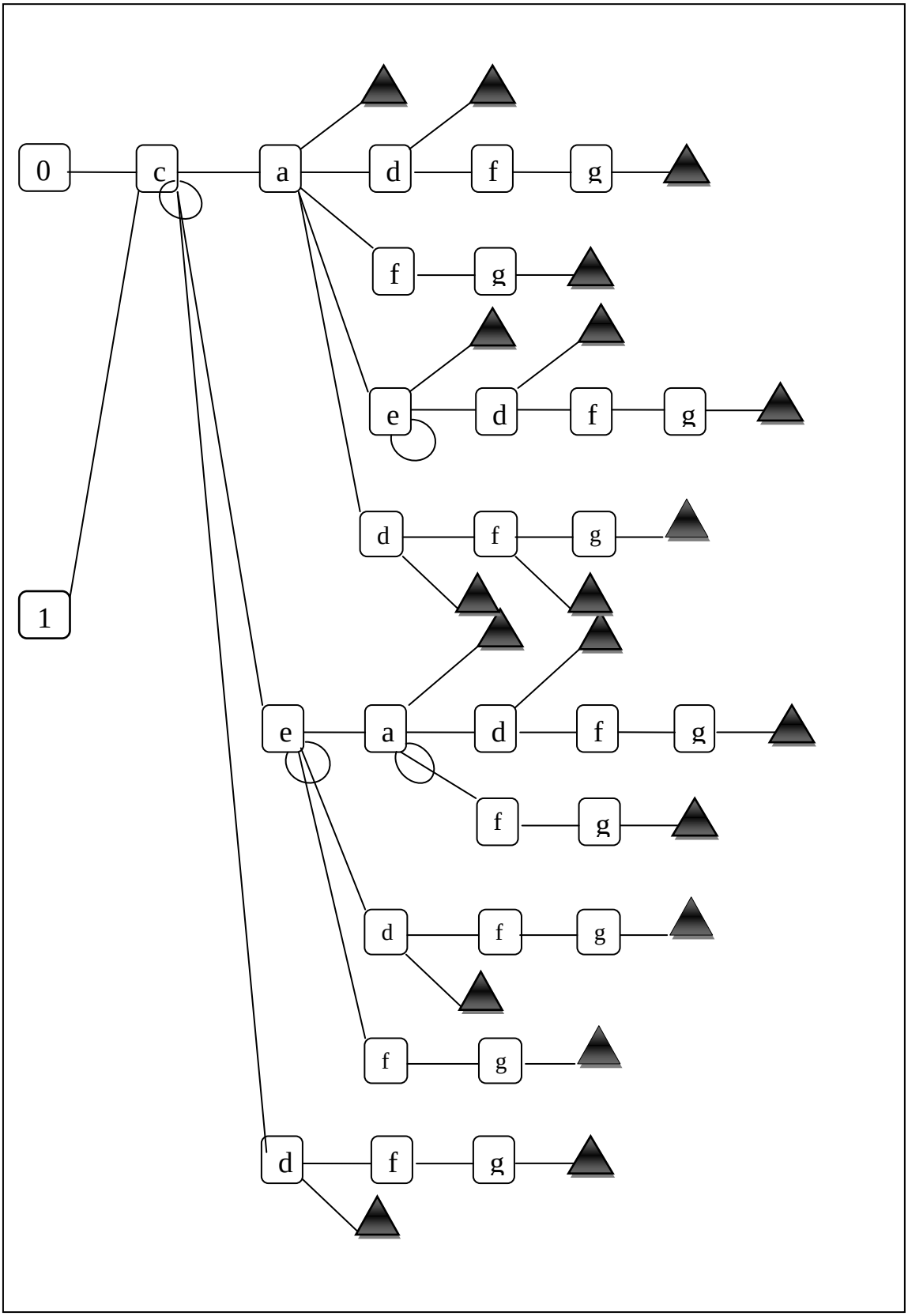
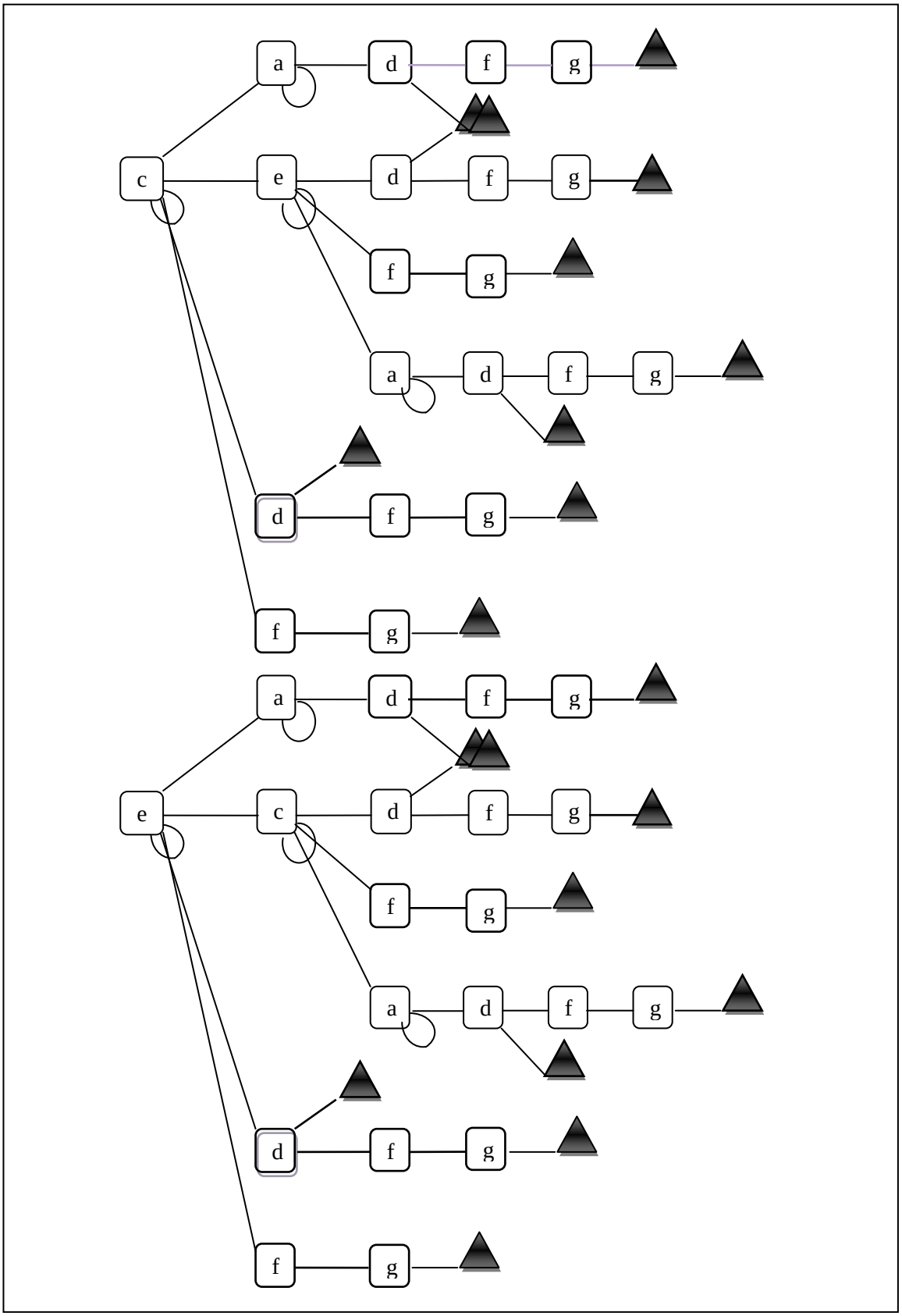


Illustration 8 : Combinatoire des sous-motifs traitant le Français majoritairement (suite).



Légende 1 :

 Fin du motif

6.3. La grammaire en œuvre : quelques exemples

6.3.1. Exemple 1 :

Hervé Morel
Consultant Financier
23 rue Maréchal Joffre
(4^{e}étage) porte 1
07013 Paris
| Paris 13(e)* → Paris 13
Paris 13e

(R15→)R18→

Hervé Morel

Consultant Financier

23 rue Maréchal Joffre

(<ORDINAL>4^{e}</ORDINAL >étage) porte 1

07013 Paris

| Paris 13(e)* → Paris 13

Paris < ORDINAL >13e</ORDINAL >

(R36→)R38→(R37→)R38→

Hervé Morel

Consultant Financier

23 rue Maréchal Joffre

(<ORDINAL>4^{e}</ ORDINAL >étage) porte 1

<CP>07013</CP> Paris

| Paris 13(e)* → Paris 13

Paris < ORDINAL >13e</ ORDINAL >

(R9→R11^{+})→R14→

Hervé Morel
Consultant Financier
<NUM>23</ NUM > rue Maréchal Joffre
(<ORDINAL>4^{e}</ ORDINAL >étage) porte < NUM >1</ NUM >
<CP>07013</CP> Paris
| Paris 13(e)* → Paris < NUM >13</ NUM >
Paris < ORDINAL >13e</ ORDINAL >

R31 →

Hervé Morel
<PROFESSION>Consultant</PROFESSION > Financier
<NUM>23</ NUM > rue Maréchal Joffre
(<ORDINAL>4^{e}</ ORDINAL > étage) porte <NUM >1</ NUM >
<NUM>07013</NUM> Paris
| Paris 13(e)* → Paris < NUM >13</ NUM >
Paris < ORDINAL >13e</ ORDINAL >

(R43→)R44→

Hervé Morel
<PROFESSION>Consultant</PROFESSION > Financier
<NUM>23</ NUM > <AREA_0>rue< /AREA_0> Maréchal Joffre
(<ORDINAL>4^{e}</ ORDINAL > <AREA_0>_étage</AREA_0>)
<AREA_0>porte</AREA_0> <NUM >1</ NUM >
<NUM>07013</NUM> Paris
| Paris 13(e)* → Paris < NUM >13</ NUM >
Paris < ORDINAL >13e</ ORDINAL >

R27→R28→

<EN>Hervé Morel</EN>

<PROFESSION>Consultant</PROFESSION> <EN>Financier</EN>

<NUM>23</ NUM > <AREA_0>rue< /AREA_0> <EN>Maréchal Joffre</EN>
(<ORDINAL>4e</ORDINAL > <AREA_0>_étage</AREA_0>
<AREA_0>porte</AREA_0> <NUM>1</ NUM >
<CODE_POSTAL>07013</CODE_POSTAL> <EN>Paris</EN>

| Paris 13(e)* → <EN>Paris</EN> < NUM >13</ NUM >
<EN>Paris</EN> < ORDINAL >13e</ ORDINAL >

R32→

<EN>Hervé Morel</EN>
<PROFESSION>Consultant <EN>Financier</EN></PROFESSION>
<NUM>23</ NUM > <AREA_0>rue< /AREA_0> <EN>Maréchal Joffre</EN>
(<ORDINAL>4e</ORDINAL > <AREA_0>_étage</AREA_0>
<AREA_0>porte</AREA_0> </AREA_0>
<CODE_POSTAL>07013</CODE_POSTAL> <EN>Paris</EN>

| Paris 13(e)* → <EN>Paris</EN> < NUM >13</ NUM >
<EN>Paris</EN> < ORDINAL >13e</ ORDINAL >

R45→

<EN>Hervé Morel</EN>
<PROFESSION>Consultant <EN> Financier</EN></PROFESSION>
<AREA_0><NUM>23</ NUM > rue <EN>Maréchal Joffre</EN></AREA_0>
(<AREA_0><ORDINAL>4e</ ORDINAL> _étage</AREA_0>
<AREA_0>porte<NUM>1</ NUM > </AREA_0>
<CODE_POSTAL>07013</CODE_POSTAL> <EN>Paris</EN>

| Paris 13(e)* → <EN>Paris</EN> < NUM >13</ NUM >
<EN>Paris</EN> < ORDINAL >13e</ ORDINAL>

R5→ (nettoyage de la chaîne textuelle)

Soit la ligne de traitement vérifiée:

[(R15→)R18→(R36→)R38→(R37→)R38→(R9→R11^{+})→R14→ R31→ (R43→)R44→ R27→R28→ R32→ R45→ R5→]

Soit le motif d'extraction validé : 0 a^{+} d

6.3.2. Exemple 2

Hélène Grand-Gens Cité de la CAF Services comptables 4 rue des Alliés 38 000 grenoble

R36→ R38→

Hélène Grand-Gens

Cité de la CAF

Services comptables

4 rue des Alliés

<CP>38 000</CP> Grenoble

(R9→R11^{+}→) R13→ [R36→ R38→(R9→R11^{+}→) R13→ R14

Hélène Grand-Gens

Cité de la CAF

Services comptables

<NUM>4</NUM> rue des Alliés

<CP>38 000</CP> Grenoble

R29→ [R36→ R38→(R9→R11^{+}→) R13→ R14→ R29

Hélène Grand-Gens

Cité de la CAF

<SSEN>Services</SSEN> comptables

<NUM>4</NUM> rue des Alliés

<CP>38 000</CP> Grenoble

R43→R44→ [R36→ R38→(R9→R11^{+}→) R13→ R14→ R29→ R43→R44

Hélène Grand-Gens
Cité de la CAF
<SSEN>Services</SSEN> comptables
<NUM>4</NUM> <AREA_0>rue</AREA_0> des Alliés
<CP>38 000</CP> Grenoble

(R47→)R48→
[R36→ R38→(R9→R11^{+}→) R13→ R14→ R29→ R43→R44→(R47→)R48

Hélène Grand-Gens
< AREA_1>Cité</AREA_1> de la CAF
<SSEN>Services</SSEN>
<NUM>4</NUM> <AREA_0>rue</AREA_0> des Alliés
<CP>38 000</CP> Grenoble

R27→R28→
[R36→ R38→(R9→R11^{+}→) R13→ R14→ R29→ R43→R44→(R47→)R48
→ R27→R28
<EN>Hélène Grand-Gens</EN>
<AREA_1>Cité</AREA_1> de la <EN>CAF</EN>
<SSEN>Services</SSEN> comptables
<NUM>4</NUM> <AREA_0>rue</AREA_0> des <EN>Alliés</EN>
<CP>38 000</CP> <EN>Grenoble</EN>

R30→
[R36→ R38→(R9→R11^{+}→) R13→ R14→ R29→ R43→R44→(R47→)R48
→ R27→R28→R30

<EN>Hélène Grand-Gens</EN>
<AREA_1>Cité</AREA_1> de la <EN>CAF</EN>

<SSEN>Services comptables</SSEN>
<NUM>4</NUM> <AREA_0>rue</AREA_0> des <EN>Alliés</EN>
<CP>38 000</CP> <EN>Grenoble</EN>

R43→R44→
[R40→ R42→ R41→R42→(R8→R10→)R13→
R33→R46→R47→(R50→)R51→R31→R34→R48→ R27→R28→R30→
R43→R44

<EN>Hélène Grand-Gens</EN>
<AREA_1>Cité</AREA_1> de la <EN>CAF</EN>
<SSEN>Services comptables</SSEN>
<NUM>4</NUM> <AREA_0>rue des <EN>Alliés</EN></AREA_0>
<CP>38 000</CP> <EN>Grenoble</EN>

R49→R50→

<EN>Hélène Grand-Gens</EN>
<AREA_1>Cité de la <EN>CAF</EN></AREA_1>
<SSEN>Services comptables</SSEN>
<NUM>4</NUM> <AREA_0>rue des <EN>Alliés</EN></AREA_0>
<CP>38 000</CP> <EN>Grenoble</EN>
R5→ (nettoyage de la chaîne textuelle)

Soit la chaîne de règles activées :
[R40→R42→R41→R42→(R8→R10→)R13→R33→R46→R47→(R50→)R51→R31→R34→R48→R27→R28→R30→ R43→R44→ R49→R50→R5]

Soit le motif d'extraction validé : 0 c e a d.

6.4. Discussion du modèle

Le modèle où seules les entités nommées sont identifiées a ses avantages. Il permet à priori de mettre en place une chaîne de collecte sur des noms mais aussi sur des variations, à l'intérieur des contraintes de casse. Celui du second modèle est d'identifier à

coup sûr les entités nommées pour ce qu'elles sont : des pays, des villes etc, abstraction faites des ambiguïtés des noms partagés par les différentes catégories et pourvu que, là aussi, la variation ait été répertoriée : variation de forme, de notation, de dénomination particulièrement. La résolution alternative a l'avantage supplémentaire de profiter du filtre d'identification en <EN></EN>, qui a pour effet mécanique de récupérer ce qui n'aura pas été indexé à l'étape précédente.

La nécessité de comparer les entrées du texte au contenu exhaustif de 4 listes introduit quelques paramètres supplémentaires à surveiller tels l'optimisation du tri pour limiter l'accroissement peut-être significatif des temps de traitement conduisant à l'extraction.

La grammaire précédente telle qu'elle a été définie laisse largement la place à l'amélioration. Amélioration deu modèle destiné à l'extraction en anglais notemment. En effet la reconnaissance des groupes SSEN ne prend pas encore en charge la syntaxe inversée (par rapport au français de l'anglais) pour la qualification des noms par exemple. Une fois lz modèle optimisé pour les deux langues, l'ordre dans lequel se fera la reconnaissance sera primordial. De lui dépendra les chances d'une reconnaissance opportune, sans que pour autant les ambigüités de reconnaissance soient complètement écartées.

