Maître de stage : Guillaume Cabanac
Tuteur : Mohand Boughanem

Romain Panza
10 avril – 22 juin 2012

Rapport de stage

Développement d'une application de scientométrie

Extraction et indexation
d'informations de revues scientifiques

Institut de Recherche en Informatique de Toulouse
118, route de Narbonne
31062 Toulouse Cedex 9

Maître de stage : Guillaume Cabanac
Tuteur : Mohand Boughanem

Romain Panza
10 avril - 22 juin 2012

Rapport de stage

Développement d'une application de scientométrie

Extraction et indexation
d'informations de revues scientifiques

Institut de Recherche en Informatique de Toulouse
118, route de Narbonne
31062 Toulouse Cedex 9

Remerciements

Je remercie M. Mohand Boughanem, professeur à Université Paul Sabatier pour m'avoir permis d'effectuer mon stage à l'IRIT.

Je tiens à remercier mon maître de stage, M. Guillaume Cabanac pour m'avoir accueilli au sien de l'équipe SIG de l'IRIT, ainsi que pour sa patience et sa grande disponibilité dont il a fait preuve, ses nombreuses remarques d'ordre professionnel et sa rigueur qui m'ont permis de progresser.

Enfin, une attention particulière pour Jérémie Clos pour son aide continuelle pendant ces onze semaines, ainsi que pour Clément et David stagiaires de DUT eux aussi, pour leur soutien et la convivialité qu'ils ont instaurée, mais toujours dans un esprit professionnel.

Je remercie également ma mère pour son soutien inconditionnel.

Table des matières

Chapitre 1

Introduction

Pour achever ma formation de DUT Informatique, j'ai réalisé un stage de onze semaines en milieu professionnel. C'est suite à une proposition du Professeur Mohand Boughanem que j'ai eu l'opportunité de faire ce stage au sein de l'Institut de Recherche en Informatique de Toulouse (IRIT). Ce stage me paraissait intéressant à plus d'un titre, mais surtout il me permettait tout en travaillant dans un milieu professionnel, d'approcher et de me familiariser avec le monde de la recherche.

Durant ce stage, trois principales missions m'ont été confiés. Premièrement, transformer des données XML dans un format adapté à la mise a jour de ces données dans une base de données documentaires scientifiques (DBLP). En second lieu, nous devons à partir d'un fichier *pdf* qui rassemblait les premières pages de publications scientifiques de la revue ISI, extraire une série de données pour pouvoir ensuite les exploiter et étudier l'historique de l'affiliation des laboratoires pour un auteur donné en fonction de la chronologie. Et pour finir, nous extrairons des données sur des conférences scientifiques dans une page HTML.

Dans le second chapitre, nous exposerons le contexte professionnel, l'environnement de travail et les objectifs du stage. Les chapitres 3, 4 et 5 détaillent les différentes étapes de chaque mission. Pour conclure ce rapport, nous terminerons par bilan technique et personnel.

Chapitre 2

Présentation du contexte professionnel : un laboratoire de recherche

Ce chapitre décrit le contexte professionnel : en premier lieu l'IRIT (section 2.1 et 2.2), puis le cadre de travail (section 2.3) et enfin les objectifs détaillés du stage (section 2.4).

2.1 L'Institut de Recherche en Informatique de Toulouse (IRIT)

L'Institut de Recherche en Informatique de Toulouse, l'IRIT, est une unité mixte de recherche, commune au Centre National de Recherche Scientifique (CNRS), à l'Institut National Polytechnique de Toulouse (INPT), à l'Université Paul Sabatier Toulouse 3 (UPS), à l'Université des Sciences Sociales Toulouse 1 (UT1), à l'Université Toulouse 2-Le Mirail (UTM) et à l'ENSEEIHT. Il est dirigé par M. Michel Daydé et a été fondé en 1990. L'IRIT représente un des plus forts potentiels de la recherche en informatique en France. Actuellement, environ 700 personnes y travaillent, dont 225 enseignants-chercheurs, 28 chercheurs, 277 doctorants et post-doctorants, et 165 personnels administratifs, techniques et autres personnels [1, 2]. Les 18 équipes de recherche sont structurées selon sept thèmes scientifiques qui couvrent l'ensemble des domaines de l'informatique actuelle :

- analyse et synthèse de l'information,
- indexation et recherche d'informations,
- interaction, autonomie, dialogue et coopération,
- raisonnement et décision,
- modélisation, algorithmes et calcul haute performance,
- architecture, systèmes et réseaux,
- sûreté de développement du logiciel.

2.2 L'équipe « Systèmes d'Informations Généralisés » (SIG)

Mon stage s'est déroulé au sein de l'équipe SIG (Systèmes d'Informations Généralisés) qui est rattachée au thème 2 de l'IRIT : Indexation et recherche d'informations. Son activité est centrée sur la modélisation et l'accès à l'information hétérogène, s'appuyant sur les problèmes de modélisation et de manipulation de systèmes d'information et de bases de données [3].

L'équipe SIG est organisée en 4 composantes (cf figure 2.1) :

- DDSS : Documents, données semi-structurées et usages
- ED : Conception de systèmes d'informations décisionnels
- EVI : Exploration et visualisation d'information
- RFI : Recherche et Filtrage d'Informations

J'ai intégré le groupe de la composante RFI où j'ai travaillé avec Guillaume Cabanac, mon maître de stage. Il a défini les objectifs de mon stage qui sont détaillés dans la section suivante.