En affinant et en étendant les lexiques, le modèle défini pour l'extraction de contacts pourra être étendu pour les langues déjà prises en charge et activé pour les autres langues du projet que sont l'espagnol, l'italien, par implémentation des lexiques. Ces langues, en effet, ne s'écartent pas fondamentalement de ce modèle contrairement à allemand qui avec son phénomène d'agglutination exigera une reconnaissance particulière. Même chose pour le chinois, avec une phénoménologie différente, cependant.

Chapitre 7 – L'extraction d'articles de presse

7.1. Analyse contextuelle

Les dispositions typiques d'un article ou plutôt d'une annonce d'article sont les suivantes :

- Le titre ou l'accroche consiste toujours en une phrase, soit 4 à 24 mots (critères larges qui ne se vérifient pas toujours, il est vrai) mais qui sont toujours mis en

valeur par un système de balises, ce qui est le plus important. Balise de lien, de paragraphe, de mise en forme, mais balise toujours !

Pour illustrer ce propos, voici quelques exemples des mises en page rencontrées:

« L'activité bancaire et Les Assurances, seuls secteurs de l'économie sn progression » est un titre possible

- L'activité bancaire et les Assurances, seuls secteurs de l'économie en progression
- L'activité bancaire et les Assurances, seuls secteurs de l'économie en progression
- <a>L'activité bancaire et les Assurances, seuls secteurs de l'économie en progression
- L'activité bancaire et Les Assurances, seuls secteurs de l'économie sn progression
- L'activité bancaire et Les Assurances, seuls secteurs de l'économie sn progression

Ce qui fait la stabilité de ce type de déclaration, c'est la nécessité de mettre en valeur, l'information principale -le titre- à destination de l'humain. Or cette action n'est possible en HTML qu'avec l'aide de balises et plus significativement, qu'avec les deux opérateurs diamant <,>.

- Les premières lignes de l'article qui ne font pas l'objet d'une reconnaissance.
- Un lien vers le contenu plein de l'article, annoncé par un attribut « href ». De la variation existe dans l'occurrence de cette structure. En effet, le lien n'existe pas systématiquement sur le titre mais alors est présent sur un texte récurrent du type « A lire », « La suite ici », « Ici », « Lire l'article », « Voir l'article » etc…

Afin de supporter les variantes dans l'enchaînement des informations propres aux infra langages, des éléments de contraintes relativement lâches ont été arrêtés pour la rétention des blocs d'informations potentiellement pertinents. Si la consécution minimale respectant les conditions évoquées plus haut n'est pas réunie, le bloc n'est plus candidat à l'extraction. Au contraire, une fois le bloc qualifié, on lui applique les motifs d'extractions.

7.2. Définition formelle : grammaires et motifs d'extraction

Dans le cas singulier de l'extraction d'articles de presse parus sur internet, un nombre limité de dictionnaires est requis, la reconnaissance se basant essentiellement sur la structure des énoncés.

7.2.1. Lexique et vocabulaire de la grammaire

Tableau 4 : Lexiques de l'extraction d'article

Ressources de type dictionnaire	Code
La ressource rassemble l'ensemble des suites lexicales, textes signalant un lien vers un contenu complet d'article.	LEX_introductif
Lexique de quatre à 5 mots (regroupe toutes les suites linguistiques outils indiquant une redirection type « Votre devis en ligne », « Une agence près de chez vous », « Découvrez nos solutions santé » etc…	LEX_invalid
Liste des mois en français et anglais	LEX_mois

L'entrée de la chaîne prototypique est le code html délivré par le crawler. Ce code, est pré-traité de sorte à simplifier les balises html présentes dans le texte, en n'en préservant que les attributs utiles, soit l'attribut href, des balises <A>. Pour le reste, le prétraitement reprend globalement les opérations de la reconnaissance précédente : simplification des « blancs » multiples en un unique espace simple, sans épargne des retours à la ligne cependant. Les sorties de l'extraction sont consignées sous forme d'XMLs.

Le vocabulaire de base de cette grammaire d'extraction est le suivant :

Tableau 5 : Vocabulaire de la grammaire

Libellé	Vocabulaire utilisé
Le titre de l'article	<TITRE_article></TITRE_article>
La date de l'article	<DATE> </DATE>
L'adresse de l'article entier	<URL_article></URL_article>
L'adresse de l'article entier (fortes présomptions)	<URLF_article></URLF_article>

7.2.2. La grammaire : les règles

Les règles de cette grammaire s'énoncent comme suit :

R 1 : Ø→ mot vide, blancs (tabulations, espace, simples et multiples) composés de blancs

R 2 : x→w| W (mots non vides)

R 3 : x′→w′|Ẇ′ (mots non vides qui ne contiennent ni l'opérateur <, ni l'opérateur>)

R 4 : x″→w′|Ẇ′ (mots non vides qui ne contiennent ni l'opérateur <, ni l'opérateur >, ni ′, ni ″)

Sachant que dans ces définitions, les nombres comptent pour des mots.

R 5 : y → >x′^{4,26}<

R 6 : NUM →[0-9]

R 7 : NUM2 →[0-9]

R 8 : NUM →NUM Ø NUM

R 9 : NUM →NUM NUM

R 10 : (Href|href|HREF) Ø = (′*|″) Ø x″(′*|″)Ø y → <URL_article>x</URL_article>

R 11 : (Href|href|HREF) Ø = Ø x Ø> Ø LEX_introductif Ø < → <URLF_article>x</URLF_article>

R 12 : y → <TITRE_article >y</TITRE_article>

R 13 : z → (((0$^{(0,1)}$[1-9]|[1-2]NUM2|3[01]) [-/Ø.])$^{(0,1)}$ ([0]$^{(0,1)}$ [1-9]|1[12]|Lex_mois) ([-/Ø.] ([1-9]^{2}| (19 NUM2^{2}|200NUM2|201[0-3]) | ((0$^{(0,1)}$[1-9]|[1-2]NUM2|3[01]) [-/Ø.]) ([0]$^{(0,1)}$ [1-9]|1[12]|Lex_mois) | (19 NUM2^{2}|200NUM2|201[0-3]) [-/Ø.,])$^{(0,1)}$ (([0]$^{(0,1)}$ [1-9]|1[12]|LEX_mois) [-/Ø.]) ([-/Ø.]0$^{(0,1)}$[1-9]|[1-2]NUM2|3[01]) | (([0]$^{(0,1)}$ [1-

9]|1[12]|LEX_mois) [-/Ø.]) (0$^{(0,1)}$[1-9]|[1-2]NUM2|3[01]) [-/Ø.,]) (19 NUM2^{2}|200NUM2|201[0-3])$^{(0,1)}$)

R 14 : Ø z → Ø <DATE>z</DATE>[Ø PUNCT]

R15 : (<TITRE_article>y<TITRE_article>) Ø <DATE>z</DATE> (Ø <TITRE_article>s<TITRE_article>)* → (<TITRE_article>y Ø z Ø s*<TITRE_article>)*

R16 : <DATE>z</DATE> (Ø <TITRE_article>y<TITRE_article>)* → (<TITRE_article>z Ø y<TITRE_article>)*

R 17 : < /*¬ (DATE|TITRE_article|URL_article)> → Ø

R18 : PUNCT →[-?./§, ;:!)(>'< » «'']

L'automate permettant de réaliser la reconnaissance préalable à l'extraction étant :

R13→R14→	R5→R12→	R15→R19→	R11→	R10→	R17→
Reconnaissance des dates	Reconnaissance des titres	Gestion des inclusions de dates à l'intérieur des titres et à leurs marges (internes et externes)	Reconnaissance des cibles des hyperliens spécifiques des textes introductifs.	Reconnaissances des cibles de tous les autres hyperliens	Effacement de toutes les balises n'appartenant pas à la grammaire et différentes de <A></A>

7.2.3. Motifs d'extraction

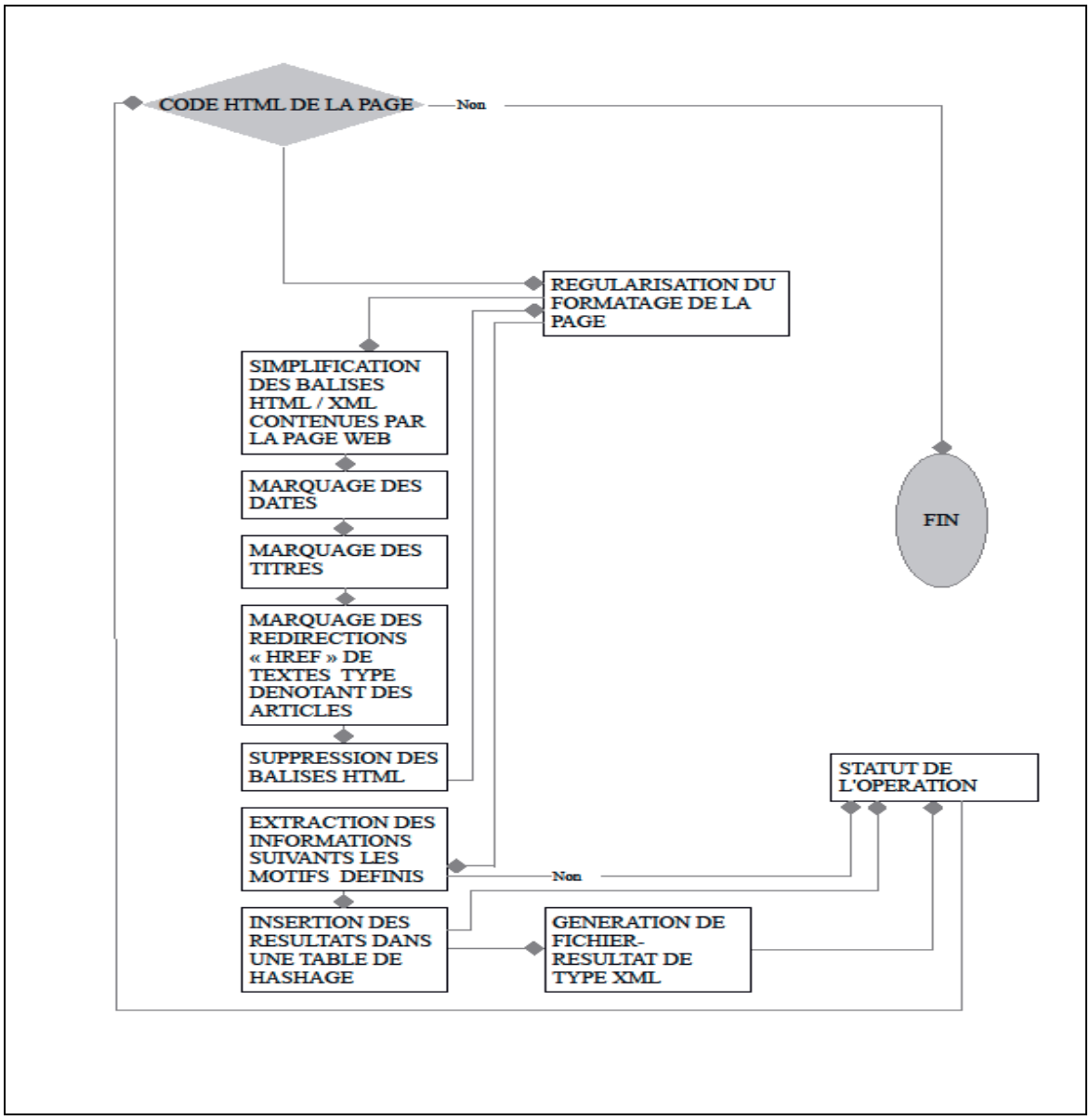
1)

([- :/.Ø]$^{+}$<DATE>z</DATE> [- :/.Ø])$^{(0,1)}$ (<URL_article>x</URL_article>[- :/.Ø]$^{+}$(<DATE>z</DATE> [- :/.Ø]$^{+}$ (</(A|a)> [- :/.Ø]$^{+}$)* (<URL_article>x</URL_article>[- :/.Ø])* <TITRE_article> z Ø y</TITRE_article> (([- :/.Ø]* (<DATE>z</DATE>)) (<URL_article>x</URL_article> [- :/.Ø]$^{+}$<DATE>z</DATE> [- :/.Ø] </(A|a)>)$^{(0,1)}$ ([- :/.Ø]$^{+}$<DATE>z</DATE> [- :/.Ø])$^{(0,1)}$ ([- :/.Ø]$^{+}$ (W'|w')$^{(1,29)}$ […]$^{(0,1)}$ <URLF_article>x</URLF_article>)* ([- :/.Ø]$^{+}$<DATE>z</DATE>])$^{(0,1)}$

2)

(<DATE>z</DATE>[- :/.Ø]$^{+}$) <TITRE_article> z Ø y</TITRE_article> ([- :/.Ø]* (<DATE>z</DATE> [- :/.Ø])* (<URL_article>x</URL_article> [- :/.Ø]$^{+}$<DATE>z</DATE> [- :/.Ø] </(A|a)>)$^{(0,1)}$ ([- :/.Ø]$^{+}$<DATE>z</DATE> [- :/.Ø])$^{(0,1)}$ ([- :/.Ø]$^{+}$ (W'|w')$^{(1,29)}$ […]$^{(0,1)}$ <URLF_article>x</URLF_article>)$^{(0,1)}$ ([- :/.Ø]$^{+}$<DATE>z</DATE> [- :/.Ø])$^{(0,1)}$

Illustration 9 : Schématisation du processus d'extraction des articles



7.3. Mise en œuvre de la grammaire et critique du modèle

7.3.1. Exemple

Illustration 10 : Extrait d'une page web traitée en extraction[16]
(http://www.agefi.fr/fiche-actualite-eclairages/solvabilite-ii-la-pression-monte-144.html)

<h3><a href=<**URL_article**>"../articles/solvabilite-2-pourrait-provoquer-le-retrait-des-assureurs-de-la-titrisation-1249478.html"</**URL_article**><**TITRE_article**>Solvabilité 2 pourrait provoquer le retrait des assureurs de la titrisation</**TITRE_article**></h3> 1

<**TITRE_article**>Selon Standard & Poor's, ceux-ci commencent déjà à se tourner vers les obligations sécurisées, plus rémunératrices car moins pénalisées</**TITRE_article**> 2

<**DATE**>12/12/2012</**DATE**> dans L'AGEFI Quotidien - Edition de 7H

Solvabilité 2, toujours en cours de négociation, pourrait faire sentir ses effets bien au-delà du secteur de l'assurance et affecter le marché de la titrisation. Se rapportant à la dernière ébauche te...

   <h2></h2> 3

<h3><a href=<**URL_article**>"../articles/les-assureurs-francais-envoient-un-message-fort-sur-leur-investissement-en-actions-1240752.html"</**URL_article**><**TITRE_article**>Les assureurs français envoient un message fort sur leur investissement en actions</**TITRE_article**></h3>

<**TITRE_article**>CNP, Predica, Cardif et Sogecap créent leur Fonds stratégique en pleine réflexion sur la fiscalité de l'épargne longue et l'évolution de l'assurance vie</**TITRE_article**> 4

<**DATE**>10/10/2012</**DATE**> dans L'AGEFI Quotidien - Edition de 7H

A l'heure où le gouvernement a engagé une réflexion sur la fiscalité de l'épargne longue qui devrait affecter l'assurance vie, l'initiative tombe à pic. La création d'un Fonds stratégique de participa...

   <h2></h2>

<h3><a href=<**URL_article**>"../articles/les-assureurs-maintiennent-leur-confiance-dans-le-private-equity-1227863.html "</**URL_article**><**TITRE_article**>Les assureurs maintiennent leur confiance dans le private equity</**TITRE_article**></h3> 5

<**TITRE_article**>Près de 90% d'entre eux entendent investir dans cette classe d'actifs. Solvabilité 2 semble surtout préoccuper les groupes européens</**TITRE_article**> 6

<**DATE**>15/06/2012<**DATE**> dans L'AGEFI Quotidien - Edition de 7H

Alors que les banques font preuve d'une grande prudence face aux échéances réglementaires qui les attendent, les groupes d'assurances semblent dans l'ensemble plus sereins dans leur allocation à l'éga...

   <h2></h2> 7

<h3><a href=<**URL_article**>"../articles/les-titrisations-risquent-d-etre-fortement-penalisees-par-solvabilite-2-1218687.html"</**URL_article**><**TITRE_article**>Les titrisations risquent d'être fortement pénalisées par Solvabilité 2</**TITRE_article**></h3>

<**TITRE_article**>Alors que les exigences en capital liées aux ABS sont élevées, le lobby financier AFME espère encore infléchir la position des législateurs< /**TITRE_article**> 8

<**DATE**>04/04/2012</**DATE**> dans L'AGEFI Quotidien - Edition de 7H

Les acteurs du marché de la titrisation en Europe se préoccupent des règles qui vont être appliquées aux assureurs. De la même manière qu'ils redoutent les contraintes prudentielles de Bâle III qui vo...