1. Source : `http://www.irit.fr/Personnel,618?lang=fr`
2. Source : `http://fr.wikipedia.org/wiki/Institut_de_recherche_en_informatique_de_Toulouse`
3. Source : `http://www.irit.fr/-Equipe-SIG-?lang=fr`

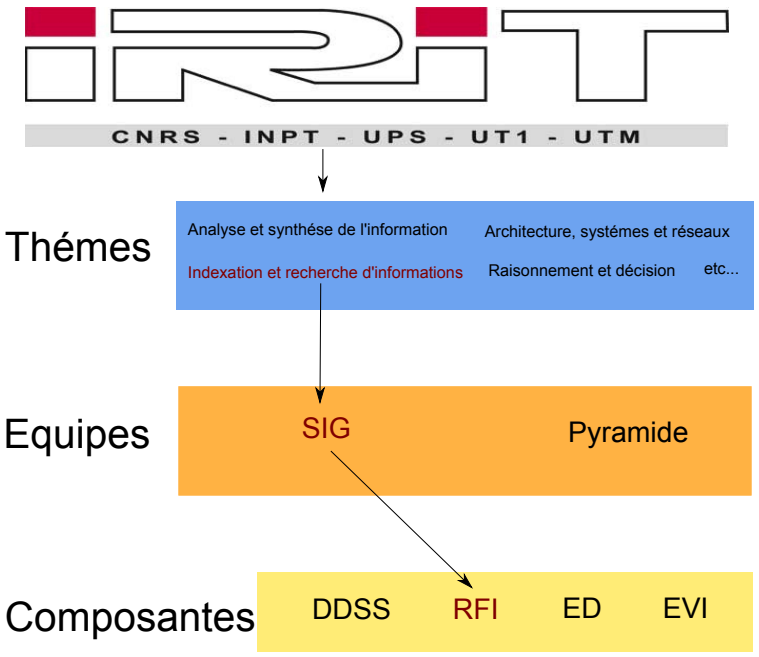

FIGURE 2.1 – Situation de l'équipe SIG dans l'organigramme de l'IRIT

2.3 Cadre de travail

Ce stage a s'est déroulé dans une salle comportant une dizaine d'ordinateurs et principalement utilisés par les stagiaires et doctorants de l'équipe SIG.

Les horaires de travail convenus avec le maître de stage ont été les suivants :

Jour	Horaires
Lundi	: 9h - 12h et 13h - 18h
Mardi	: 9h - 12h et 13h - 18h
Mercredi	: 9h - 12h et 13h - 18h
Jeudi	: 9h - 12h et 13h - 18h
Vendredi	: 9h - 12h

La planification initiale est illustrée sur la figure 2.2 qui présente la durée des 3 missions qui m'ont été confiées. Les missions CORIA et ACM son deux tâches rapide en terme de temps, elles se limitaient à de la transformation et de l'extraction d'une petite quantité de données. Dans la mission ISI nous avons mis en œuvre conjointement avec mon maître de stage une période d'analyse et de conception de la base de données, puis la phase de développement d'un programme, capable d'extraire toutes les informations nécessaires à notre base de données.

À la fin de chaque semaine, je fournissais à mon maître de stage un compte-rendu qui synthétisait mon travail de la semaine et indiquait le travail restant. C'est à partir de ces comptes-rendus que nous discutions de l'avancement des missions et des orientations à prendre pour la semaine à venir (une sélection des compte-rendus est disponible en Annexes). Mon maître de stage à imposé la rédaction de tout document devant lui être remis en LaTeX [4], ce qui est le cas de ce rapport.

Après avoir précisé le cadre de travail du stage, la section suivante présente ses objectifs.

4. LaTeXest un langage et un système de composition de documents, il est pendant de Word ou OpenOffice, et très utilisé dans le monde universitaire.

2.4 Missions du stage

Trois missions ont été mises en œuvre durant ce stage, chacune avec des objectifs différents. Néanmoins toutes ces missions concernent la scientométrie, il est donc essentiel de définir ce terme. La scientométrie est la mesure et l'analyse de la science par une démarche scientifique [HW01]. Ce stage s'inscrit dans le cadre du travail de mon maître de stage dans ce domaine [Cab12]. « *La scientométrie est souvent en partie liée avec la bibliométrie et peut être considérée à la fois comme une réduction et une extension de celle-ci :*

- *Réduction, puisqu'elle n'applique les techniques bibliométriques qu'au champ des études de la science et de la technologie, en comptabilisant les publications scientifiques.*
- *Extension, puisqu'elle n'analyse pas seulement les publications mais également des financements, ressources humaines, brevets, etc.* » [5]

2.4.1 Objectif de la mission CORIA

CORIA, COnférence en Recherche d'Information et Applications, est une conférence scientifique annuelle sur le thème du document numérique est de la recherche d'information. Cette conférence permet aux chercheurs de communiquer leurs ressources scientifiques et de les publier.

Or, toutes les publications des chercheurs en informatique sont regroupées dans une base de données internationale DBLP [6] hébergée par l'université Universität Trier, en Allemagne et qui regroupe actuellement plus de deux millions de publications scientifiques.

Les chercheurs envoient donc leurs publications aux administrateurs de la base, sous un format imposé par DBLP dans le but d'être référencées. Mon objectif était de fournir le fichier au format imposé comportant l'ensemble des publications de la conférence CORIA 2012. Le format imposé par DBLP nous contraint à manipuler un fichier XML et utiliser le langage XSLT pour cette mission.

2.4.2 Objectif de la mission ISI

ISI, Ingénierie des Systèmes d'Information, est une revue scientifique mensuelle. Elle regroupe des publications scientifiques traitant de problématiques de recherche autour des systèmes d'information parmi lesquelles la spécification de besoins, la modélisation et la métamodélisation des SI, les méthodologies de conception, reconception et maintenance, les processus d'intégration de SI, les problématiques de sécurité, les problématiques de bases de données et/ou de recherche d'information.

Lors d'une publication scientifique plusieurs éléments sont importants et caractérisent la publication. Premièrement le/les auteur(s), le/les laboratoire(s), ensuite le/les affiliation(s) du/des auteur(s) à leur laboratoires ou université, puis le titre de la publication et enfin la date.

Grâce à la base de données DBLP, les chercheurs peuvent ainsi tracer les liens entre les différents auteurs des articles et observer les collaborations entre auteurs.