   <h2></h2>

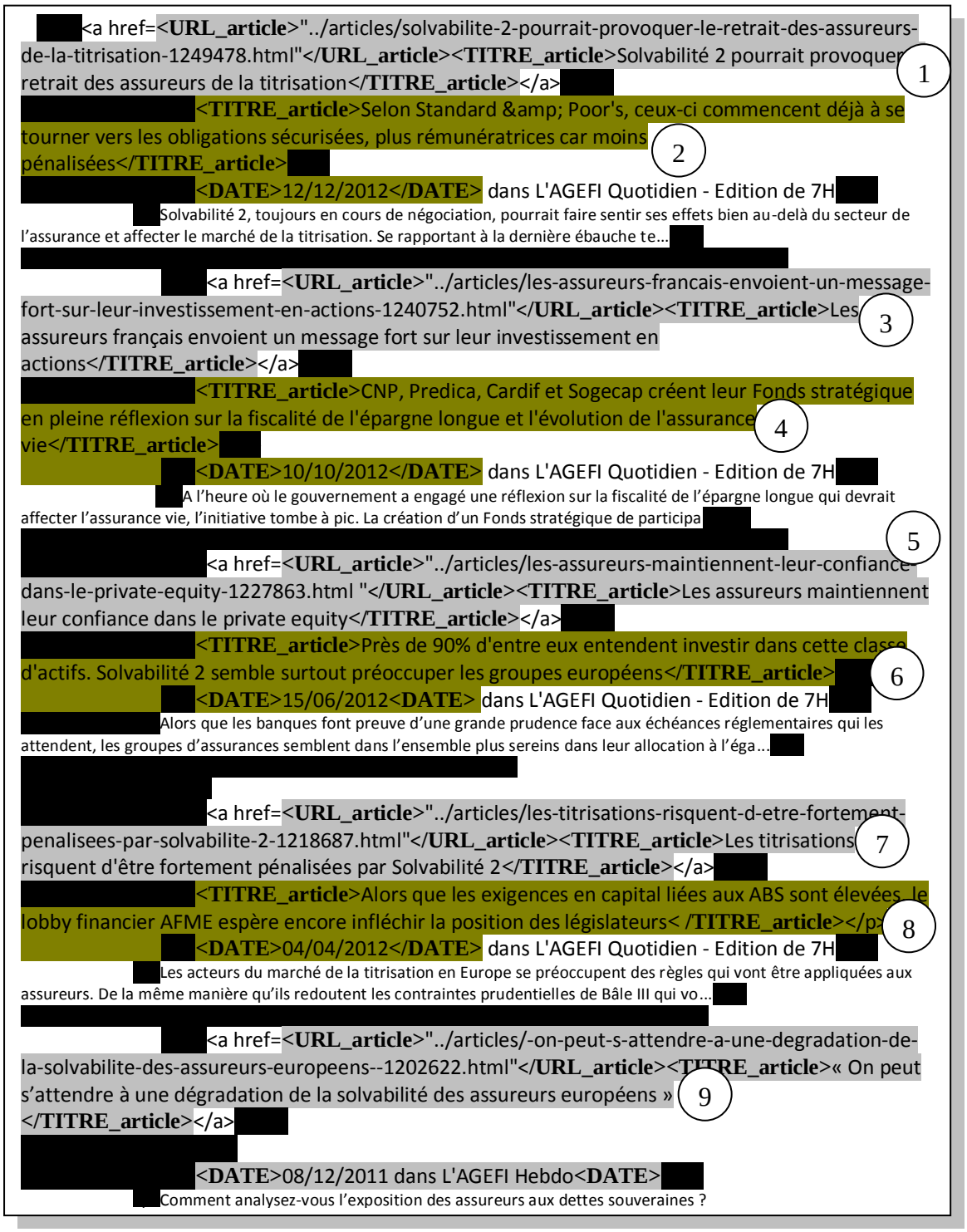
<h3><a href=<**URL_article**>"../articles/-on-peut-s-attendre-a-une-degradation-de-la-solvabilite-des-assureurs-europeens--1202622.html"</**URL_article**><**TITRE_article**>« On peut s'attendre à une dégradation de la solvabilité des assureurs européens » </**TITRE_article**></h3>

<**DATE**>08/12/2011 dans L'AGEFI Hebdo<**DATE**> 9

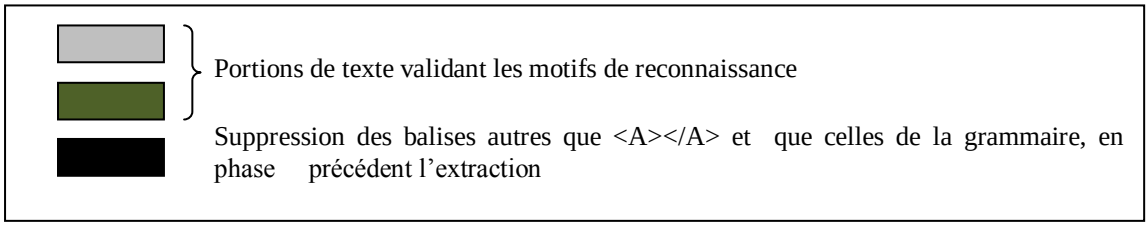
Comment analysez-vous l'exposition des assureurs aux dettes souveraines ?

[16] Pour le confort de la lecture les blancs multiples et les retours à la ligne sont présents ici, bien qu'absents dans les entrées traitées où ils sont remplacés par un espace simple et un seul, même dans des incidences contigües et multiples.

Illustration 11 : Transformation finale avant extraction de la page http://www.agefi.fr/fiche-actualite-eclairages/solvabilite-ii-la-pression-monte-144.html)



Légende :



7.3.2. Les limites du modèle

Les limites de ce modèle sont relativement évidentes : n'ayant rien d'autre pour spécifier un titre, que la longueur d'un énoncé (26 mots) et critère plus pertinent, la quasi impossibilité qu'un point final apparaisse à l'intérieur d'un titre (pas en fin, bien évidemment), la possibilité de reconnaitre une amorce d'article comme un titre, est importante, comme le montre les exemples de l'échantillon traité, fourni plus haut : exemples 2, 4, 6, 8 contre l'exemple 9 dont la reconnaissance est correcte, pour un contexte équivalent.

Le risque existe pareillement de reconnaitre à tort des textes courts (5-8 mots) comme des titres, alors qu'ils ne sont que des entrées de glossaire, des liens produits ou d'autres liens usuels. Le bruit qu'ils induisent dans la reconnaissance peut être aisément maîtrisé en introduisant un filtre solide sur les résultats des extractions. Ce filtre, déjà présent dans le projet tel qu'il existe aujourd'hui devra s'enrichir des entrées de glossaire récupérées dans le volet « extraction des glossaires du monde des assurances ».

Pour ce qui est des titres plus longs, il sera toujours possible d'assortir la longueur de la reconnaissance à une valeur statistique, référence établie sur l'observation des pratiques du domaine

Partie 4

Analyse critique de N5 et retours d’expériences

Chapitre 8 – Analyse critique de N5

N5 a été mon cadre de travail pendant 5 mois. Je me suis habituée insensiblement à ce que la société offrait de mieux comme à ses faiblesses. Identifier les problèmes dont elle souffrait fut facile parce qu'ils transpiraient des détails dérangeants du quotidien. Ils furent faciles à identifier également parce que récurrents, imprimant des stratégies de contournement systématiques.

8.1. Critiques négatives de N5

8.1.1. Insuffisance de la planification matérielle

Le premier problème fut l'insuffisance de moyens matériels. Le caractère artisanal de la société est manifeste à travers une série de détails convergents :

- Le siège social de l'entreprise qui correspond dans les faits à une boîte postale.
- Le lieu d'exercice de l'activité de N5 qui est le domicile de l'employeur, même si une zone dédiée y était aménagée,
- La profusion de documents accumulés dans les fonds d'archives de la société mais qui ne sont pas franchement organisés. Ceux-ci se déclinent en documents informatifs de toute sorte, lectures générales et spécialisées distribuées par grands domaines d'application, annuaires, dictionnaires, inventaires de concepts métiers faits personnellement par le directeur de la structure, peu de documents techniques explicatifs associés à l'exercice des spécialités informatique ou linguistique.

Ce caractère artisanal n'est pas une tare ou un mal en soit, j'aurai l'occasion de l'exposer plus tard. Ce qu'il faut déplorer, c'est la faiblesse de l'équipement de l'entreprise en machines. Le minimum en matériel informatique n'étant pas assuré au sein d'une société dont la principale activité est justement le développement informatique. Ce manque fut rendu manifeste dans une série de mésaventures qui m'ont amené à travailler sur trois ordinateurs différents. Concernant le pôle d'« édition de site », ce « minimum » existe et est même assez solidement constitué. N5 dispose en effet d'un serveur récent, puissant tournant à partir du système linux, Ubuntu. Il est équipé d'un serveur Apache remplissant les emplois attendus de serveur web, php et mysql. Cette machine est également le poste de travail attitré du développeur informaticien : un poste de travail qui n'autorise jamais

qu'une personne à travailler. Ma contribution souhaitée dans la société, promettait d'être informatique avec une contribution possible en traitement des chaînes linguistiques. Une machine était donc un préalable incontournable. Il fut conclu que cette machine serait ma machine d'étudiant, l'entreprise ne disposant pas d'autres ressources. Après trois semaines d'activité, mon ordinateur portable connut une panne irréparable. Une machine de remplacement, trouvée dans l'entourage familial du directeur, connu le même sort après deux mois d'affectation. Chaque changement de machine constitua une perte de temps non négligeable. Outre l'adaptation successive des environnements logiciels, il fallut chaque fois procéder au rapatriement des données locales. Ainsi la première limitation dans l'environnement de N5 réside bien dans l'insuffisance des postes de travail qu'elle fournit. La société ne dispose en propre que d'un seul ordinateur, ce qui est tout de même déplorable !

L'insuffisance d'équipement de la société déborde encore sur sa capacité à proposer un réseau fiable à son personnel et lui fournir une liaison Internet. Cette disposition particulière fut source de retards en cascade, dans mon propre travail, quand il fallut procéder aux phases de test notamment dans la simulation de fonctionnalités portées par d'autres phases du projet et non encore réalisées. Nier que ce réseau existe serait mensonger. Néanmoins, il reste par trop intriqué avec les ressources personnelles du foyer qui accueille la structure, ceci sur des détails aussi insignifiants que le câblage du réseau par exemple.

Le cadre matériel de N5 est clairement fait pour accueillir une ressource permanente bien que l'espace en autorise trois fois plus. Jusqu'ici la bonne volonté de chacun a permis de trouver dans son entourage les solutions de dépannage requises.

8.1.2. Insuffisance d'encadrement technique

Néanmoins, ce n'est pas là que réside le problème majeur de la société. Celui-ci réside dans l'absence de formalisation technique fixant les contours du projet qui a longtemps caractérisé la réalité de l'entreprise. Elle s'est maintenue d'une certaine manière après la prise de conscience du caractère impératif de la documentation technique du projet. Les documents de formalisation sont systématiquement produits après coup. C'est-à-dire après la phase d'implémentation. Il est difficile dans ces conditions d'établir un découpage modulaire prévisionnel des différentes tâches à effectuer, distinctives du

développement modulaire pourtant particulièrement adapté aux modalités de travail en vigueur au sein de cette entreprise.

Je ne doute pas qu'un document consignant un synopsis posant l'ambition d'AssurGroup et plus particulièrement d'AssurWeb, présentant ses prétentions fonctionnelles, en un mot, tenant lieu d'avant-projet, existe ; quoique je ne l'ai jamais vu. Ce que je sais ne pas exister pour l'avoir cherché et demandé, c'est un document de planification, un document directeur fixant les choix (même à minima) et formalisant les sous-parties du développement à effectuer. Circonscrire un projet, définir les limites de l'intervention de chaque intervenant constitue l'un des premiers attendus de la conduite de projets. Il est ainsi possible, à tout moment de l'étape de développement, de mesurer l'écart au but à atteindre. Ceci à tous les niveaux de détails. Comme j'ai eu l'occasion de le dire ailleurs, d'importants travaux de développement avaient déjà été menés à N5. Celui précédant mon arrivée avait été mené pendant à peu près un an et manifestait des manques inacceptables. Le principal étant de n'être absolument pas commenté et de n'être accompagné d'aucune documentation même informelle.

Tous ces traits manifestent l'absence de direction technique dont a souffert N5, de manière générale. Bien qu'il ait rempli de fait tous les rôles, M. Navellou ne pouvait pas porter le projet dans ses dimensions techniques, porter un jugement éclairé sur les choix de développement effectués, ou simplement suivre les réalisations réalisées pour son compte.

Cette faiblesse d'encadrement commençait déjà à se résorber durant mon stage, avec l'arrivée d'un ingénieur réseau, rompu à la direction d'équipe de développeurs et à la conduite de projets.

Un dernier point mérite d'être évoqué, en lien avec le domaine du traitement du langage, c'est l'absence d'investissement de la société dans les outils de traitement de la chaîne écrite du langage naturel ; un manque persistant, malgré des travaux visant à en les sélectionner. Ces travaux ne furent pas exploités au temps de ma présence. Cependant, on peut espérer qu'ils le seront par la suite.

N5 connaît quelques freins structurels, freins d'organisation et freins de compétences qui heureusement pour elle, sont peu à peu en train d'être levés. Cette évolution favorable vient compléter un ensemble de dispositions heureuses existant au sein de la société et qui peuvent contribuer durablement à sa solidité.

8.2. Ce qui marche à N5

N5 a pour elle des qualités d'organisation sur le plan humain que lui envierait plus d'une société. Peut-être que seule sa petite taille en est la cause mais le fait est que l'entreprise ne souffre d'aucun problème de communication. Tout le monde sait ce sur quoi travaillent les autres et aucune barrière n'est entretenue entre les ressources au travail. Les points de communication avec la direction sont fréquents et personne n'est laissée à travailler absolument seul dans son coin pendant des semaines. En même temps, toute autonomie est laissée à chacun de mener ses travaux à sa façon. Cette configuration particulière est sans doute ce qui a longtemps maintenu fait la cohérence de N5, alors même que la communication formelle lu manquait. Il est à noter malgré tout que l'accompagnement fut plus fortement marqué à l'endroit du développement informatique traditionnel.

D'une manière générale, l'atmosphère générale au travail y est agréable.

Chapitre 9 – Retours d'expérience

Entre missions de stage, réussies ou non, relations entre collègues de travail, les expériences intégrées par les individus au travail ont nécessairement des échos à plusieurs niveaux : celui élémentaire des simples compétences, de la maturité intérieure et relationnelle, du regard que l'on porte sur sa future profession. Mon expérience personnelle n'a pas déroger à cette règle.

9.1. Vécus négatifs du stage

Je crois avoir connu ce que redoute tout étudiant qui entreprend une période de stage. Conscient du bagage qu'il possède, il est aussi conscient de tout ce qu'il ignore. Il craint d'être incapable de réaliser ce qu'on lui demande, faute des connaissances adéquates. Celles qui lui par exemple, permettraient de reconnaître d'emblée ce qui est techniquement possible de ce qui ne l'est pas. Le danger de ne pas savoir opérer cette distinction est de se retrouver souvent balloté entre les demandes du commanditaire et les limites réelles de la technique. Un risque qui est décuplé quand une trop grande proximité existe avec son commanditaire. Celui-ci n'est jamais vraiment conscient de la somme de travail, des ressources qu'engagent ses choix, ses velléités. Les répercussions peuvent être autrement lourdes quand il cumule, de surcroît, le rôle de chef de projet, sans la maîtrise

technique que suppose ce le statut. Ce fut exactement la configuration professionnelle dans laquelle je me trouvai. A cause des raisons évoquées précédemment, j'eus donc, largement à déplorer d'avoir :

- mes tâches redéfinies quasiment toutes les semaines et demies sans que le travail réalisé soit jamais exploité ou mené à terme ;

- de voir des tâches que je jugeais nécessaires, évacuées par une précipitation à mon sens mal venue. Exemple, la préparation d'un document de navigation sur les fonctionnalités que j'entendais développer dans le cadre de mon intervention. Malheureusement parce que le travail nécessaire était sous-estimé ou son à-propos méconnu, de nombreux documents de prévision de ce type ne purent jamais être achevés ou même entamés.

Les tâches que l'on me demandait parfois d'accomplir étaient dignes d'un chef de projet expérimenté. Pour autant que l'on prenait son travail à cœur, on réalisait rapidement que bon nombre des aspects d'AssurWeb demandait à être complètement refondé. La tâche était d'autant plus herculéenne qu'aucune donnée technique n'était autrement accessible que par immersion dans le corps du code. La base de données d'AssurWeb quant à elle existait entre une version papier et une version ancienne, implémentée mais non documentée.

Toutes ses conditions n'ont pas manqué de susciter un sentiment horrifié quant à l'ampleur de la tâche supposée, la valse hésitante qui semblait définir ma contribution. Cependant cette phase a été surmontée, grâce notamment à une redistribution conséquente des rôles au sein de la société. On peut aussi penser que tout cela relevait d'une phase d'ajustement inévitable.

9.2. Apports positifs du stage

9.2.1. Compétences acquises en informatique

En matière de connaissance nouvelle, le stage à N5 m'a permis de d'appréhender l'étendue des champs d'exercice du traitement automatique avec certains secteurs de l'industrie informatique traditionnelle. Des liens que je ne soupçonnais pas me sont apparus dans les relations entretenues par certains types de fichiers et les langages de

balises : ceux notamment du format bien connu du PDF et le langage de balises LATEX[17]. Ces recherches m'ont fait réalisé que les techniques du TAL avaient déjà produit, de longue date, des outils de conversion entre formats de fichiers. Le TAL, c'était aussi les logiciels de génération automatique de documents techniques dynamiques, permettant plus d'interactivité que le linéaire du papier. Ces recherches m'ont permis également d'aborder des outils informatiques ignorés jusque-là, les moteurs de recherche, avec *SOLR* et un peu *Lucene*, quels principes sous-tendaient leur fonctionnement, quels principes gouvernaient les moteurs d'indexation, comment ils trouvaient leur place dans les travaux de linguiste informaticien, comment les installer. Tout cela fut un apport majeur !

D'une manière plus générale, on peut dire de ce stage qu'il m'a permis d'expérimenter le développement informatique dans un environnement moins lisse, moins maîtrisé que celui d'une conduite de projet en milieu universitaire. Il m'a permis de combler des lacunes dans mes habitudes de développement, d'acquérir des habitudes fiables, indépendantes du contexte dans lequel elles doivent être mises en œuvre. J'entends par là aussi bien le contexte matériel que l'environnement humain. Je fus en effet contrainte de fonder des stratégies de programmation qui tout en me restant propres (correspondant à ma perception du développement), me permirent d'être efficace et de limiter la déperdition d'énergie corollaire des refontes de code plus ou moins fréquentes dans ce champ d'activité.

La dualité des environnements de travail (Linux, Windows) en me contraignant à évoluer sur différentes plateformes (OS), m'a permis par ailleurs d'acquérir plus d'aisance sur chacune d'elles. Il m'a fallu apprendre à faire sur l'un ce qui ne présente aucune difficulté particulière sur l'autre. J'ai ainsi appris à créer des exécutables sous Windows quand sous linux, seule la procédure d'installation prévaut (pour l'interpréteur perl est à connaître) et une fonction principale à écrire.

9.2.2. Des compétences transversales

Un apport non négligeable de ce stage fut de m'apprendre à communiquer, sur un plan professionnel avec des personnes venant d'un horizon différent du mien. En fait de personnes différentes, les seules que j'avais rencontrées jusque-là étaient des informaticiens. Partageant un peu de leur bagage, je n'ai jamais eu de difficultés particulières pour échanger avec eux. Même plus, grâce à mes formations antérieures, j'ai

[17] Voir Lexique page 129.

toujours eu l'avantage de percevoir très vite les limites humaines dans les raisonnements qu'ils opposent : interprétations, postures privilégiées) Je parvenais ainsi à discerner les lacunes de connaissances, l'obstination ou la mauvaise volonté à l'œuvre, sans les confondre avec les limites réelles de la discipline. Je partageais toujours suffisamment de leur langage pour les comprendre et m'en faire comprendre. Cela ne s'est pas trouvé vrai pour mes relations avec le commanditaire du projet notamment dans le cadre de l'attribution de certaines missions, dans la formulation de certaines tâches. Ce fut le chef de projet Frédéric Lasnier qui fut le lien entre nous.

En lien avec l'expérience précédente, ce stage de 5 mois a eu pour effet de renforcer la conviction de la vérité entourant un retour d'expérience partagé avec nous en cours. Celui-ci constitue un fondamental de l'exercice professionnel : savoir tenir à distance un commanditaire, une fois que sa demande ayant été formulée, elle a été acceptée par la société prestataire de services. C'est la garantie que les étapes normales du document contractuel engageant les deux parties ne soient pas brûlées ou ignorées ou redéfinies trop souvent. Le donneur d'ordre peut être trop envahissant, impatient, sans mauvaise volonté délibérée et être une force désorganisatrice.

9.2.3. Sur le plan des acquis humains

Au-delà d'un vécu initial quelque peu difficile, le nombre élevé de tâches et la rapidité avec laquelle ces tâches se sont succédées ont eu leur part d'effets positifs. L'un d'entre eux fut de m'apprendre de manière très pratique que d'autres modes de fonctionnement que ceux abordés en cours étaient possibles. Le cadre de N5 a eu le mérite de me laisser entrevoir les aménagements nécessaires au mode de fonctionnement « artisanal ». La dynamique d'un petit groupe autorise en effet des aménagements inenvisageables dans une organisation plus importante.

Un autre aspect positif de l'errance qui caractérisa l'attribution des tâches dans les premiers temps de mon stage fut de définir à mes propres yeux les conditions dans lesquelles je me crois capable d'exercer mes capacités. La prise de conscience à l'origine de cette avancée ne fut pas aisée. Je suis depuis toujours victime du complexe qui veut que je dois être capable de tout faire avec un minimum de moyens voire pas de moyens du tout. J'ai donc appris à poser mes besoins et à oser les exposer à mon employeur pour avoir une chance les voir satisfaits. La réalisation d'un projet s'accompagne en effet de contraintes qu'il faut « oser » lever, tout simplement.

Mon expérience en entreprise m'a permis de me forger une nouvelle assurance, parce que mon jugement ayant été posé hors du milieu universitaire, par définition milieu protégé, il a rencontré l'approbation d'un professionnel expérimenté. L'atmosphère de coopération dans laquelle s'est déroulé mon stage y a grandement participé. Nous étions encouragés à nous exprimer, à poser nos caractères, à exprimer nos convictions, prudemment pour certains, mais fondamentalement sans jugement. Cette dynamique de groupe s'est transposée jusque dans l'organisation pratique de nos journées de travail. Nous prenions nos repas en commun. Cette manière de faire nous a, là encore, été profitable puisque c'est lors de ces occasions informelles que nos « leçons » les plus durables ont été échangées. Les déjeuners que nous partagions étaient des occasions pour notre commanditaire de nous communiquer sa conviction quant au projet, d'échanger des savoirs de spécialistes, d'engranger des retours d'expérience sur les pratiques professionnelles émaillant la vie d'entrepreneur ou de chef de projet : le vol d'idées, les opportunités qui naissent de réseaux de contacts, professionnels, amicaux ou familiaux. De plus, à ces occasions nous avons eu des « cours » sur l'art d'organiser la visibilité commerciale de son entreprise, comment construire un business modèle quand on est société éditrice de site, et sur l'art du management en général.

A titre personnel, j'ai entrevu comment travailler autrement, moi qui n'envisage largement l'exercice d'une profession que comme un temps distinct du reste de ma vie, de ma vraie vie. Le temps professionnel est un temps avant tout dévolu à un tiers, la propriété de celui qui m'emploie. Difficile dans ce cas de l'éprouver comme librement vécu. Cependant continuer de penser, de vivre comme cela, c'est mettre de côté 8 heures de ma vie, chaque jour ouvrable, pendant 40 ans. Seule conclusion pour éviter ce gâchis c'est commencer à exister au travail, prendre le temps d'exister dans ses relations à ses collègues, dans ses choix même au travail. C'est un effort que j'ai résolu de tenir et dont j'espère qu'il ne sera pas vain.

Conclusion

La s.a.r.l N5 a été créée avec la préoccupation immédiate de fournir un cadre fonctionnel à la réalisation de l'idée porteuse de M. Jean-Luc Navellou, créer un pôle de services sectoriels consacré aux assurances. La forme juridique de la s.a.r.l devait permettre de supporter intelligemment les efforts financiers requis par la conception et le développement informatique de son projet. Elle lui donnait ainsi accès aux facilités offertes à toutes les entreprises de ce type: « don » de la TVA, réduction d'impôts à titre personnel pour les sommes investies dans le capital de l'entreprise. N5 embauche une à deux personnes simultanément, depuis 4 ans au moins, embauches uniquement supportées par l'afflux d'argent apporté par son directeur. Or celui-ci représente un investissement important, rapporté aux seules ressources d'un particulier. Effort d'autant plus conséquent quand on considère la structure de l'exercice annuel de cette société. Celle-ci n'a pas encore atteint la phase de commercialisation de ses services. Elle ne possède pas encore d'entrées liées à l'activité de vente qu'elle entend mettre en place. Cette configuration à largement influencé ses modalités matérielles de travail, lui faisant privilégier les outils *open source* de manière générale.

L'évolution récente a vu s'opérer des changements majeurs dans l'organisation de N5 mais également dans le devenir de la société. Avec l'arrivée de M. Frédéric Lasnier, non seulement N5 gagna un nouvel associé mais gagna aussi et surtout un ingénieur informaticien aguerri, apportant l'encadrement technique manquant jusque-là à la société.

La contribution avérée de cette force nouvelle fut d'être notamment d'un « valideur » de fonctionnalités. Grâce à lui, il devînt possible à la société, d'évaluer le coût financier, matériel et temporel des différents modules du projet AssurGroup, de les définir autrement qu'en des concepts généraux. Une de ces premières tâches fut de restreindre le projet à une seule réalisation, AssurWeb. Avec lui, il devînt possible de juger des réalisations de l'entreprise, de la qualité mais aussi de la pertinence de ses développements informatiques. Une part importante du travail de ce nouveau chef de projet consista, en coïncidence avec mes trois derniers mois de stage, à abandonner définitivement des pans entiers de développement parce que sans plus-value réelle pour l'entreprise, rapportés aux offres librement accessibles sur marché et parce qu'inappropriés par leur langage de développement.

En même temps qu'un directeur de projet, N5 gagna la mémoire technique qui lui manquait.

D'autres limites que des limites d'encadrement technique pesaient et pèsent encore sur l'entreprise. Des limites matérielles dont ses dirigeants ne semblent pas avoir compris qu'elles existent. La société manque de postes de travail lui appartenant en propres et où le système d'exploitation du projet est dûment représenté. N5 repose essentiellement sur le matériel informatique apporté par ses employés ou stagiaires. Elle ne parvient pas à proposer un environnement de travail stable à plus d'un intervenant. L'énergie, le temps consacrés à mettre en place des solutions palliatives à ce contexte d'exploitation ont un coût: ils nuisent au moral des collaborateurs et nuisent à l'efficacité des efforts engagés. Un épisode de ma période de stage figure bien cette réalité. J'eus en effet une journée de stage travaillée uniquement 1h30 suite à l'installation problématique d'une base de documents, MongoDb. La tentative d'installation entraîna la « disparition » - non pas l'écrasement - du système d'exploitation de la machine. L'opération avait pourtant été menée par notre ingénieur informaticien. Entre tentatives de récupération, opérations de sauvegarde du disque local - qui contenait des années de travaux de celui qui avait fait le prêt de ce matériel-, je ne pus travailler de tout le reste de la journée.

Autre épisode, durant la mise au point des modules d'extraction, mon environnement de travail changea soudainement. Du jour au lendemain, je me trouvais privée de connexion réseau, de connexion Internet stable et continue. En attendant que la société trouvât un moyen pour me rétablir ces prérogatives indispensables, je dus redéfinir les sources alimentant mon système et suspendre la réalisation des spécifications originales du prototype. Le délai introduit fut de 3 jours environ.

Le stage que j'ai effectué fut donc placé à plusieurs titres sous le signe de la précarité : précarité matérielle, caractérisée par le manque d'ordinateurs, la sous représentation du système sous lequel devait pourtant se décliner le projet, précarité technique avec l'absence d'un cahier des charges ou d'un document directeur fixant les contours du projet. Cependant de cette précarité est née la contrepartie que j'eus de « tout » expérimenter, découvrant ainsi de nouvelles dimensions au terme « flexibilité ». J'eus de surcroît, l'heureuse opportunité d'augmenter mon savoir à propos d'un langage déjà rencontré, optimisé pour le traitement des chaînes de caractères, c'est-à-dire PERL. J'ai appris à en manipuler les outils de traitement logique non abordés jusque-là. J'eus l'occasion de mesurer la « transposabilité » de ce langage au monde industriel, d'apprendre

comment un langage comme Ruby qui reprend les dispositions si spéciales de PERL, lui est préféré parce qu'il propose de plus grandes possibilités d'interfaçage avec les autres langages privilégiés dans le monde de la programmation. Il est en effet, plus souvent en adéquation avec les habitudes de développement des entreprises informatiques.

L'autre satisfaction importante de ce stage fut de trouver à y accomplir une mission alliant idéalement les deux attendus de la formation de linguiste informaticien, la description informatique et la description systémique.

Mon embauche au sein de la société N5 est la preuve parmi d'autres que la société possède une dynamique positive. Celle-ci se retrouve premièrement dans l'enthousiasme et la bonne volonté qui unissent ses membres dans la réalisation d'un même but et l'atmosphère de travail détendue et informelle, deuxièmement dans l'entente entre les « employés » et la direction et le soucis qu'on y a d'échanger du haut au bas de la pyramide, courte il est vrai. Cette dynamique qui soude le groupe s'exprime dans des moments informels tel que celui du déjeuner par exemple, qui est une affaire commune.

Maintenant en tant que petite entreprise, quel soutien a-t-elle rencontré et trouvé dans les réalisations logicielles *open source* du TAL pour soutenir son évolution. Les outils du TAL sont nombreux et si chaque domaine est représenté, ils ne le sont pas chacun forcément sur le même plan. Ainsi de nombreux services en ligne gratuits sont disponibles, service de démonstration et ou de consultation, services domiciliés comme ceux d'Open Calais, qui permet de réaliser les ontologies à partir d'entrées textes notamment. Cependant, quand il s'agit d'obtenir un retour exploitable informatiquement sur les étapes d'un processus général de travail sur une chaîne écrite pour une traitement local ou en ligne, peu de solutions existent gratuitement. Il ne s'agit pas bien évidemment d'exiger des éditeurs TAL de livrer gratuitement leurs technologies mais d'ouvrir des perspectives nouvelles à leurs outils, en faisant émerger par là une nouvelle main-d'œuvre, celle des talistes comme spécialistes de ce « nouveau domaine obscur », à la manière de ce qu'a fait l'informatique générale en imposant des informaticiens de spécialité, exemples ceux des réseaux, les administrateurs de bases de données, jusqu'aux web designers.

Or ce qui a manqué à N5, ce sont des outils, bridés pourquoi pas, mais capables de servir de vitrines sensibles à ce que peut vraiment faire le TAL. Particulièrement dans l'étape décisive de la description grammaticale et syntaxique de l'écrit. Rien d'étonnant quand on sait que toute la plus-value du traitement de l'écrit se trouve à cet endroit. Si de

telles vitrines sont rares, le constat de leur disponibilité ou de leur indisponibilité n'est pas uniforme. Une différence de pratique, sans doute commerciale (taille des marchés peut-être) éloigne les deux mondes anglophone et francophone. Les solutions à la fois « clefs en main » et *open source*, originaires des USA notamment, existent plus fréquemment en anglais (voir *Tnt*, qui livré entier, est un outil d'étiquetage et d'analyse syntaxique statistique). En Europe et en France, particulièrement, ce type d'aller-retour donnant-donnant au consommateur est vraiment exceptionnel et le degré de partage bien moindre. Si tout le monde universitaire connait le fameux TreeTagger qui a rendu possible nombre de descriptions linguistiques, les règles de ces descriptions restent des savoirs strictement privés, sauf rares exceptions concernant des langues elles-mêmes rares, dont l'intérêt se confine à de petites communautés. Ainsi, le galicien, le kurde kurmanji (dictionnaires Leff). Ainsi en traduction, hors services de consultation en ligne, le déplacement des ressources vers l'utilisateur consiste au mieux, en la livraison d'interfaces d'utilisation très simple des dictionnaires de mots alignés, listes avec ou sans catégorisation (nom, adjectif et pronoms personnels etc). Ce qui est loin d'être une contribution négligeable quand on monte un prototype.

Des fonctions, des morceaux de code remplissant un office limité quoique conséquent, dans une stratégie de résolution globale, sont également partagées par certains chercheurs et certains centres. Ainsi, les initiatives de Didier Bourrigault (chargé de recherche au CNRS), dont la fonction *getTermAndHead.pl* permet d'extraire des relations sémantiques du type hyperonyme et participe à la modélisation des processus de généralisation mis en œuvre par la langue.

L'autre contribution généreuse de l'open source au TAL, se situe au plan de l'informatique générale, avec la provision de *frameworks*, à la fois boîtes à outils pour les développeurs et gardiens de la conformité des développements. Techniquement parlant ces frameworks peuvent être aussi bien de simples bibliothèques que des environnements de travail multi-fenêtres, au mode graphique avancé. Parmi les plus récents figurent *Symfony*, en utilisation au sein de l'entreprise, ou d'autres comme Drupal. De ce point de vue, il en sort tous les ans. En tant que solutions open source ces outils avancés, s'ils sont performants, sont souvent mis à disposition des utilisateurs avec peu de documentation sur leur utilisation experte ou semi experte. Ces outils. reposent sur une communauté de développeurs réduite en raison même de leur jeunesse et de la concurrence qui caractérise

ces segments (marchés). Cependant, ils ont le mérite de laisser la place à une vraie utilisation d'expérimentation, de développement également.

L'ensemble de ces outils offre un champ d'intégration inégal et surtout discontinu, rapporté à l'ambition de modéliser les processus de surface de la langue jusqu'à sa production de sens. Très souvent ainsi a-t-on une bonne représentation des éditeurs ontologiques évolués mais peu d'étiqueteurs et d'analyseurs syntaxiques intégrant d'office les grammaires requises pour fonctionner. Malgré tout, là encore, tous ces outils ont le mérite d'exister. Ils sont largement renseignés dans un document figurant en annexe de ce mémoire, en tant qu'inventaire de ressources, organisées par grandes fonctions et secteurs majeurs du TAL dans un premier temps et dans un deuxième, en nœuds d'interfaçage potentiel. Ces deux classements tiennent compte des contraintes techniques liées aux outils, celles du langage informatique utilisé et les conditions d'extension de services par rapport à l'objet central.

Aujourd'hui, qui veut travailler le français à partir de l'open source est confronté *ad minimum* à cet incontournable : développer ses propres descriptions après avoir en amont résolu le problème du moteur gérant la reconnaissance de cette description. Au vu des sommes à engager pour l'obtention d'outils fiables et efficaces, gérant la morphologie et la syntaxe de la langue, l'entreprise doit avoir une vision claire de ses prétentions dans le domaine. Car non seulement il faut compter avec l'investissement matériel mais avec celui requis par l'embauche d'un personnel dédié. La délégation de service étant encore une autre option. Le temps de mon stage m'a prouvé que le traitement automatique de la chaîne écrite (a fortiori de la chaîne parlée) est possible par un programmeur linguiste, mais le coût de son intervention est considérable et consiste à réinventé ce qui a déjà été fait à créer son étiqueteur morphologique et son analyseur syntaxique maison. Le travail peut en valoir l'effort (financier et en temps) quand l'entreprise entend capitaliser dessus. La tâche transposé à l'anglais, parait cependant moins ardue qu'il s'agisse de construire une chaîne de traitement de l'écrit ou un prototype.