L'objectif de cette mission était de créer une base de données, à partir des informations données par ISI, avec les informations suivantes :

5. Définition scientométrie : `http://fr.wikipedia.org/wiki/Scientom%C3%A9trie`
6. DBLP : Computer Science Bibliography - `http://www.informatik.uni-trier.de/~ley/db/`

Table	Champ
Auteur	Nom, prénom
Article	Titre
Laboratoire	Nom, ville, université, pays
Numero	Volume, numéro, année

TABLE 2.1 – Champs de données devant ce trouver dans la base de données ISI

Cette base de données vise principalement de retracer l'historique de l'affiliation des laboratoires pour un auteur donné en fonction de la chronologie. Elle pourra aussi servir à retracer les collaborations entre auteurs comme le fait actuellement DBLP.

2.4.3 Objectif de la mission ACM

ACM, Association for Computing Machinery, littéralement « Association pour la machinerie informatique » est une association américaine crée en 1947, son but consiste à développer et soutenir la recherche scientifique en informatique et l'innovation. Elle est structurée par antennes thématiques ou « Spécial Interest Group » (SIG) en anglais [7]. Dans notre cas, nous nous intéresserons à l'antenne thématique SIGIR (Antenne en recherche d'information). SIGIR sélectionne et publie chaque année un certain nombre de publications de chercheurs du monde entier. Sur le site d'ACM sont disponibles par année, les chiffres des publications, avec le nombre de publications envoyées, le nombre de publications sélectionnées par SIGIR et son taux d'acceptation.

L'objectif est d'extraire, depuis le site d'ACM, les différents chiffres de publication de SIGIR et des autres conférences ACM de notre choix.

Les trois chapitres suivants présentent de manière détaillée les différentes missions et leurs objectifs, en énonçant les différentes possibilités et en justifiant nos choix.

7. Liste des antennes : `http://fr.wikipedia.org/wiki/Association_for_Computing_Machinery`

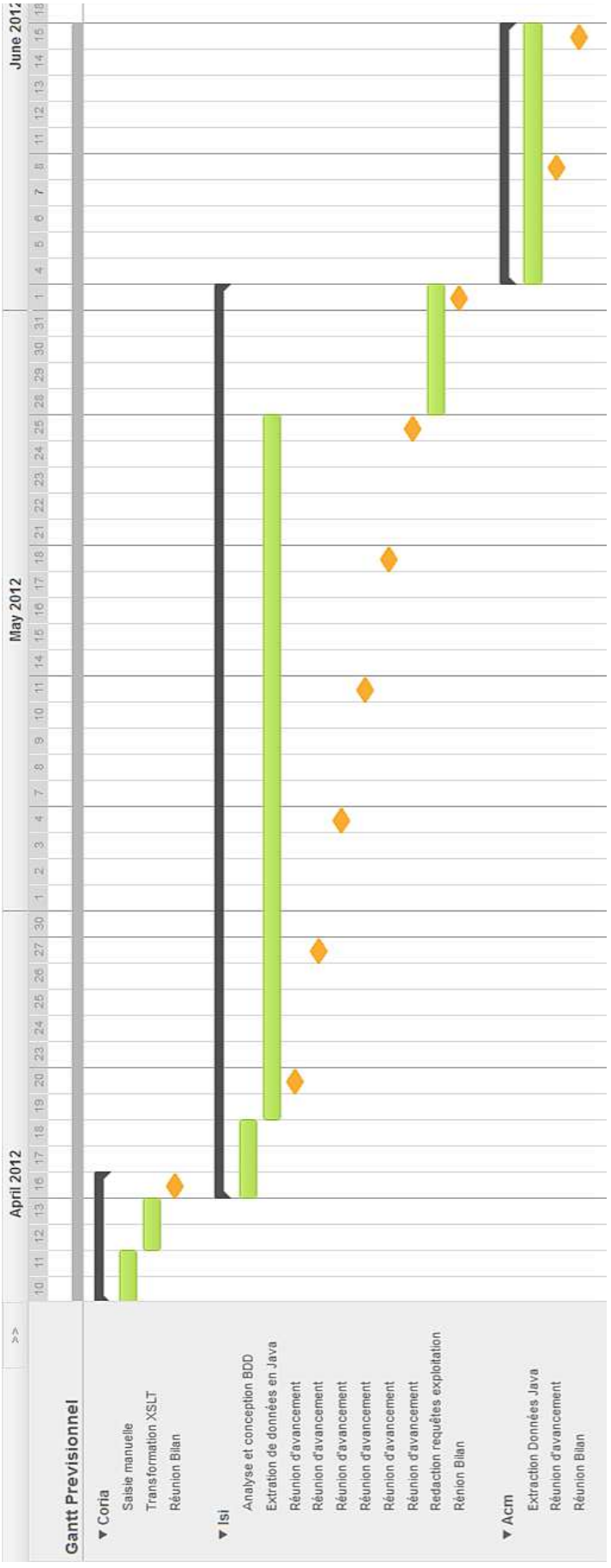

FIGURE 2.2 – Diagramme de GANTT présentant la planification initiale des 3 missions du stage

Chapitre 3

Mission CORIA

Ce chapitre expose comment nous avons résolu cette mission. Pour plus de détails sur les objectifs de cette mission voir section 2.4.1.

3.1 Acquisition des données

Le format des données que nous avions à transformer pour la base DBLP était sous forme d'un livre rassemblant toutes les actes de la conférence, il a fallu saisir toutes les informations dans un fichier XML dont la DTD nous a été fournie par mon maître de stage.

3.2 Transformation des données avec XSLT

Une fois les données saisies et prêtes dans un fichier XML, nous disposions d'un script en XSLT nous permettant de transformer les données XML dans le format demandé par DBLP, le format demandé par DBLP était une page HTML structurée dans un ordre précis.

« XSLT (Extensible Stylesheet Language Transformations) est un langage permettant de manipuler et de transformer des documents XML dans divers formats, comme le HTML, ou encore le XML. XSLT est un langage de programmation complet : on peut créer des « fonctions », des boucles, calculer un maximum, faire des recherches dans un document XML, compter le nombre de résultats, etc. Mais XSLT est avant tout orienté vers le traitement d'un fichier XML. On peut appliquer des modèles (templates) sur les balises XML, puis leur appliquer des traitements divers. » [8]

Nous n'avions donc pas à développer directement en HTML. Notre objectif était plutôt de modifier le code fourni afin d'y inclure un nouveau champs d'information à traiter : les éditeurs du livre CORIA 2012. L'information n'existait pas dans la DTD fournie et elle n'était pas gérée par le script XSLT. Nous avons donc dû modifier la DTD et nous former à ce nouveau langage qu'est XSLT afin d'adapter le code pour y inclure d'éventuels éditeurs.