Pour ce qui est de N5, Il lui reste encore à définir ce que sera ses objectifs dans le traitement de la langue, à l'intérieur de la construction de son prototype et dans la réalisation son projet à long terme.

Sources

Bibliographie

BOULANGER Thierry, HEURTEL Olivier, *Symfony: Maîtrisez le développement PHP avec le framework Symfony*, Saint-Herblain : Eni Editions, 2010, 807 p.

GOSPODNETIÆ Otis., HATCHER Erik, *Lucene In Action*, Lieu d'édition: Manning Publications Co, 2010, coll. Action Series, 475 p.

HAMON Hugo, POTENCIER Fabien, *Symfony : Mieux développer en PHP avec Symfony 1.2 et Doctrine*, Paris : Eyrolles, 2009, coll. Les cahiers du programmeur, 486 p.

PUGH E., SMILEY David, *Solr 1.4 Enterprise Search Server*, Birmingham: Packt Publishing Limited, 2009, 336 p.

Sitographie

ATALA. ATALA : Association pour le Traitement Automatique des Langues. [en ligne]. http://www.atala.org/ [page consultée le 30/10/2011]

BRANTS Thorsten. TnT Statistical Part-of-Speech Tagging. [document électronique].Saarbrüchen, Universität des Saarlandes Computational Linguistics, 1998, http://www.coli.uni-saarland.de/~thorsten/tnt

Dibril. Application Perl/TK non figée… les threads : Utilisation des méthodes internes et modules externes. [document électronique]. Developpez.com, 2011, http://djibril.developpez.com/tutoriels/perl/application-perl-tk-non-figee-threads-win32/

Djibril. Créer un exécutable à partir de sources Perl. [document électronique]. Developpez.com, 2011, http://djibril.developpez.com/tutoriels/perl/creer-executable-sources-perl/

Djibril. Installation des modules Perl CPAN. [document électronique]. Developpez.com, 2008, http://djibril.developpez.com/tutoriels/perl/installation-modules/

DEVELOPPEZ.COM. Les meilleurs cours et tutoriels de la rubrique Perl : consultez tous les cours. [en ligne]. http://perl.developpez.com/cours/#TutorielsDebuter [page consultée le 28/11/2011]

DEVELOPPEZ.COM. Perl et les interfaces graphiques. [en ligne]. http://perl.developpez.com/cours/#TutorielsPerlGUI [page consultée le 28/11/2011] DEVELOPPEZ.COM. Perl et les modules CPAN. [en ligne]. http://perl.developpez.com/cours/#TutorielsPerlCPAN [page consultée le 28/11/2011]

ITRnews.com, 2010, http://www.itrnews.com/articles/101491/sens-semantique-fabrice-lacroix-president-antidot.html

JONES, Richard. A ReStructuredText Primer. [document électronique]. SourceForge, http://docutils.sourceforge.net/docs/user/rst/quickstart.html

JONES, Richard. Introduction à ReStructuredText. [document électronique]. SourceForge, Traduction : Dode William, http://docutils.sourceforge.net/sandbox/wilk/french/quickstart-fr.html

LACROIX Fabrice. *Le sens de la sémantique*.[document électronique]. ITRnews.com, 2010, http://www.itrnews.com/articles/101491/sens-semantique-fabrice-lacroix-president-antidot.html

LHULLIER, Sylvain. Guide de programmation Perl ou comment débuter en Perl. [document électronique]. Developpez.com, 2006, http://lhullier.developpez.com/tutoriels/perl/intro/

LIEUZE, François. Documentez vos modules Perl avec POD. [document électronique]. Developpez.com club des développeurs et IT pro, 2007, http://woufeil.developpez.com/tutoriels/perl/pod/

NUSSBAUM, Lucas. Le format RST : reStructuredText. NUSSBAUM Lucas, 2005, http://www.lucas-nussbaum.net/blog/?p=137

SALAUM, Olivier. Débuter en Perl. [document électronique]. Université de Rennes 1, 2002, http://perso.univ-rennes1.fr/francois.dagorn/perl/ [page consultée le 28/11/2011]

SILBERZTEIN, Max. NOOJ. [en ligne]. http://www.nooj4nlp.net/pages/nooj.html, [page consultée le 28/11/2011]

THE STANFORD NATURAL LANGUAGE PROCESSING GROUP. The Stanford Parser : A statistical parser. Stanford University, http://nlp.stanford.edu/software/lex-parser.shtml [page consultée le 28/11/2011]

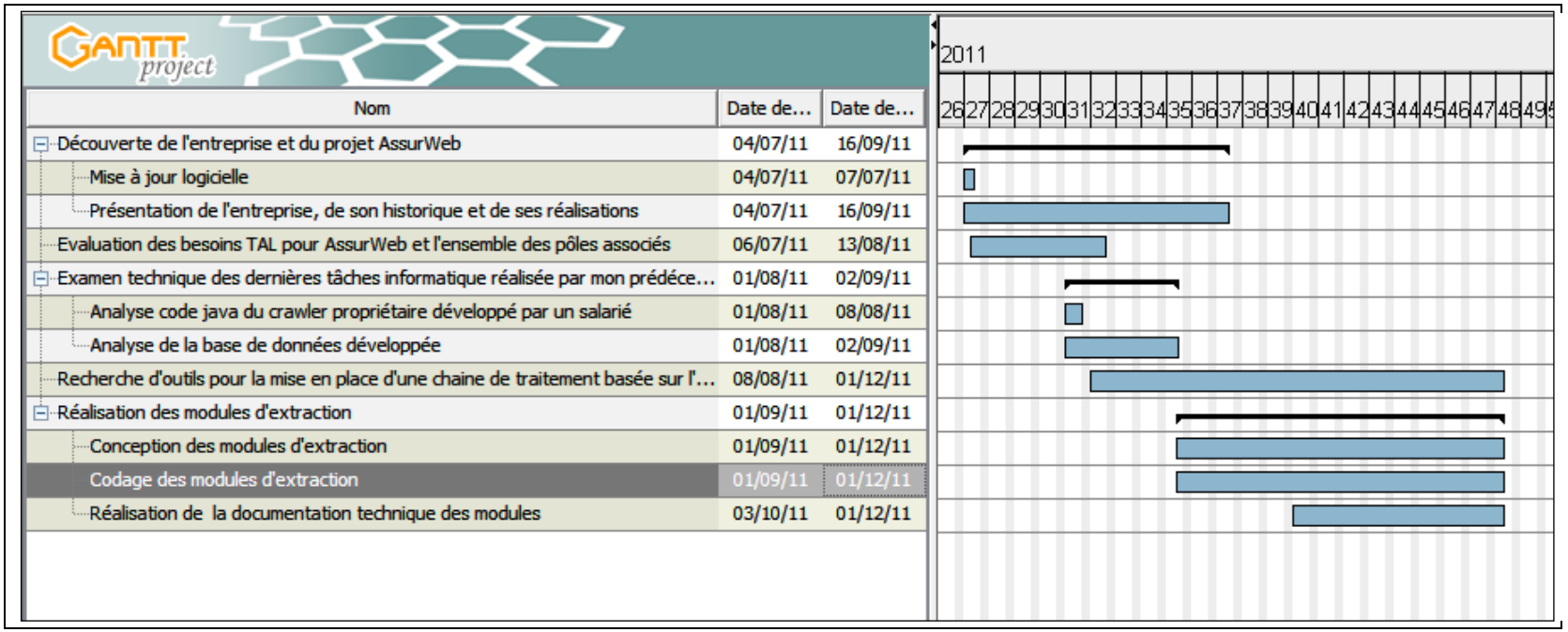
VANHEESCH, Dimitri. Doxygen : Generate documention from source code.[en ligne] www.doxygen.org [page consultée le 28/11/2011]

Contenu soumis à la licence CC-BY-SA 3.0. Traitement automatique du langage naturel. [document électronique]. Wikipedia, http://fr.wikipedia.org/wiki/Traitement_automatique_du_langage_naturel

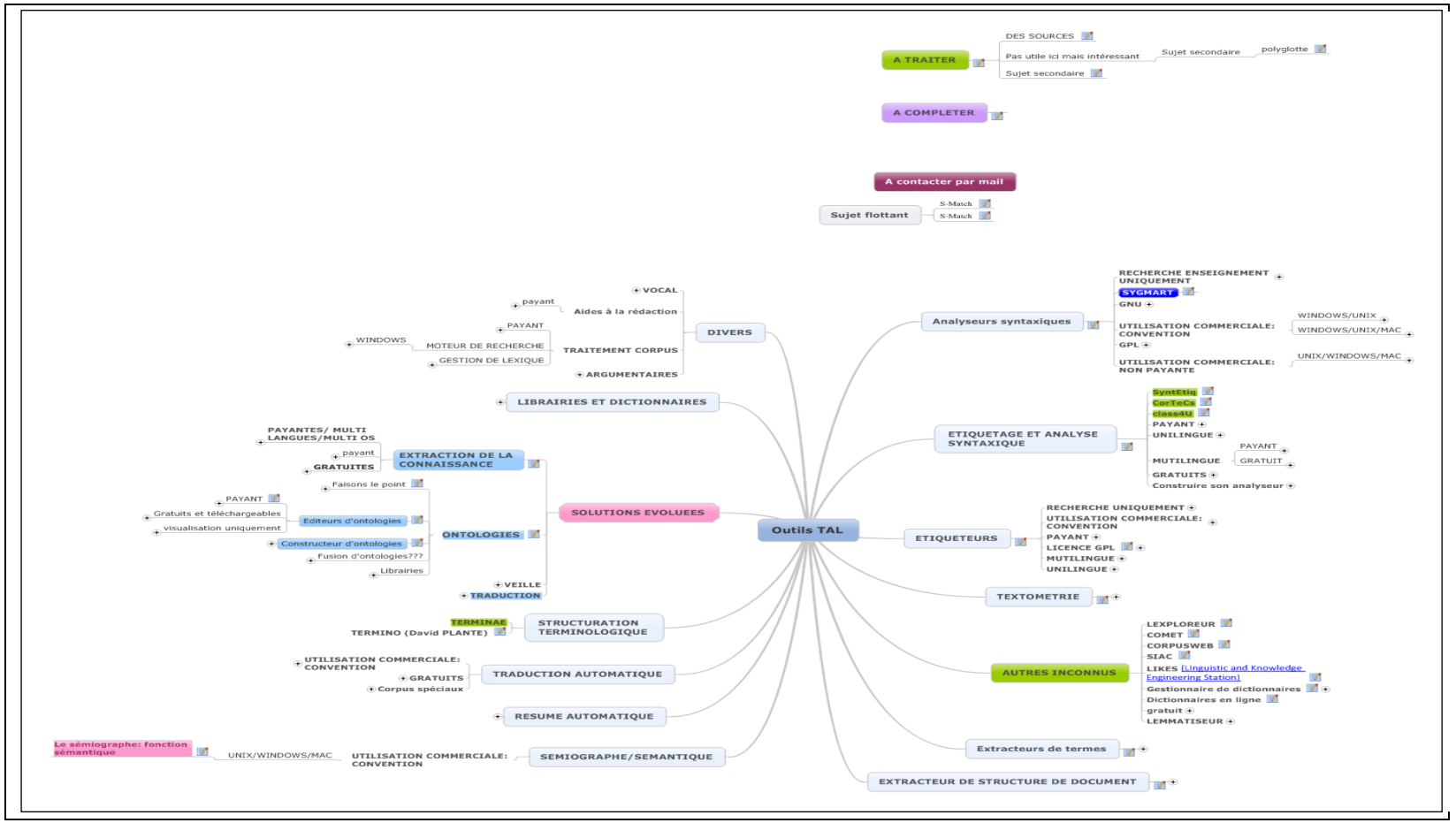
Quick ReStructuredText. SourceForge, 2009, http://docutils.sourceforge.net/docs/user/rst/quickref.html ReStructuredText : Markup Syntax and Parser Component of Docutils. [document électronique]. SourceForge, 2010, http://docutils.sourceforge.net/rst.html

Table des annexes

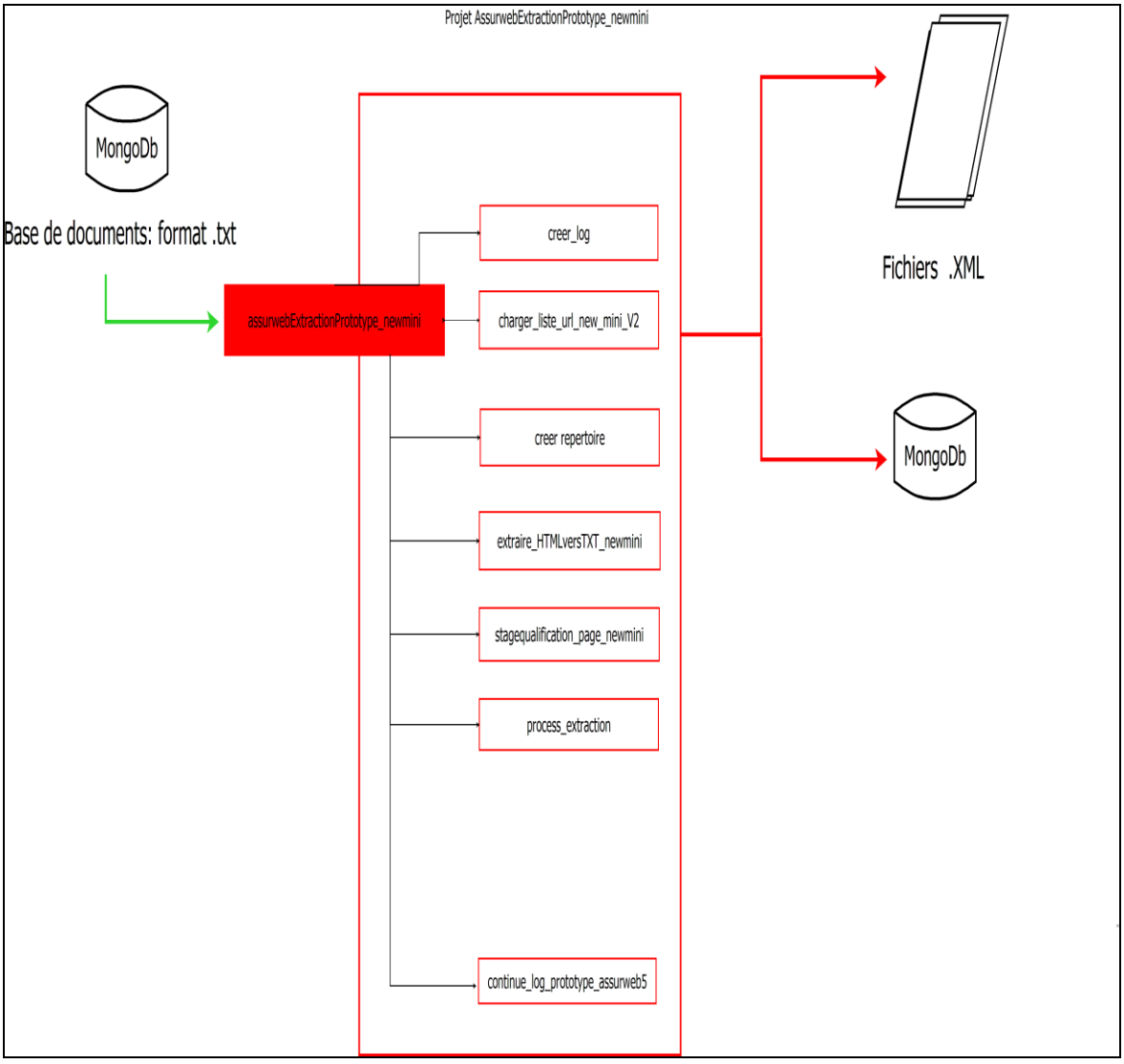
Annexe 1
Distribution des tâches pendant le stage



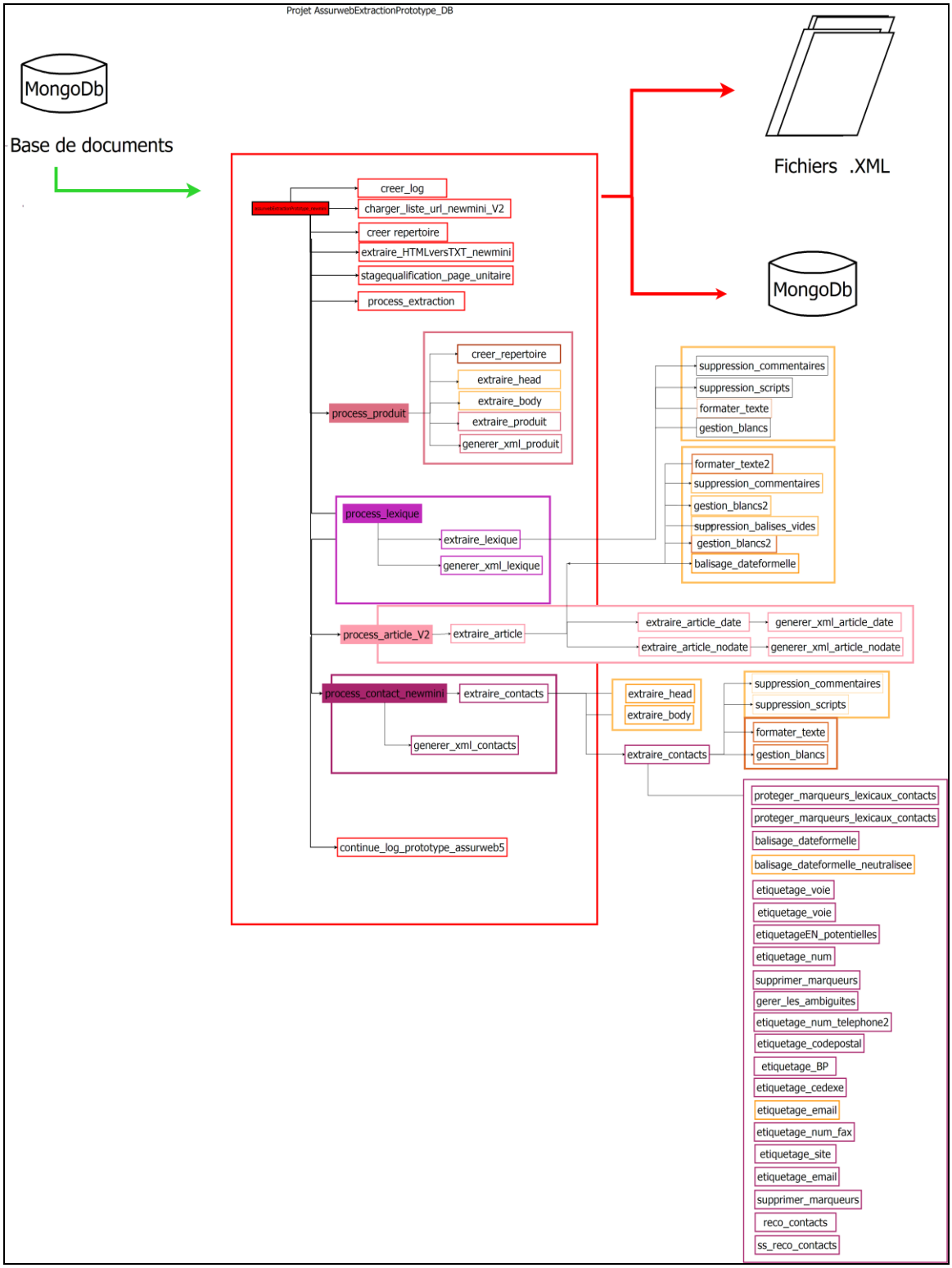
Annexe 2
Recherche d'outils TAL



Annexe 3
Diagramme fonctionnel du prototype d'extraction fin novembre 2011



Annexe 4
Diagramme fonctionnel détaillé du prototype d'extraction fin novembre 2011



Annexe 5
Couverture linguistique des fonctions d'extraction

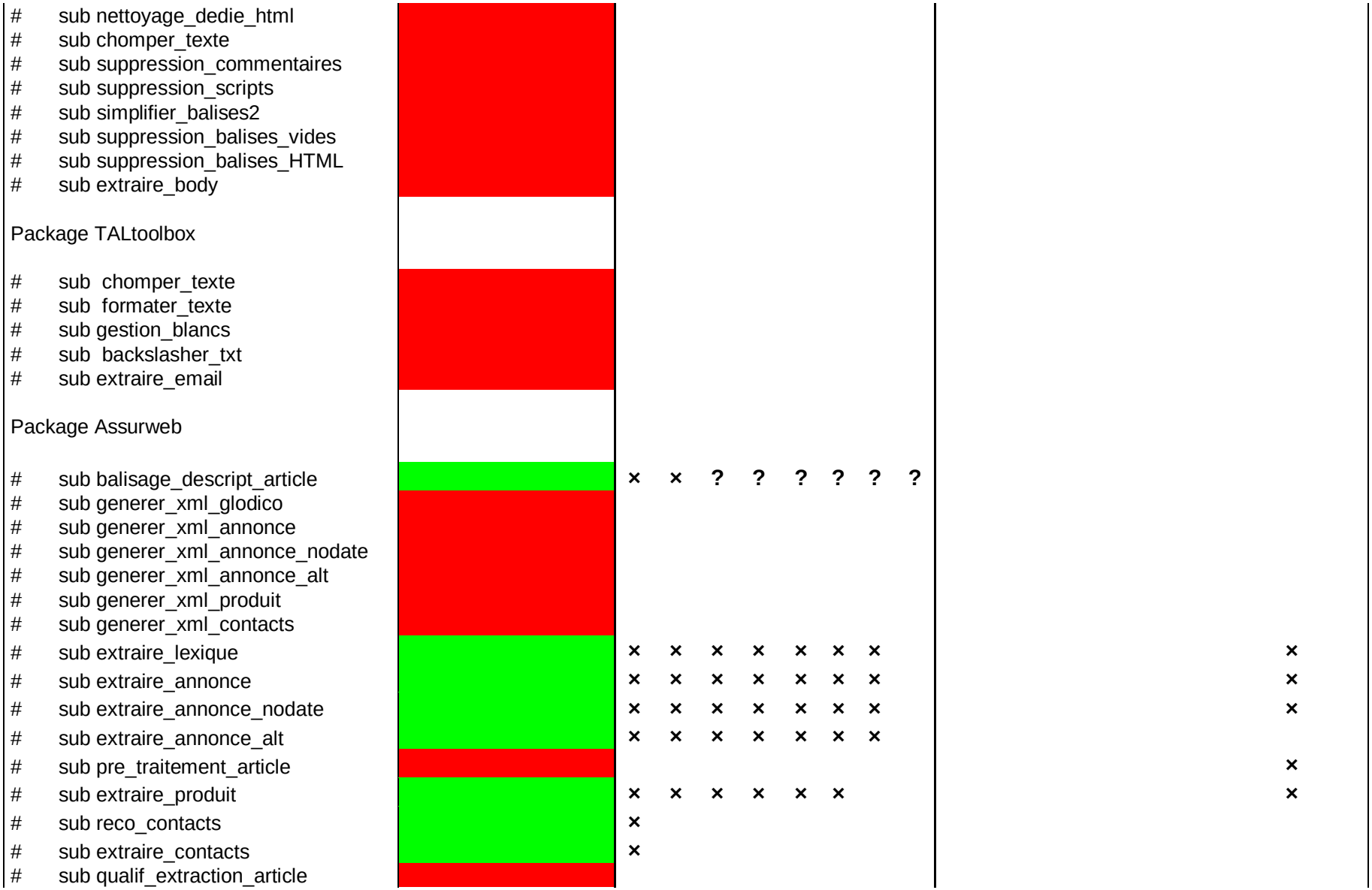
	Langues gérées								Extension Multilingue envisageable							
	FR	EN	SP	DE	PO	IT	SE	CH	FR	EN	SP	DE	PO	IT	SE	CH
Fonctions d'extraction																
1.1- extraction des contacts																
Reconnaissance d'entités nommées																
reconnaissance d'entités nommées potentielles	×	×	×	×	×	×	×									
reconnaissance d'entités nommées de type pays	×	×	×	×	×	×	×									
reconnaissance d'entités nommées de type ville	×	×	×	×	×	×	×									
reconnaissance d'entités nommées de type voie	×	×														
Reconnaissance de numéros de :																
reconnaissance de numéros de téléphone	×	×														
reconnaissance de numéro de fax	×	×														
reconnaissance de nombres	×	×	×	×	×	×	×									
Reconnaissance des types de voie	×	×														

Reconnaissance des blocs "contact"	×						
1.2- Extraction d'articles							
extraction de blocs informatifs de type article							
reconnaissance de date	×	×					
reconnaissance de texte pointant l'emplacement des pleins contenus d'article	×	×					
1.3- Extraction de produits							
extraction de noms de produits	×	×	×	×	×	×	×
1.4- Extraction de lexique	×	×	×	×	×	×	×
Fonctions de stockage de l'information							
2.1- création de xml pour le stockage des contacts extraits	×	×	×	×	×	×	×
2.2- création de xml pour le stockage des informations types des articles	×	×	×	×	×	×	×

2.3- création de xml pour le stockage de noms de produits extraits	×	×	×	×	×	×	×		
2.4- création de xml pour le stockage des lexiques extraits	×	×	×	×	×	×	×		

Annexe 6
Qualification linguistique des fonctions du projet AssurwebExtractionPrototype

	Propriétés linguistiques	Langues gérées								Extension Multilingue envisageable							
		FR	EN	SP	DE	PO	IT	SE	Ch	FR	EN	SP	DE	PO	IT	SE	Ch
Package OrdinaryToolbox																	
# sub creer_repertoire																	
# sub push_hash2																	
# sub push_hash																	
# sub tableau_sortieecran																	
# sub monitoring																	
# sub monitoringparallele																	
Package TALMarkupLanguage																	
#																	
# sub balisage_dateformelle		×	×	?	?	?	?	?		×	×	×	×	×	×	×	×
# sub debalisage																	
# sub etiquetage_email																	
# sub etiquetageEN_potentielles		×	×	×	×	×	×	×		×	×	×	×	×	×	×	×
# sub etiquetageEN_pays		×	×	×	×	×	×	×		×	×	×	×	×	×	×	×
# sub etiquetageEN_ville		×	×	×	×	×	×	×		×	×	×	×	×	×	×	×
# sub etiquetage_num		×	×	×	×	×	×	×									
# sub etiquetage_codepostal		×								?	?	?	?	?	?	?	×
# sub etiquetage_voie		×								×	×	×	×	×	×	×	×
# sub etiquetage_site		×	×	×	×	×	×	×		×	×	×	×	×	×	×	×
# sub etiquetage_num_telephone2		×	×							×	×	×	×	×	×	×	×
Package TALHTMLtoolbox																	


sub nettoyage_dedie_html
sub chomper_texte
sub suppression_commentaires
sub suppression_scripts
sub simplifier_balises2
sub suppression_balises_vides
sub suppression_balises_HTML
sub extraire_body
Package TALtoolbox
sub chomper_texte
sub formater_texte
sub gestion_blancs
sub backslasher_txt
sub extraire_email
Package Assurweb
sub balisage_descript_article
sub generer_xml_glodico
sub generer_xml_annonce
sub generer_xml_annonce_nodate
sub generer_xml_annonce_alt
sub generer_xml_produit
sub generer_xml_contacts
sub extraire_lexique
sub extraire_annonce
sub extraire_annonce_nodate
sub extraire_annonce_alt
sub pre_traitement_article
sub extraire_produit
sub reco_contacts
sub extraire_contacts
sub qualif_extraction_article

# sub stagequalification_page		×		×	×	×	×	×	×	×		×		
# sub generer_glodico														
# sub process_annonces														
# sub process_produit														
# sub process_contacts														
# sub proteger_marqueurs_textuels														
# sub supprimer_marqueurs														
# sub gerer_les_ambiguites		×												
# sub creer_log														
# sub continue_log														
# sub perlcommentstoDoxygen														
Package AssurwebExtractionPrototype														
# sub charger_liste_url														
# sub assurwebextractionprototype														

Légende :

Les fonctions sans actions linguistiques

Les fonctions linguistiques

Annexe 7
Licence d'agrément pour TnT

Licence Agreement for TnT

This HTML version is for information purposes only. Please print the postscript version or the PDF version.

This agreement is made and entered into as of by and between the parties:

a)

Thorsten Brants
Postal address:
Saarland University, FR 8.7 Computational Linguistics
P.O.Box 151150, D-66041 Saarbrücken, Germany
thorsten@brants.net
referred to as: "LICENSER"

b)

Name: JOSEPHINE Rylenc
Address: 3 rue des frères Lumière 77 000 Melun
Email: m.josephine@gmx.fr
referred to as: "USER"

Please print the PS or PDF version and fax the license form to +1-815-846-0652.

Product

TnT, the short form of Trigrams'n'Tags, is a very efficient statistical part-of-speech tagger that is trainable on different languages and virtually any tagset. The component for parameter generation trains on tagged corpora. The system incorporates several methods of smoothing and of handling unknown words.

TnT consists of four modules: parameter generation (`ext-para`), the tagger (`tnt`), comparison of original and tagged files (`txt-diff`), and token/type/tag counting (`tnt-wc`). Two pre-compiled language models are supplied, one for German and one for English.

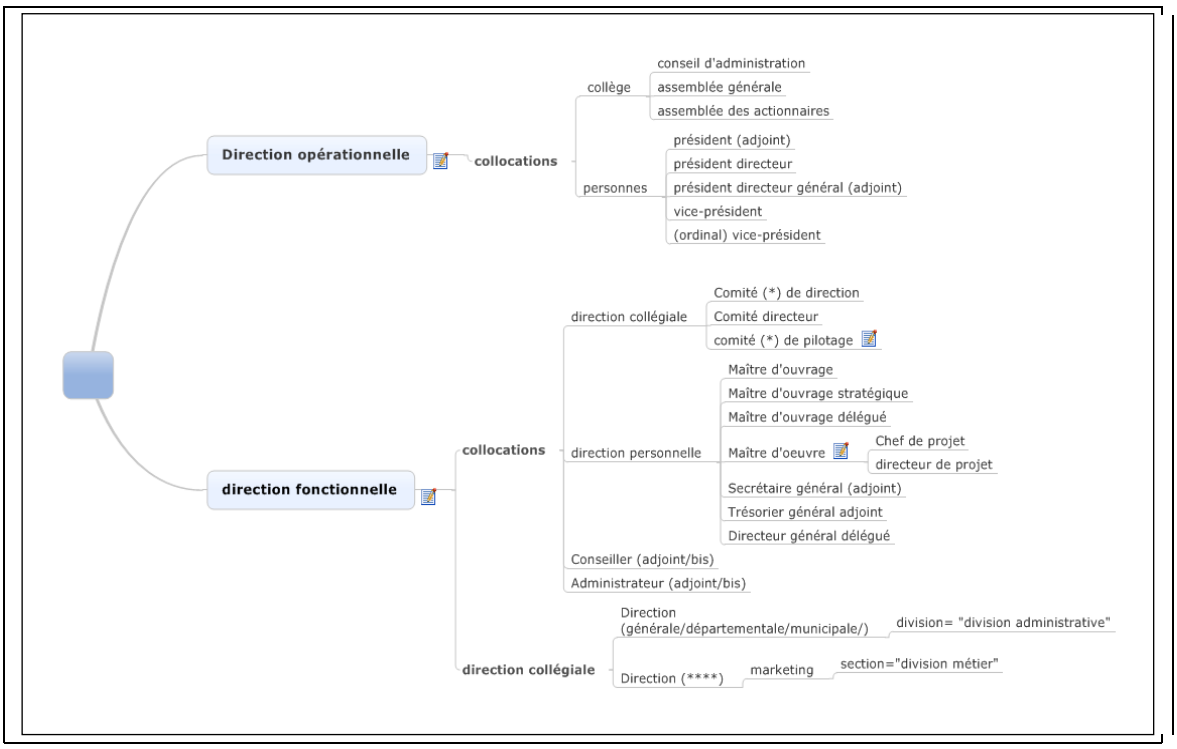
The work on TnT was made possible by support of Hans Uszkoreit at the department of Computational Linguistics, Saarland University, and by a grant from the Deutsche Forschungsgemeinschaft in the Graduiertenkolleg Kognitionswissenschaft Saarbrücken.

The German language model that is delivered with TnT is generated from the NEGRA corpus. This corpus was part-of-speech tagged and manually corrected at the Saarland University, Computational Linguistics, Saarbrücken, and University Stuttgart, Institut für maschinelle Sprachverarbeitung. It was also syntactically (structurally) annotated in Saarbrücken.

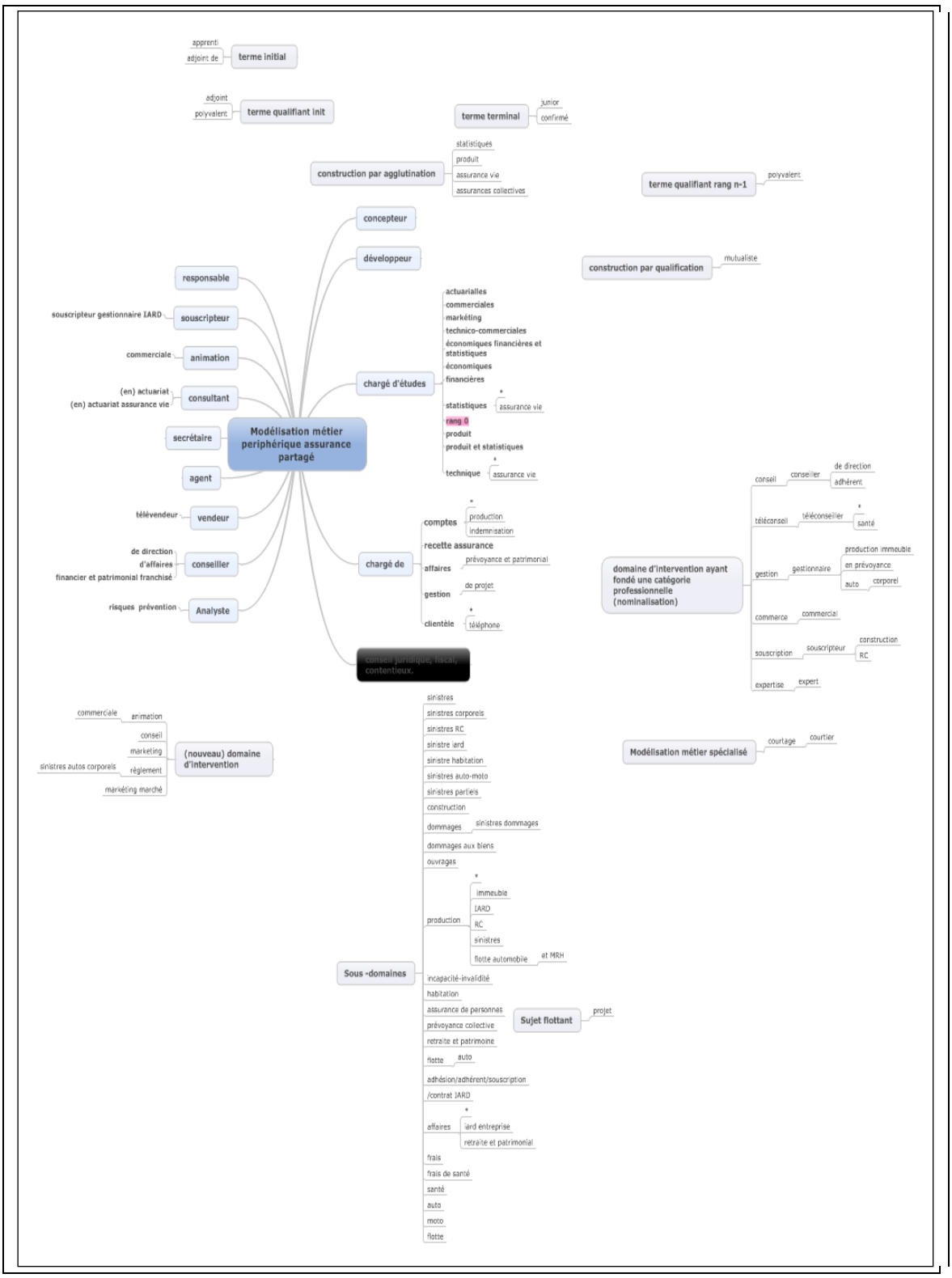
The English language model that is delivered with TnT is generated from the Susanne corpus. This corpus was published by Geoffrey Sampson and is described in his book *English for the Computer*, Clarendon Press, Oxford, 1995.