3.3 Bilan de la mission

Cette mission avait pour principale contrainte, une rigueur extrême dans la saisie du fichier XML qui devait absolument ne contenir aucune erreur. Pour s'assurer de la fiabilité du fichier XML, de nombreuses et minutieuses relectures ont dû être effectuées. Cette mission nous a permis d'approcher et de connaitre un nouveau langage de programmation, le XSLT très pratique pour traiter des données XML en vu d'autres traitements.

8. Source : `http://haypo.developpez.com/tutoriel/xml/xslt/`

Chapitre 4

Mission ISI

Ce chapitre présente comment nous avons résolu cette mission. Pour plus de détails sur les objectifs de cette mission voir section 2.4.2.

4.1 Analyse et conception de la base de données

Une fois la liste des données à extraire établie (voir section 2.4.2, et table 2.1) nous nous avons réalisé l'analyse et la conception d'une base de données capable de répondre à la demande du maître de stage. Notre base de données devait être capable de tracer l'historique des affiliations des laboratoires pour un auteur donné en fonction de la chronologie, tout en prenant en compte des informations complémentaires, qui sont l'appartenance à un article, à une parution dans un numéro de la revue ISI et le lien d'écriture entre un auteur et un article.

Pour répondre à ces objectifs, nous avons proposé un modèle conceptuel de données et un modèle logique de données (voir respectivement figures 4.1 et 4.2) [ST05].

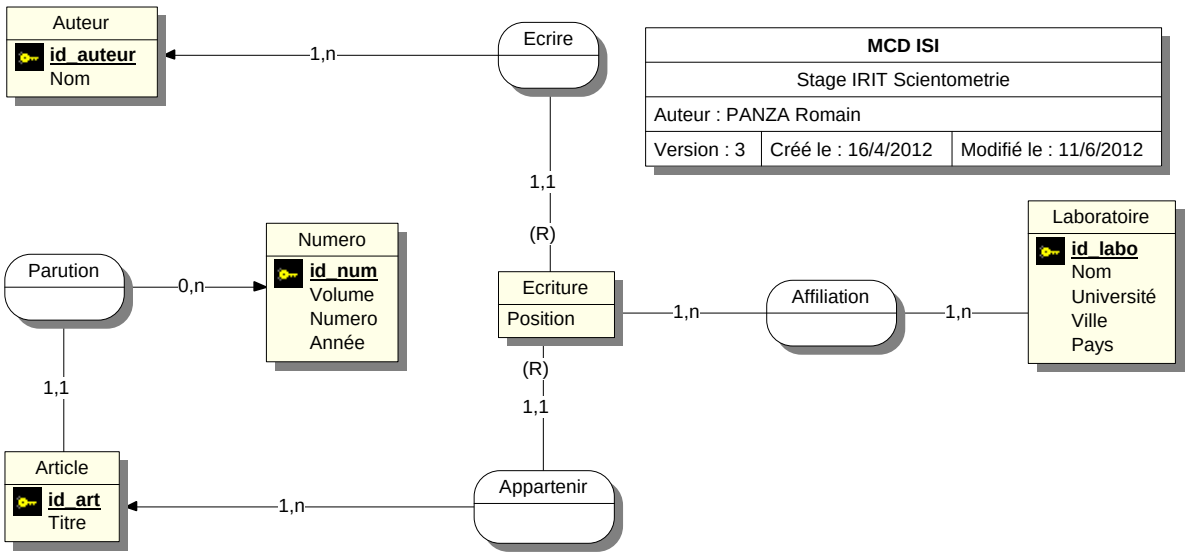

FIGURE 4.1 – Modèle Conceptuel de Données ISI

À partir du modèle conceptuel de données et en fonction des règles de transformation de modèle, on obtient le modèle logique de données suivant :

Auteur	=	{id_auteur, nom}
Article	=	{id_article, #id_num, titre}
Laboratoire	=	{id_labo, nom, universite, ville, pays}
Numero	=	{id_num, volume, numero, annee}
Ecriture	=	{#id_auteur, #id_article, position}
Affiliation	=	{#[#id_auteur, #id_article], #id_labo}

FIGURE 4.2 – Modèle Logique de Données ISI

Nous avons implémenté le MLD sur une base de données *Oracle*. Pour s'assurer que notre base de données fonctionne bien et réponde bien aux demandes du maître de stage, nous avons crée un jeu d'essai, avec deux articles test présentant des données exploitant entièrement la base, sont joint en annexe les requêtes du jeu d'essai. La section suivante détaille l'alimentation de la base de données.

4.2 Extraction des données

Pour extraire les données demandé, nous avions à notre disposition un échantillon fourni par le maître de stage sous la forme d'un fichier *pdf* contenant la première page des publications. C'est cette première page qui contient toutes les informations nécessaires à notre base de données. L'extraction des données devait donc être appliquée à ce fichier. Le choix du langage de programmation a été imposé par le maître de stage, qui souhaitait un programme Java, pour optimiser la gestion de document *pdf*.

En étudiant les possibilités de gestion d'un fichier *pdf* en Java, nous avons rapidement trouvé des bibliothèques adéquates capables de répondre aux demandes, comme PDFbox [9] développé par Apache. Par contre, extraire du texte directement depuis le fichier *pdf* s'est relevé difficile :
- En raison des problèmes d'encodage en fonction des versions d'*Adobe Acrobat* utilisées.
- Et parce que le fichier que nous manipulons a été constitué à partir de plusieurs fichiers *pdf*.

Nous avons donc choisi de transformer l'ensemble du fichier *pdf* en un document texte brut, fichier simple à utiliser et sans formatage particulier.

Une fois le fichier *pdf* transformé en texte, nous avons décidé de structurer notre programme Java en plusieurs classes pour répondre aux objectifs. Nous avons donc créé un organigramme, représentant l'ensemble des fonctions et procédures cf figure 4.5. En regardant cet organigramme on s'aperçoit que le programme n'est pas orienté objet, comme le favorise le langage Java, puisque à la conception même de notre programme nous l'avons réfléchi procédural. Il suffisait d'un ensemble de fonctions et de procédures pour récupérer les informations et les insérer. Nous avons en conséquence, décidé de faire de nos classes des singletons [10].

Notre programme comprend 4 classes, leurs rôles se rapprochent de ceux des paquetages, rassembler des procédures concourant à un même traitement.
Rôles résumés ci-après :
- **ISI** : C'est le programme principal qui appelle les méthodes des autres classes.
- **PdfGestionnaire** : Cette classe sert à transformer le fichier *pdf* en fichier texte et à créer l'ensemble des "*pageX.txt*".
- **BddOracle** : Gère les connexions aux bases de données DBLP et la base qui stocke les données extraites qui sont deux bases de données distinctes. Gère plusieurs fonctions d'insertions dans la

9. PDFbox est une API favorisant la manipulation et la création de fichiers *pdf* en Java.
10. Singleton : Conception au sein d'une classe dont on veut s'assurer qu'il n'existera qu'une et une seule instance dans l'espace et dans le temps du cycle de vie de l'application.

base ISI et vérifie si des auteurs existent déjà.
– **Extracteur** : Permet d'extraire toutes les informations utiles à insérer dans la base ISI, sauf les laboratoires.

Pour notre traitement, nous avons analysé qu'il fallait traiter le fichier page par page plutôt que l'ensemble des pages (puisqu'il se compose de 316 pages). Il nous a semblé plus approprié et facile de créer à partir du fichier texte contenant toutes les pages des fichiers nommés « *pageX.txt* » ou X est le numéro de la page. Nous avons développé fonction nous a permis de découper les pages et de créer les fichiers "*pageX.txt*".

Une fois toutes nos publications dans des fichiers séparés, il faut que notre programme traite page par page l'extraction de chaque information nous intéressant, et l'insère dans notre base de données.

4.2.1 Étude de la structure des pages

Pour pourvoir extraire les informations dans un fichier texte, nous devons analyser la structure générale des articles, pour savoir où se trouvent les informations que nous voulons. Ces articles ont été rédigés sur une longue période (10 ans). Ce délai à pour conséquence de produire des changements de structure et des cas particuliers (approximativement 5% de cas particuliers) et seront détaillés dans la section 4.2.4.

Présentation de la structure type d'une page dans la figure 4.3.

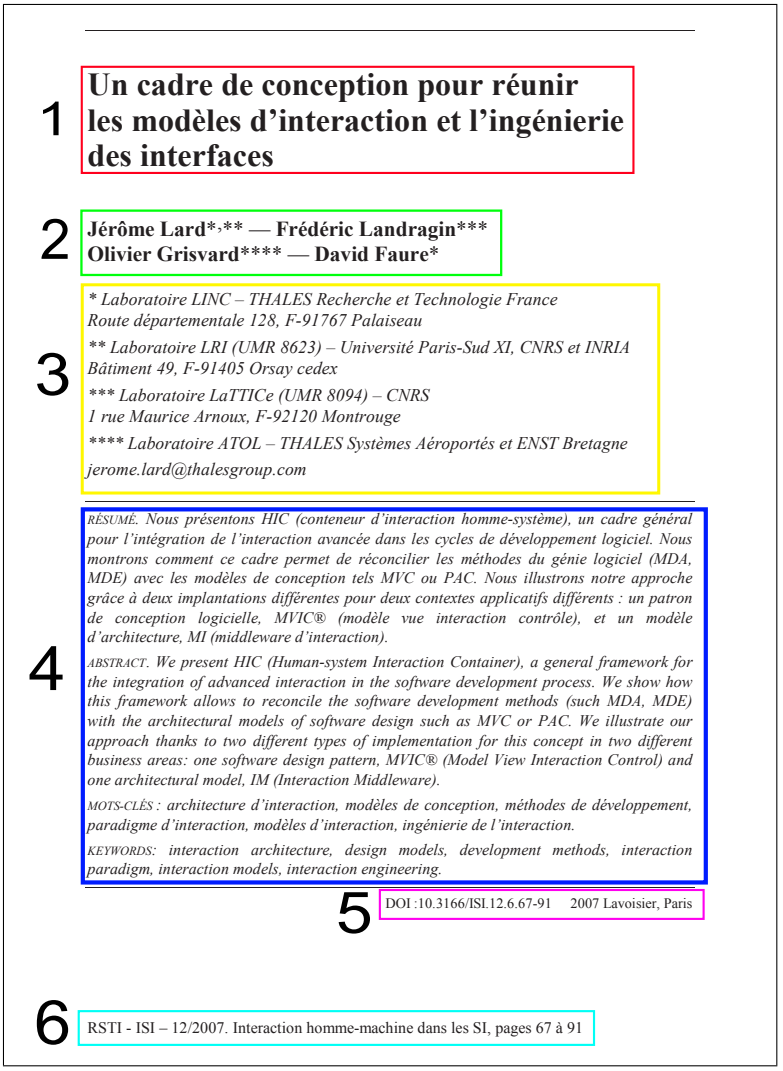

FIGURE 4.3 – Exemple type de première page d'un article

Légende :

1. Cadre rouge : Titre de l'article
2. Cadre vert : Auteur(s) de l'article
3. Cadre Jaune : Laboratoire(s)
4. Cadre bleu : Résumé de l'article
5. Cadre rose : Information sur la revue ISI
6. Cadre cyan : Autres information sur la revue ISI

Nous pouvons noter un certain nombre de remarques sur cette structure :

– Sur le cadre 2, pour les auteurs : un auteur est tout le temps rattaché à un ou plusieurs laboratoire(s) grâce aux étoiles à côté de chaque nom d'auteur. Lorsque tous les auteurs sont affiliés au même laboratoire il n'y a pas d'étoile. Dans le cas où un auteur est rattaché à plusieurs laboratoires, une suite d'étoiles séparées par une virgule apparait à côté de son nom.
– Sur le cadre 3 pour les laboratoires : un laboratoire est une adresse postale, ce qui pose le problème d'être non normalisée.
– Sur le cadre 4 pour le résumé : toute cette partie de la page ne nous est pas nécessaire elle sera ignorée, ou supprimée en fonction du traitement à faire.