Copyright

Annexe 8
Exemple de travail préparatoire à la description des connaissances : Mouvements de nomination/démission



Annexe 9
Exemple de travail préparatoire à la description des connaissances des métiers de l'assurance et de ses actuaires



Annexe 10
Adresses canadiennes : entre standards et observations sur le Net

LAVERLY, DE BILLY
bureau 500
925 Grande Allée Ouest
QC – GS 1C1 Québec
Canada

Monsieur le Premier Ministre
Édifice Langevin
80, rue Wellington
Ottawa, Ontario K1A 0A3
Canada

Monsieur le Premier Ministre
Édifice Langevin
80, rue Wellington
Ontario K1A 0A3
Canada

Premier Ministre du Canada
Edifice Langevin
80 Wellington Street
Ottawa, ON KIA OA3
Canada

Billy MAPLES
22 Collingwood Crescent
MB –RH 3L1 Winnipeg
+1 204 261 14 15

Observations tirées du Net

Standards postaux

Variantes dans les numéros

Annexe 11
Adresses belges et suisses

Tower Watson N.V
Avenue Edmond Van
Nieuwheyse 2
1160 Audergherm

Bâtiment administratif de la Pontaise
Avenue des Casernes 2
10 14 Lausanne
Tel. +41(0)21 316 42 00
Fax + 41 (0) 21 316 42 78

Buck Consultants NV
Jan Emiel Mommaert siann 16B
1831 Diegem (Machelen)

Ethias S.A.
Rue des Croisiers 24
4000 Liège

B.B.A.A
Rue de la Charité, 33
Bte 2
BE- 1210 Bruxelles
Téléphone 02/2 57.18.00
Fax 021 316 42 78

Service de la santé publique
Bâtiment administratif de la Pontaise
Avenue de Caserne 2,
CH-1014 Lausanne
Tél. 0 800 106 106
Fax 021 316 42 78

Belgique

Suisse

Variantes dans les numéros

- Rue Jules Cocks 8-10
- Avenue des arts 1-2-3

Variantes dans les codes postaux
Code alphabétique (facultatif) + numéro

En Suisse le code postal (composé de quatre chiffres) est placé avant le nom de la ville de destination et le numéro de l'habitation se trouve après le nom de la rue (contrairement à la France)

Annexe 12
Exemples de fonctionnement de la grammaire « extraction de contacts »

Exemples mixtes de règles francophones et anglophones

Exemple 1 : adresse québécoise

Premier Ministre du Canada
Edifice Langevin
80 Wellington Street
Ottawa, ON KIA OA3
Canada

(R34→)R35
Premier Ministre du Canada
Edifice Langevin
80 Wellington Street
Ottawa, <CODE>ON-KIA-OA3</CODE>
Canada

(R9→R11^{+})→R14

Premier Ministre du Canada
Edifice Langevin
<NUM>80</NUM> Wellington Street
Ottawa, <CODE>ON-KIA-OA3</CODE>
Canada

R31
<PROFESSION>**Premier Ministre**</PROFESSION> **du Canada**
Edifice Langevin
<NUM>80</NUM> Wellington Street
Ottawa, <CODE>ON-KIA-OA3</CODE>
Canada

R43→R44
<PROFESSION>**Premier Ministre**</PROFESSION> **du Canada**
Edifice Langevin
<NUM>80</NUM> Wellington <AREA_0>Street</AREA_0>
Ottawa, <CODE>ON-KIA-OA3</CODE>
Canada

(R46→)R47
<PROFESSION>**Premier Ministre**</PROFESSION> **du Canada**
<AREA_1>Edifice</AREA_1> Langevin
<NUM>80</NUM> Wellington <AREA_0>Street</AREA_0>
Ottawa, <CODE>ON-KIA-OA3</CODE>
Canada

R27→R28

<PROFESSION>**Premier Ministre**</PROFESSION> **du** <EN>**Canada**</EN>
<AREA_1>Edifice</AREA_1> <EN>Langevin</EN>
<NUM>80</NUM> <EN>Wellington</EN> <AREA_0>Street</AREA_0>
<EN>Ottawa</EN>, <CODE>ON-KIA-OA3</CODE>
<EN>Canada</EN>

Nécessité d'une règle de qui s'appuie sur la langue du vocabulaire activé (ici, « street »).
C'est l'exception québécoise, partagée entre normes francophones et normes anglophones.

R46 (Reconnaissance de l'anglais)
<PROFESSION>**Premier Ministre du Canada**</PROFESSION>
<AREA_1>Edifice</AREA_1> <EN>Langevin</EN>

<AREA_0><NUM>80<NUM> <EN>Wellington</EN> Street</AREA_0>

<EN>Ottawa</EN>, <CODE>ON-KIA-OA3</CODE>
<EN>Canada</EN>

R49
<PROFESSION>**Premier Ministre du Canada**</PROFESSION>
<AREA_1>Edifice <EN>Langevin</EN></AREA_1>
<AREA_0><NUM>80<NUM> <EN>Wellington</EN> Street</AREA_0>
<EN>Ottawa</EN>, <CODE>ON-KIA-OA3</CODE>
<EN>Canada</EN>

R5 : nettoyage de la chaîne.
<PROFESSION>**Premier Ministre du Canada**</PROFESSION>
<AREA_1>Edifice <EN>Langevin</EN></AREA_1>
<AREA_0><NUM>80<NUM> <EN>Wellington</EN> Street</AREA_0>
<EN>Ottawa</EN> <CODE>ON-KIA-OA3</CODE>
<EN>Canada</EN>

Chaîne de règles activées :
(R34→)R35→(R9→R11^{+})→R14→ R31→ R43→R44→(R46→)R47→ R27→R28

[R38→)R39 →(R8→R10^{+})→R13→R35→ (R46→)R47→r50→R51→R31→R32 →R36→R48→ R52→R53]→ R46→ R49→ →R5]

Soit le motif d'extraction validé : 0 c a d

Exemple 2 : adresse québécoise

LAVERLY, DE BILLY
bureau 500
925 Grande Allée Ouest
QC – GS 1C1 Québec
Canada

(R34→)R35
LAVERLY, DE BILLY
bureau 500
925 Grande Allée Ouest
<CODE>QC – GS 1C1</CODE> Québec
Canada

(R9→R11^{+}→)R14
LAVERLY, DE BILLY
bureau <NUM>500</NUM>
<NUM>925</NUM> Grande Allée Ouest
<CODE>QC – GS 1C1</CODE Québec
Canada

R29
LAVERLY, DE BILLY
<SSEN>bureau</SSEN> <NUM>500</NUM>
<NUM>925</NUM> Grande Allée Ouest
<CODE>QC – GS 1C1</CODE Québec
Canada

(R43→)R44→
LAVERLY, DE BILLY
<SSEN>bureau</SSEN> <NUM>500</NUM>
<NUM>925</NUM> <AREA_0>Grande Allée</AREA_0> Ouest
<CODE>QC – GS 1C1</CODE > Québec
Canada

R27→R28
<EN>**LAVERLY**</EN>**,** <EN> **DE BILLY**</EN>
<SSEN>bureau</SSEN> <NUM>500</NUM>
<NUM>925</NUM> <AREA_0>Grande Allée</AREA_0> <EN>Ouest
<CODE>QC – GS 1C1</CODE> <EN>Québec
Canada</EN>

R30
<EN>**LAVERLY**</EN>**,** <EN> **DE BILLY**</EN>
<SSEN>bureau <NUM>500</NUM></SSEN>
<NUM>925</NUM> <AREA_0>Grande Allée</AREA_0> <EN>Ouest
<CODE>QC – GS 1C1</CODE> <EN>Québec
Canada</EN>

R45
<EN>**LAVERLY**</EN>**,** <EN> **DE BILLY**</EN>
<SSEN>bureau <NUM>500</NUM></SSEN>
<AREA_0><NUM>925</NUM> Grande Allée <EN>Ouest</EN></AREA_0>
<CODE>QC – GS 1C1</CODE> <EN>Québec
Canada</EN>

R5 (nettoyage de la chaîne):
<EN>**LAVERLY**</EN> <EN>**DE BILLY**</EN>
<SSEN>bureau <NUM>500</NUM></SSEN>
<AREA_0><NUM>925</NUM> Grande Allée <EN>Ouest</EN></AREA_0>
<CODE>QC – GS 1C1</CODE> <EN>Québec
Canada</EN>

[(R34→)R35 →(R9→R11^{+}→)R14→R29→ (R43→)R44→R27→R28 →R30→ R45→R5]

Soit le motif d'extraction validé : 0 e a d

Exemple 3 : adresses anglophone canadienne

Billy MAPLES
22 Collingwood Crescent
MB –RH 3L1 Winnipeg
+1 204 261 14 15

R34→R35
Billy MAPLES
22 Collingwood Crescent
<CODE>MB –RH– 3L1</CODE> Winnipeg
+1 204 261 14 15

R9→R11^{+}→R14→
<NUM>22</NUM> Collingwood Crescent
<CODE>MB –RH –3L1</CODE> Winnipeg
+1 204 261 14 15

R21→ R25→ R26→R20→ R22→ R23→R24
Billy MAPLES
<NUM>22</NUM> Collingwood Crescent
<CODE>MB –RH –3L1</CODE> Winnipeg
<TEL>+1 204 261 14 15</TEL>

R43→R44
Billy MAPLES
<NUM>22</NUM> Collingwood <AREA_0>Crescent</AREA_0>
<CODE>MB –RH –3L1</CODE0> Winnipeg
<TEL>+1 204 261 14 15</TEL>

R27→R28
<EN>**Billy MAPLES**</EN>
<NUM>22</NUM> <EN>Collingwood</EN> <AREA_0>Crescent</AREA_0>
<CODE>MB –RH– 3L1</CODE> <EN>Winnipeg</EN>
<TEL>+1 204 261 14 15</TEL>

R46
<EN>**Billy MAPLES**</EN>
<AREA_0><NUM>22</NUM> <EN>Collingwood</EN> Crescent</AREA_0>
<CODE>MB –RH– 3L1</CODE> <EN>Winnipeg</EN>
<TEL>+1 204 261 14 15</TEL>

R5 :
Nettoyage de la chaîne

Soit l'ensemble des règles activées :
[R34→R35
→R9→R11^{+}→R14→R21→ R25→ R26→R20→ R22→ R23→R24→R43→R44→R27→R28→
R46→R5]
Et le motif validé : 0 a' d'f

Exemple 4 : adresses USA

Phone: (415) 353-2273
UCSF Med Ctr, Dept Neurology
400 Parnassus Ave Fl, 8, Box 0348
San Francisco, CA 94143

R36→R38→R37→R38
Phone: (415) 353-2273
UCSF Med Ctr, Dept Neurology
400 Parnassus Ave Fl, 8, Box 0348
San Francisco, CA <CP>94143</CP>

(R9→R11^{+})→R14
Phone: (415) 353-2273
UCSF Med Ctr, Dept Neurology
<NUM>400</NUM> Parnassus Ave Fl, <NUM>8</NUM>, Box <NUM>0348</NUM>
San Francisco, CA <CP>94143</CP>

R41
Phone: (415) 353-2273
UCSF Med Ctr, Dept Neurology
<NUM>400</NUM> Parnassus Ave Fl, <NUM>8</NUM>, Box <BP>0348</BP>
San Francisco, CA<CP>94143</CP>

(R20→) R22→ R23→R24
Phone: <TEL>(415) 353-2273</TEL>
UCSF Med Ctr, Dept Neurology
<NUM>400</NUM> Parnassus Ave Fl, <NUM>8</NUM>, Box <BP>0348</BP>
San Francisco, CA<CP>94143</CP>

R29→
Phone: <TEL>(415) 353-2273</TEL>
UCSF Med Ctr, <SSEN>Dept</SSEN> Neurology
<NUM>400</NUM> Parnassus Ave Fl, <NUM>8</NUM>, Box <BP>0348</BP>
San Francisco, CA<CP>94143</CP>

(R43→)R44→
Phone: <TEL>(415) 353-2273</TEL>
UCSF Med Ctr, <SSEN>Dept</SSEN> Neurology
<NUM>400</NUM> Parnassus <AREA_0>Ave</AREA_0> <AREA_0> Fl</AREA_0>,
<NUM>8</NUM>, Box <BP>0348</BP>
San Francisco, CA<CP>94143</CP>

(R47→R48)
Phone: <TEL>(415) 353-2273</TEL>
UCSF Med <AREA_1>Ctr</AREA_1>, <SSEN>Dept</SSEN>
<NUM>400</NUM> Parnassus <AREA_0>Ave</AREA_0> <AREA_0> Fl</AREA_0>,
<NUM>8</NUM>, Box <BP>0348</BP>
San Francisco, CA<CP>94143</CP>

R27→R28
Phone: <TEL>(415) 353-2273</TEL>
<EN>UCSF Med</EN> <AREA_1>Ctr</AREA_1>, <SSEN>Dept</SSEN>
<EN>Neurology</EN>

<NUM>400</NUM> <EN>Parnassus</EN> <AREA_0>Ave</AREA_0> <EN>Fl</EN>,
<NUM>8</NUM>, Box <BP>0348</BP>
<EN>San Francisco</EN>, <EN>CA</EN> <CP>94143</CP>

R30
Phone: <TEL>(415) 353-2273</TEL>
<EN>UCSF Med</EN> <AREA_1>Ctr</AREA_1>, <SSEN>Dept
<EN>Neurology</EN></SSEN>
<NUM>400</NUM> <EN>Parnassus</EN> <AREA_0>Ave</AREA_0>
<AREA_0>Fl</AREA_0>, <NUM>8</NUM>, Box <BP>0348</BP>
<EN>San Francisco</EN>, <EN>CA</EN> <CP>94143</CP>

R46
Phone: <TEL>(415) 353-2273</TEL>
<EN>UCSF Med</EN> <AREA_1>Ctr</AREA_1>, <SSEN>Dept
<EN>Neurology</EN></SSEN>
<AREA_0><NUM>400</NUM><EN>Parnassus</EN> Ave</AREA_0>
<AREA_0>Fl, <NUM>8</NUM></AREA_0>, Box <BP>0348</BP>
<EN>San Francisco</EN>, <EN>CA</EN> <CP>94143</CP>

R50
Phone: <TEL>(415) 353-2273</TEL>
<AREA_1><EN>UCSF Med</EN> Ctr</AREA_1>, <SSEN>Dept
<EN>Neurology</EN></SSEN>
<AREA_0><NUM>400</NUM> <EN>Parnassus</EN> Ave</AREA_0>
<AREA_0>Fl, <NUM>8</NUM></AREA_0>,Box <BP>0348</BP>
<EN>San Francisco</EN>, <EN>CA</EN> <CP>94143</CP>

R5: nettoyage de la chaîne.
Phone: <TEL>(415) 353-2273</TEL>
<AREA_1><EN>UCSF Med</EN> Ctr</AREA_1> <SSEN>Dept
<EN>Neurology</EN></SSEN>
<AREA_0><NUM>400</NUM> <EN>Parnassus</EN> Ave <AREA_0>Fl
<NUM>8</NUM></AREA_0>, Box <BP>0348</BP>
<EN>San Francisco</EN> <EN>CA</EN> <CP>94143</CP>

Règles activées :
[R36→R38→R37→R38→(R9→R11^{+})→R14→R41→(R20→) R22→ R23→R24→
R29→(R43→)R44→ R47→R48→ R27→R28→R30→ R46→R50→R5] .