– Sur le cadre 5 : ce cadre n'existe que sur les articles publié après 2007. Il contient le numéro, le volume et l'année de la revue, information utile pour la table « numéro ».
– Sur le cadre 6 : ce cadre existe sur toutes les pages et contient l'année et le numéro des revues.

Cette structure uniforme est respectée pour quasiment tous les articles et va faciliter leur extraction. Il faudra néanmoins prendre en compte les quelques cas particuliers, qui seront traité manuellement.

4.2.2 Utilisation des informations à notre disposition

Pour extraire les informations que nous voulons, il faut être capable de pouvoir se repérer dans le texte, et savoir où l'on se situe dans la structure du texte. Pour répondre à ce problème, deux solutions ont été utilisées :

– Le repérage par caractères spéciaux. Dans chaque page sont placés de nombreux caractères spéciaux, notre programme va se repérer grâce à eux (comme le tiret long « — », ou encore l'astérisque « * »)
– L'utilisation d'une copie locale de la base de données DBLP, qui stocke tous les auteurs des publications que nous gérons, et dont le nom a été normalisé. Cela nous permet grâce à une fonction que nous avons crée de retrouver les auteurs.

4.2.3 Extraction des laboratoires

L'extraction des laboratoires a été la tâche la plus laborieuse. Pour saisir le problème de l'extraction des laboratoires voir l'exemple d'un article figure 4.3 et observer que l'écriture d'une adresse postale n'est pas normalisée. Nous avons donc conclu avec le maître de stage qu'il était très difficile, voire impossible d'extraire automatiquement les laboratoires à partir d'adresses postales non normalisées, pour avoir un résultat sans doublon et uniforme pour chaque laboratoire.
Nous avons donc dû saisir l'ensemble des requêtes d'insertion des laboratoires manuellement, avec une grande attention pour éviter les doublons.

4.2.4 Problème d'encodage et de structure de page

Comme précisé dans la section 4.2.1, la structure de la page permet à notre programme de se repérer et d'extraire les données utiles. Néanmoins deux situations restaient à résoudre :

– en cas de problème d'encodage, à cause du fichier *pdf* transformé en texte, certaines pages deviennent complètement illisibles ou certains caractères font stopper notre programme (moins de 20 pages). Il a donc fallu les traiter manuellement.
– en cas de structure différente, notre programme fonctionne si la structure détaillée section 4.2.1 est parfaitement respectée, malheureusement une dizaine de pages ne respectent pas cette structure. Là aussi il a fallu les traiter manuellement.

4.2.5 Mise en place de l'extraction

Notre classe Extracteur fonctionne sur le principe de traitement page par page, pour l'extraction des données. Ensuite il y a une fonction adaptée à l'extraction de chaque information pour la base de données ISI. Ainsi comme on peut le voir dans la figure 4.5, nous avons une liste de fonctions se complétant pour arriver à nos fins.
Par exemple, pour extraire les auteurs prêts pour être insérés dans la base de données ISI :

– Il faut utiliser la fonction *getContenuPage* qui renvoie tous les caractères de la page.
– ensuite on applique sur le résultat de *getContenuPage* les fonctions *couperResumer* et *couperTitre*.
– une fois obtenu ce résultat, on le passe en paramètre à la fonction *getAuteurBrut*, qui nous retourne les auteurs avec leurs éventuelles étoiles.
– pour terminer il faut donner en paramètre le résultat de *getAuteurBrut* à *getAuteurUtilInsert*.

Cette enchainement est illustrée par la figure 4.4.

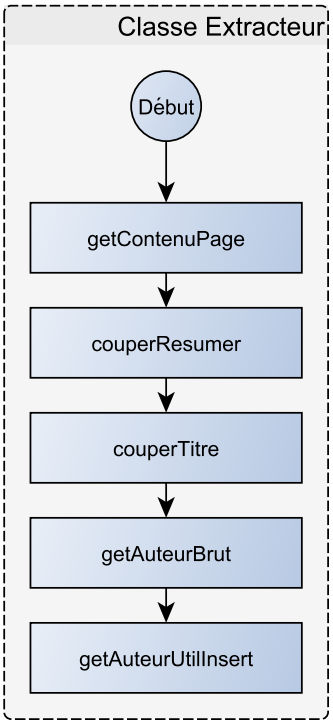

FIGURE 4.4 – Organigramme du programme pour extraire les auteurs prêt pour être insérés

4.3 Insertion des données

Grâce à la classe Extracteur nous avons toutes les donnée. La classe BddOracle va insérer les données dans la base ISI. Comme détaillé dans la section 4.2.4 l'insertion des laboratoires s'est effectuée manuellement. Nous avons écrit un fichier *sql* avec l'ensemble des insertions à faire. Nous avons donc une fonction *insertDonneesAuteursArticleNumeroEcriture* qui complète à elle seule l'ensemble des tables Auteur, Article, Numero et Ecriture. Il reste à compléter la table Affiliation qui permet de faire le lien entre les tuples de la table Ecriture et Laboratoire.

Nous nous sommes aperçus qu'il manquait à notre programme le lien entre le nombre d'astérisques sur un laboratoire et son identifiant dans la base de données. Une fois cette information connue, notre programme pouvait automatiser le lien entre le/les laboratoire(s), le/les auteur(s) , la création des requêtes d'insertion dans la table Affiliation et les insérer pour nous.

Nous avons créée la procédure *creerInsertAffiliation* dans la classe BddOracle qui demande à l'utilisateur de saisir les identifiants des laboratoires pour une page donnée. Grâce à ces informations, elle génère les requêtes d'affiliations pour une page dont le chemin d'accès est précisé en paramètre et écrit dans un fichier les requêtes d'affiliation, lui aussi donné en paramètre. Ensuite une autre procédure écrite dans la classe ISI, *ecrireToutLesInsertAffiliation* qui appelle la procédure *creerInsertAffiliation*, boucle pour traiter toutes les pages et ainsi demander manuellement tous les identifiants des laboratoires pour chaque page.

Une fois se traitement fait, la base de données ISI est alimentée.

4.4 Limites

Le programme que nous avons développé dans les sections précédentes présente comment nous avons pu créer une base de données utilisable pour des traitements futurs. La quantité de données étant limitée (316 articles) nous n'avons pas eu la nécessité d'optimiser notre code ni de mettre en place des tests d'intégration. Nous avons testé notre programme manuellement, par le biais d'un jeu d'articles type représentant l'ensemble des cas possibles, que nous testions régulièrement sur l'ensemble de notre programme. Par conséquent nous ne pouvons affirmer la fiabilité totale des informations dans cette base de données, par l'absence de test nous ne disposons pas de chiffres comme le taux de fiabilité des données.

Il serait donc intéressant dans une perspective d'amélioration du projet de prévoir des tests d'intégrations qui permettraient de s'assurer de la fiabilité du code et aussi des tests de fiabilité des données.

4.5 Bilan de la mission

A première vue cette mission ne me semblait pas longue à réaliser et sans difficulté majeure, je pensais pouvoir la résoudre rapidement. J'ai dû faire face à des imprévus et des difficultés techniques qui ont largement ralenti le déroulement de cette mission. Cependant l'objectif de la missions fus entièrement rempli. La base de données ayant été livré à mon maitre de stage avec l'ensemble des éléments pour pouvoir la reproduire.

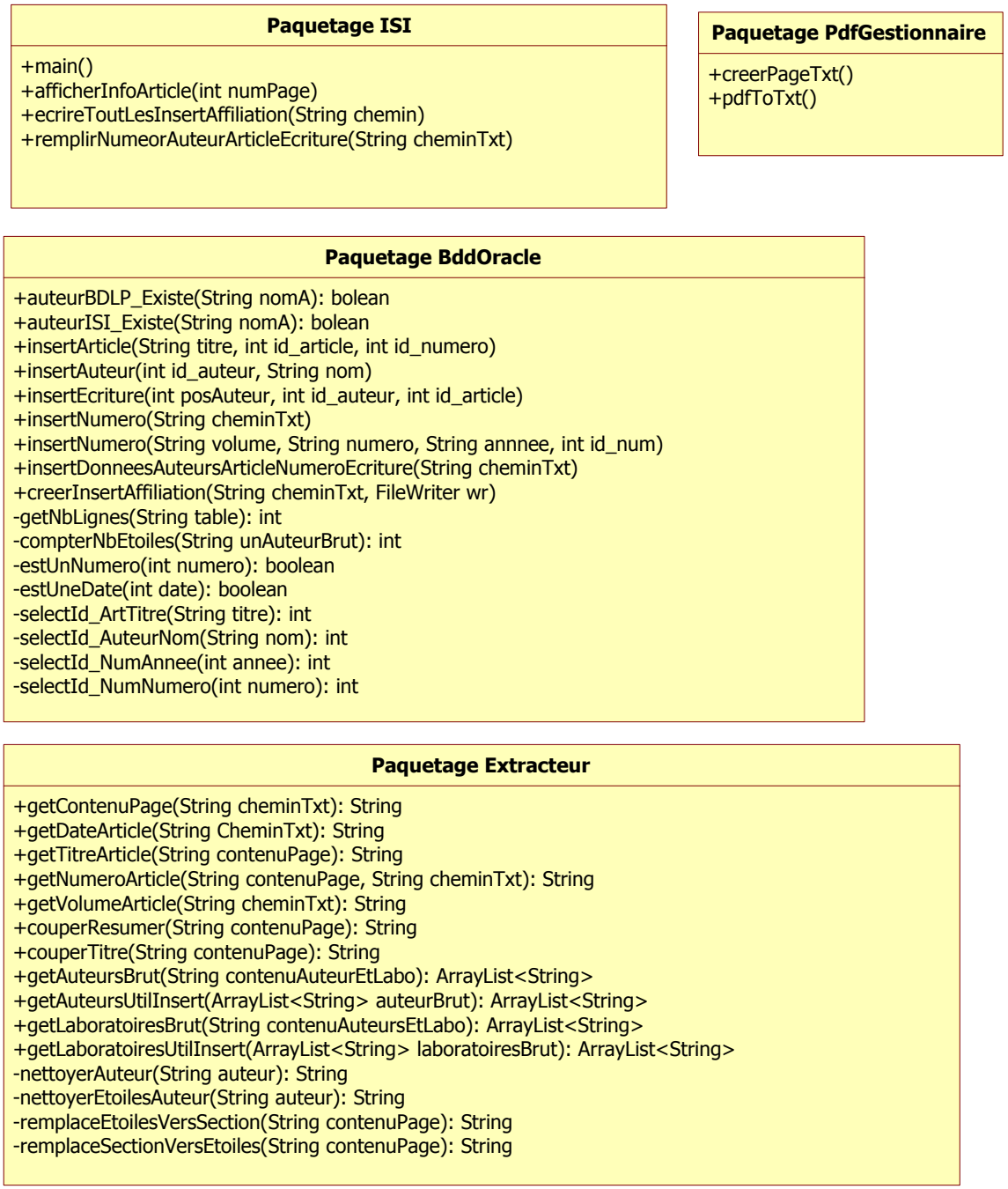

FIGURE 4.5 – Présentation des paquetages ISI

Chapitre 5

Mission ACM

Ce chapitre va vous décrire comment nous avons résolu cette mission. Pour plus de détails sur les objectifs de cette mission voir section 2.4.3.