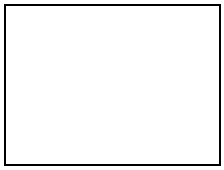
Motif d'extraction:
f c e a'^{2} d'

Annexe 13[18]
DVD « Hub TAL : recherche d'outils »

HUB TAL: RECHERCHE D'OUTILS

1 LA CHAINE ECRITE

See attachment(s): OUTILS TAL INIT.mmap ou **cliquez** sur la figure ci-dessous.



2 SPECIAL SEMANTIQUE

2.1 PROBLEMATIQUE

2.1.1 TENDANCES ANIMANT LA DISCIPLINE

See document(s): indexation-semantique

See attachment(s): Semantic Web Client Library.htm

2.1.2 Faisons le point

Le langage de spécification est l'élément central sur lequel repose l'ontologie.

La plupart de ces langages se basent sur la logique du premier ordre, et représentent donc les connaissances sous forme d'assertion (sujet, prédicat, objet). Parmi les formalismes les plus employés se basant sur la logique des prédicats, on retrouve des langages comme N3 (ou N-Triple). On peut aussi évoquer le langage DEF-*.

Par ailleurs, dans le cadre de ses travaux sur le Web sémantique, le W3C a mis en place en 2002 un groupe de travail dédié au développement de langages standards pour modéliser des ontologies utilisables et échangeables sur le Web. S'inspirant de langages précédents comme DAML+OIL et des fondements théoriques des logiques de description, ce groupe a publié en 2004 une recommandation définissant le langage OWL (Web Ontology Language), fondé sur le standard RDF et en spécifiant une syntaxe XML. Plus expressif que son prédécesseur RDFS, OWL a rapidement pris une place prépondérante dans le paysage des ontologies et est désormais, de facto, le standard le plus utilisé.

[18] Si pour des raisons de conversion de fichier les objets présents dans cette annexe n'étaient plus accessibles, vous pouvez me contacter avec l'objet de message « Mémoire Stendhal » à m.josephine@gmx.fr pour en obtenir une version numérique complète.

Bien que développé pour la représentation des vocabulaires contrôlés et structurés (thésaurus), SKOS peut être utilisé pour élaborer et gérer des ontologies légères multilingues 3.

2.1.3 Robustesse d'une ontologie

D'après Gruber, cinq critères permettent de mettre en évidence des aspects importants d'une ontologie :

La clarté : La définition d'un concept doit faire passer le sens voulu du terme, de manière aussi objective que possible (indépendante du contexte). Une définition doit de plus être complète (c'est-à-dire définie par des conditions à la fois nécessaires et suffisantes) et documentée en langage naturel.

La cohérence : Rien qui ne puisse être inféré de l'ontologie ne doit entrer en contradiction avec les définitions des concepts (y compris celles qui sont exprimées en langage naturel).

L'extensibilité : Les extensions qui pourront être ajoutées à l'ontologie doivent être anticipées. Il doit être possible d'ajouter de nouveaux concepts sans avoir à toucher aux fondations de l'ontologie.

Une déformation d'encodage minimale : Une déformation d'encodage a lieu lorsque la spécification influe la conceptualisation (un concept donné peut être plus simple à définir d'une certaine façon pour un langage d'ontologie donné, bien que cette définition ne corresponde pas exactement au sens initial). Ces déformations doivent être évitées autant que possible.

Un engagement ontologique minimal : Le but d'une ontologie est de définir un vocabulaire pour décrire un domaine, si possible de manière complète ; ni plus, ni moins. Contrairement aux bases de connaissances par exemple, on n'attend pas d'une ontologie qu'elle soit en mesure de fournir systématiquement une réponse à une question arbitraire sur le domaine. Une ontologie est la théorie la plus faible couvrant un domaine ; elle ne définit que les termes nécessaires pour partager la connaissance liée à ce domaine.

2.1.4 Ontologie: ses travaux, ses outils

See document(s): KCAP01-pdf-ACM.pdf

http://oa.upm.es/5483/1/KCAP01-pdf-ACM.pdf

There is no correspondence between existing methodologies and environments for building ontologies, except ODE and METHONTOLOGY [13]. Existing environments just give support for designing and implementing ontologies, but they do not support all
the activities of the ontology life cycle. There are a lot of isolated ontology development tools that cannot interoperate easily, because they are based on different technologies, on different nowledge models for representing ontologies, etc. Consequently, there is a need for a common workbench to ensure a wide acceptance and use of ontological technology. We foresee three main areas in this workbench, as shown in figure 1:

2.1.5 Méthodologies: les théories

Voir Ontoclean

http://en.wikipedia.org/wiki/OntoClean

Boostrapping ontomogies for Web Services

Aviv Segev and Quan Z. Sheng : http://cs.adelaide.edu.au/~qsheng/papers/TSC-segev.pdf
(Le doc en loca est : TSC-segev.pdf)

Voir Methontology

Textes : http://fr.slideshare.net/ontoini/comment-construire-les-ontologies

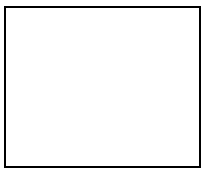
Voir OntoWeb

See document(s): frame-based-models.html,
Une approche complète, un cours complet sur les ontologies, les modèles développé (conception ergonomique en fenêtres (ou Frame based-Models)et plus encore, est dispo à http://www.obitko.com/tutorials/ontologies-semantic-web/frame-based-models.html. This document is part of a research project funded by the IST Programme of the Commission of the European Communities as project number IST-2000-29243. Il définit les limites de l'OntoWeb Consortium /résolution du problème de la sémantique.

2.2 CONNAISSANCES GENERALES

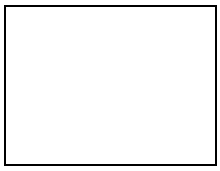
2.2.1 FORMATS COURANTS DE LA MODELISATION ONTOLOGIQUE ET SEMANTIQUE

See attachment(s): Formats en cours dans la modelisations des ontologies.mmap ou **cliquez** la figure ci-dessous.



2.2.2 PRINCIPAUX CENTRES DE RECHERCHES UNIVERSITAIRES

See attachment(s): Outils TAL par centres de recherche.mmap ou **cliquez** la figure ci-dessous.



2.2.3 Lectures afférant aux données conceptuelles

Xtract : http://hal.inria.fr/docs/00/70/42/94/PDF/PosterIC2012.pdf

Outils yatea, linguae, synoterm, syntex, etc : http://lipn.univ-paris13.fr/~szulman/manuelUtil07.pdf

Terminae, OWL:

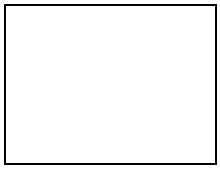
http://liris.cnrs.fr/~ic04/programme/articles/Szulman-IC2004.pdf

http://hal.archives-ouvertes.fr/docs/00/61/81/64/PDF/Szulman_demo.pdf

2.3 SELECTION D'OUTILS POUR LA CONSTRUCTION SEMANTIQUE

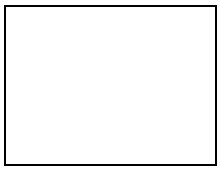
2.3.1 INVENTAIRE IN EXTENSO

See attachment(s): SEMANTIQUE W3C.mmap ou **cliquez** la figure ci-dessous.



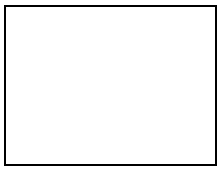
2.3.2 LIGNES D'INTEGRATION POTENTIELLES

See attachment(s): Outils semantiques par principaux langages -Integration.mmap
ou **Cliquez** la figure ci-dessous.



2.3.3 DESCRIPTIONS D'ONTOLOGIES ET SCHEMAS D'INFORMATIONS

See attachment(s): Collection d'ontologies.mmap ou **Cliquez** la figure ci-dessous.



3 LICENCES, DROITS COMMERCIAUX ET DROITS DE DEVELOPPEMENT

See attachment(s): Licences open source.mmap ou **cliquez** la figure ci-dessous.

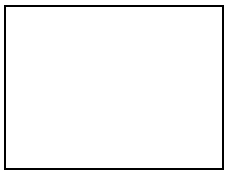


Table des illustrations

Table des tableaux

Sigles et abréviations utilisés

HTML :	HyperText Markup Language
O.S	Operating System
PME :	Petite et Moyenne Entreprise
POD :	Plain Old Documentation
POD2html:	Pod to Html.
ReST	ReStructuredText
s.a.r.l	Société A Responsabilité Limitée
TAL :	Traitement Automatique des Langues
TALN :	Traitement Automatique des Langues Naturelles

Glossaire

Actuaire — L'actuaire est le spécialiste dédié aux calculs statistiques spécifiques des activités d'assurance. Il est avant tout un mathématicien spécialisé dans l'application de cette discipline à la résolution des problèmes du monde financier. Une large partie de son activité consiste par conséquent à calculer les impacts financiers liés au risque, participer à l'élaboration de leurs produits financiers et à faire du conseil aux assurances. Son expertise s'étend à l'ensemble des systèmes d'assurance : organisation des régimes de retraite et de prévoyance, assurances incendies, accidents, risques divers.

Active Perl — est une distribution de Perl gratuite et créée par ActiveState basée sur Perl 5.6.1, ce qui lui permet de gérer tous les programmes perl 5.6.1 et antérieurs. Pour installer cette dernière version de Perl, il est nécessaire de passer par Windows Installer, sachant, toutefois, que le logiciel est disponible pour les plateformes Windows, Linux, Solaris, AIX, HP-UX.

Assurances — terme générique désignant le système visant à prémunir financièrement, un individu, une société contre la survenue d'un aléa, tout le temps contre un engagement financier bien inférieur au coût de l'aléa. On distingue différents secteurs d'assurance en fonction du risque assuré. Ainsi a-t-on la prévoyance où l'aléa assuré est la mort. Le monde des assurances accueille deux grands acteurs : les mutuelles et les assurances à proprement parler. Cf Mutuelle.

Doxygen — est un logiciel *open source* permettant de créer la documentation technique à adosser à un projet. Cette documentation peut atteindre un haut niveau d'interactivité grâce notamment à l'intégration d'autres programmes comme Graphiz qui lui permet par exemple de créer des documents graphiques répondant au caractère précédemment énoncé. Doxygen supporte nativement de nombreux langages tel C, C++, Java, Python et quelques autres encore mais pas Perl, malheureusement.

HTML — L'HyperTextMarkupLanguage est un langage de balises qui permet la mise en forme des pages constituant le Web. Il en gère l'affichage et dans

une certaine latitude, la disposition des éléments qui en constituent la partie statique.

Java est un langage de programmation interprété, créé par Sun Microsystems, ceci en 1995. Par « langage interprété » on entend que le code généré par le langage est, indépendant de la plateforme sur laquelle il s'exerce, tourne grâce à une machine virtuelle, un interprétateur du code, dans ce cas précis, la JVM (Java Virtuel Machine).

LATEX est un langage de balises qui permet de décrire la structure et/ou la sémantique d'un document. C'est un langage de mise en page des objets que sont les tableaux, les images, le texte, etc qui fondent le dit document. Les données relevant de l'apparence (détails et non pas structure) peuvent bien entendu être pris en charge, bien que ce ne soit pas la finalité du langage.

Le fichier obtenu est déclaré au format « .tex ». Il est ensuite compilé par le compilateur latex vers un le format « .dvi ». Celui-ci peut alors être utilisé pour effectuer des conversions vers les formats Pdf, PostScript, Html et bien d'autres. Il est toujours possible de visualiser les fichiers .tex d'origine grâce à certains programmes : Xdi, Kdvi ou Dvips.

Latex est un vrai langage de programmation.

Lucene « est un moteur de recherche libre écrit en Java qui permet d'indexer et de rechercher du texte. C'est un projet *open source* de la fondation Apache mis à disposition sous licence Apache. Il est également disponible pour les langages Ruby, Perl, C++, PHP. » [19]

MindjetManager est un logiciel qui entre dans la catégorie des outils de bureautique. Sa ligne d'édition est cependant plus originale que celle des environnements traditionnels puisqu'elle propose une réflexion en maps ou treillis. Mindjet est extrêmement souple puisqu'à l'intérieur de son paradigme elle permet de manipuler tous les principaux formats habituels, texte, images, documents et pdf. Ses fichiers peuvent également être exportés vers les formats pdf et word.

[19] Source : Wikipédia

Mutuelle	La mutuelle a bien entendu à voir avec l'activité d'assurance, l'activité qui consiste à se prémunir contre le préjudice financier accompagnant souvent un aléa : la maladie, la mort, un accident. L'activité d'assurance s'étend aussi à la perte ou détérioration des biens. La différence essentielle existant entre une mutuelle et un assurance, réside dans le principe de solidarité des contributions versés par les clients ou adhérents, disposition inexistante dans la seconde. Chacun paie un service en fonction de ses moyens. Les coûts et les risques sont donc répartis entre les adhérents. Dans une assurance, une unique contribution est attachée à une prestation donnée. En France deux codes distincts régisse les régimes de la mutualité et celui des assurances.
Perl	est un langage interprété, polyvalent et particulièrement adapté au traitement et à la manipulation des chaînes de caractères soit aux fichiers textes, grâce à l'intégration qu'il fait des expressions régulières De fait, il a été créé en 1987 par Larry Wall un linguiste. Il reprend un certain nombre de fonctionnalités présentes dans le langage C et dans les langages de scripts que sont awk, sed et sh (shell). Le langage a connu de nombreuses versions. Aujourd'hui, on est à la 5.12.
PME	Petite et Moyenne Entreprise
POD	Plain Old Documentation est un langage de balises à la manière d'HTML. Il permet d'écrire de la documentation à partir des sources Perl. Ce code à l'intérieur des sources sera ignoré par l'interpréteur exécutant le code perl alors que le traducteur Pod l'extraira pour créer les documents commentant techniquement le code. La documentation produite sera au format html, latex, txt selon les traducteurs appelés.
POD2htm .	est le module de Pod qui convertit les fichiers .pod en fichiers html.
ReST	ReStructuredText.est un langage de mise en page très léger, sans fioritures et relativement rudimentaire. Son utilisation est privilégiée dans la rédaction des wiki en raison justement de cette légèreté. Ni son

	vocabulaire, ni sa syntaxe ne repose sur l'utilisation des caractères chevrons (< >).
Serveur Apache	Serveur extrêmement répandu sur le net et pensé pour fonctionner sur systèmes Unix. Il assure trois types de services : serveur html, serveur php, serveur mysql. Il a été porté sur Windows en tant que « EasyPhP » et est désormais disponible pour d'autres plateformes.
SOLR	est un serveur de recherche Open Source basé sur Lucene. C'est un logiciel libre réalisé en Java, soutenu par la Fondation Apache.
Strawberry Perl	est une autre distribution possible de Perl, au même titre qu'Active Perl.
Symfony	est un framework php à destination des développeurs, qui en leur fournissant une boîte de composants leur permettant de faire l'économie de certaines phases de développement élémentaires mais consommatrices de temps. En leur offrant cette économie, elle leur permet de se concentrer sur l'objet de leur développement, tout en leur apportant des garanties de code structuré.
TAL(N)	Le Traitement Automatique des Langues Naturelles est le domaine de connaissance qui s'intéresse à la description des systèmes langagiers naturels (TALN). S'entend par là, toutes les langues humaines jamais parlées. Les systèmes de langues peuvent être des créations artificielles tel l'espéranto ou le langage mathématique. La particularité de ces langages est leur univocité. Leur structure simplifiée bannit l'ambiguïté caractéristique des langues humaines proprement naturelles.

Index des noms de personnes

Fred LASNIER : responsable SI et chef de projet d'AssurGroup

Jean-Luc Navellou : directeur de la société N5 et principal associé

Olivier Querne : stagiaire en 1e année d'informatique

Index des noms de lieux

Paris 17è arrondissement

Rue Legendre

Table des matières

MOTS-CLEFS : sélection d'outils, rapport de stage, perl, extractions d'informations

RÉSUMÉ

Le Traitement Automatique du Langage (Naturel) est une discipline jeune qui a vu le jour dans les années 60. Elle est jeune également dans sa représentation en entreprise. L'ensemble de ces caractères définissent sans doute des schémas d'intervention au sein de ces sociétés dont l'expérience au sein de la société N5 peut être un reflet. Ces services sont pour partie prévisibles. Ce sont tout d'abord toutes les missions à visée informative qui permettent à l'entreprise de réaliser son positionnement logiciel voire matériel, au regard de ses besoins. Ces missions d'information sont en effet immanquablement suivies d'une phase de sélection d'outils. Les étapes à suivre voient la mise en place et l'expérimentation des dits outils ou au contraire la commande de missions plus légères, telle la réalisation de missions d'extraction. Ce sont des missions légères au point de vue des ressources engagées, puisque qu'elles ne requièrent pas d'outils spécialement coûteux. Toute la spécificité de cette phase réside dans la particularité des besoins à satisfaire. Bien que là encore, des motifs récurrents s'observent. Les outils de base du linguiste n'étant pas toujours les mêmes, le vocabulaire du langage ?

KEYWORDS: information extraction, tools selection, internship report, perl

ABSTRACT

Automatic (Natural) Language Processing is a new discipline, born in the 60's. the discipline is new too in companies' practice. One can presume how ANLP can work for businesses. What happened in N5 is somehow significant of what kinds of interventions are required by small and medium-sized enterprises. These actions are quite predictable.

First they are informative missions that define these companies' software and material environment. Selecting tools and trying them is often a natural end in this stage but are not always achieved. Next missions command extraction services. These last activities are financially light: they don't require a lot of money, no more expensive materials. The enterprise's needs made this stage specificity. Although, even here, there are some regular patterns because the linguist's bricks are still the same lowest level: (human) vocabulary.