5.1 Extraction des données d'une page HTML

L'utilisation, pour l'extraction des données d'une page *HTML*, de *HtmlUnit* et *jQuery* fut conseillée par le maître de stage. *HtmlUnit* est une extension de *JUnit*[11], et propose de tester une application. Il est décrit comme un navigateur web sans interface. Il offre la possibilité d'appeler les éléments *HTML* d'une page et ainsi simuler une navigation web. Le développeur peut ainsi remplir des formulaires, cocher des *checkbox* ou simuler un clic sur un bouton. Cette extension de *JUnit* est donc tout à fait adaptée à cette mission dans laquelle il a fallu naviguer sur le site internet d'ACM. *jQuery* est une bibliothèque de *JavaScript* et *HTML*, son but est de simplifier les commandes communes de *JavaScript*.

Après une phase de documentation sur *HtmlUnit* et *jQuery*, nous avons développé un programme capable d'aller sur le site d'ACM et de naviguer sur les pages web. Malheureusement ayant pris du retard sur la mission ISI, cette mission d'extraction de données d'une page *HTML* n'a pas pu être terminer à l'heure de la rédaction de ce rapport. Néanmoins au vu de l'avancement du travail sur cette dernière mission (40% effectué), il sera terminé à la fin de la dernière semaine de stage.

5.2 Bilan de la mission

L'extraction de données directement depuis une page *HTML* étant nouveau dans ce stage, il a fallu un travail de documentation essentiel sur *HtmlUnit* et *jQuery* qui sont des bibliothèques issues de langage de programmation.

11. JUnit : Bibliothèque de test unitaire pour la langage de programmation *Java*

Chapitre 6

Conclusion

Pendant ce stage, nous nous sommes intéressés à l'extraction d'information sur des données scientifiques en lien directement ou indirectement, avec leurs publications et leurs indexations en vu de traitements ultérieurs.

La continuité de ce stage aurait été de produire les requêtes afin d'étudier les données extraites, et ainsi pouvoir étudier les liens entres la rédaction de publication et le laboratoire par l'intermédiaire de graphiques par exemple.

Pour aller plus loin on pourrait développer une application qui, en fonction du domaine de recherche des laboratoires et de leur collaborations pourrait définir de nouvelles collaborations pertinentes.

De plus notons que sur les trois missions à effectuer, deux sont d'ores et déjà totalement réalisées et livrées, la troisième est pratiquement terminée et sera bouclée en dernière semaine de stage. Ce rapport étant à remettre avant que le stage soit effectivement terminé, nous pouvons acter que nos objectifs de ce stage sont atteints.

Chapitre 7

Bilan technique et personnel

J'ai commencé ce stage en vu d'un double objectif personnel, le premier valoriser ma formation en DUT informatique, afin d'y apporter un supplément à mon CV, professionnel et original. En effet je trouve le stage au sein de l'IRIT un très bon défi, car c'est un environnement tout aussi professionnel qu'une entreprise en lien avec l'informatique puisque le fonctionnement de l'IRIT est comparable à celui une entreprise. De plus, avoir travaillé au sein d'un laboratoire de recherche m'offrait l'opportunité de découverte d'un monde qui m'était inconnu.

Au cours de ce stage, j'ai eu l'occasion d'appréhender de nouvelles connaissances comme le XSLT, langage de transformation de fichier XML, le HtmlUnit ou encore LaTeX. J'ai ensuite mobilisé mes connaissances acquises à l'IUT. J'ai dû aussi fournir un travail régulier de rédaction de compte-rendus d'avancement sur mon travail. Ce dernier m'a contraint à faire un effort rédactionnel de synthèse sur mon bilan hebdomadaire. Ce type de compétence n'avait pas été développé à l'IUT.

Mon second objectif personnel était de découvrir le monde de la Recherche que l'on est amené à côtoyer au sein de l'IUT. Pendant mes deux années de formation en DUT, j'ai eu l'occasion de discuter avec de nombreux enseignants-chercheurs qui ont suscité ma curiosité pour ce métier et la démarche scientifique qu'il développe.

Ensuite d'un point de vue personnel, ce fut la première fois depuis mon entrée à l'IUT que j'ai dû mener un projet informatique d'un bout à l'autre, seul. Puisque la formation au sein de l'IUT privilégie le travail en binôme ou en quadrinome, je n'ai jamais eu à gérer de manière autonome un projet. La formation de l'IUT est positive est stimulante, puisque j'aime le travail en équipe. Ce dernier permet de mettre en avant les compétences de chacun et la répartition des tâches en fonction des membres d'un groupe. Ici j'ai géré seul l'ensemble du projet (sauf aide ponctuelle de la part de mon maître de stage ou de la part d'autres stagiaires). Ce travail m'a donc permis de développer mes compétences transversales dans la gestion d'un projet d'un bout à l'autre.

Étant seul, mais au sein d'une équipe de recherche, et devant faire mon compte-tenu hebdomadaire, j'ai aussi développé mes compétences en communication puisqu'avec mon maître de stage nous avons effectué plusieurs analyses sur mon expression orale (en vu de l'améliorer constamment). Ces exercices d'expression ont permis de maintenir un bon contact avec mon maître de stage et de m'assurer que mon travail s'orientait bien vers ce qu'il souhaitait. De plus, les échanges avec d'autres stagiaires de l'IRIT de l'équipe SIG de tous les niveaux (master, doctorat) furent un partage intéressant sur les différents parcours professionnels qui s'offrent à moi. Ceci m'a conduit à davantage réfléchir à mon avenir et à m'a poursuite d'études.

Ce stage m'a permis de parfaire ma formation professionnelle et technologique en informatique. Je reste indécis sur ma poursuite d'études et mon parcours professionnel. Ce stage m'a apporter des éléments nouveaux, nourrissant ma réflexion. Je reste partagé entre études longues et courtes. Prenant un compte les arguments sur la poursuite d'études je pense, à l'heure actuelle, poursuivre mes études au moins jusqu'au master.

Bibliographie

[Cab12] Cabanac G. : Shaping the landscape of research in Information Systems from the perspective of editorial boards : A scientometric study of 77 leading journals. *Journal of the American Society for Information Science and Technology*, 63(5):977–996, mai 2012.

[HW01] Hood W. et Wilson C. : The literature of Bibliometrics, Scientometrics, and Informetrics. *Scientometr.*, 52(2):291–314, 2001.

[ST05] Soutou C. et Teste O. : *SQL pour Oracle.* Eyrolles, 2005.