



Université  
Paul Sabatier  
TOULOUSE III

# Rapport de stage

Version 6

---

Étude scientométrique sur la  
représentation des femmes dans  
les comités de rédaction de  
journaux scientifiques

Ophélie FRAISIER

Du 8 avril au 28 juin 2013



Maître de stage :  
Guillaume Cabanac

Tuteur de stage :  
Sébastien Gerchinovitz





# Rapport de stage

Version 6

---

Étude scientométrique sur la  
représentation des femmes dans  
les comités de rédaction de  
journaux scientifiques

Ophélie FRAISIER

Du 8 avril au 28 juin 2013



Maître de stage :  
Guillaume Cabanac

Tuteur de stage :  
Sébastien Gerchinovitz



## Historique des versions de ce document

Version	Cause de la modification	Chapitre(s) concerné(s)
1	Création du document	1 – 3
2	Ajout des chapitres 4 et 6	4, 6
3	Ajout des chapitres 5 et 7	5, 7
4	Modification du chapitre 5	5
5	Finalisation des chapitres 5 et 7	5, 7
6	Ajout du chapitre 8	8



Je tiens à remercier ici toutes les personnes m'ayant permises de réaliser ce stage.

En premier lieu je remercie Michel Daydé, directeur de l'IRIT, de m'avoir acceptée comme stagiaire au sein de son établissement et Josiane Mothe, responsable de l'équipe SIG, de m'avoir permise de rejoindre son équipe.

Je remercie ensuite Karen Pinel-Sauvagnat et Xavier Gendre, ainsi que tous les autres membres de l'équipe pédagogique nous ayant encadrés tout au long de l'année, pour les connaissances qu'ils nous ont apportées. Je tiens également à remercier Sébastien Gershinovitz, mon tuteur de stage, pour son aide.

Enfin je tiens à remercier tout particulièrement Guillaume Cabanac, mon maître de stage, tout d'abord pour m'avoir proposé ce stage, mais également pour l'aide et les conseils qu'il m'a apporté. Il a toujours été présent et attentif à mes remarques ou suggestions, et ce fut un véritable plaisir de travailler à ses côtés durant trois mois.



# Table des matières

<b>1</b>	<b>Objet et but du document</b>	<b>13</b>
1.1	Présentation du projet	13
1.2	Présentation du document	13
<b>2</b>	<b>Documents de référence</b>	<b>15</b>
2.1	<i>Shaping the landscape of research in information systems from the perspective of editorial boards: A scientometric study of 77 leading journals</i>	15
2.2	Cours de concepts fondamentaux de Bases de Données	15
2.3	Cours d'optimisation de requête	15
2.4	Cours de langage de requêtes	15
2.5	Cours de Statistiques Exploratoires et Inférentielles	15
2.6	DUT Informatique	16
<b>3</b>	<b>Terminologie</b>	<b>17</b>
3.1	Projet	17
3.2	Base de données	18
3.3	Statistiques	18
<b>4</b>	<b>Organisation</b>	<b>19</b>
4.1	Entreprise	19
4.2	Équipe du projet	19
4.3	Planification	19
4.3.1	Planning prévisionnel	19

4.3.2	Planning effectif . . . . .	19
<b>5</b>	<b>Inforsid . . . . .</b>	<b>21</b>
<b>5.1</b>	<b>Présentation du contexte</b>	<b>21</b>
<b>5.2</b>	<b>Intégration des données des éditions 2012 et 2013</b>	<b>22</b>
5.2.1	Fonctionnement de la transformation des données . . . . .	22
5.2.2	Modifications apportées au processus de traitement des données . . . . .	23
5.2.3	Gestion des «synonymes» . . . . .	23
5.2.4	Gestion des pays . . . . .	24
<b>5.3</b>	<b>Valorisation du congrès</b>	<b>24</b>
5.3.1	Pôles principaux . . . . .	24
5.3.2	Thèmes du congrès . . . . .	25
5.3.3	Influence des membres . . . . .	28
5.3.4	Villes ayant accueilli le congrès . . . . .	28
<b>5.4</b>	<b>Modification du site web</b>	<b>28</b>
5.4.1	Optimisation des procédures PL/SQL . . . . .	28
5.4.2	Création d'un nouveau style graphique . . . . .	32
<b>6</b>	<b>Étude de genre . . . . .</b>	<b>35</b>
<b>6.1</b>	<b>Présentation du contexte</b>	<b>35</b>
<b>6.2</b>	<b>Récupération et mise en forme des données</b>	<b>35</b>
<b>6.3</b>	<b>Présentation des données et des tests utilisés</b>	<b>36</b>
6.3.1	Présentation générale des données . . . . .	36
6.3.2	Présentation des tests statistiques . . . . .	38
<b>6.4</b>	<b>Comparaison de générations</b>	<b>38</b>
<b>6.5</b>	<b>Conclusion de l'étude</b>	<b>43</b>
<b>7</b>	<b>Méthodes et outils utilisés . . . . .</b>	<b>45</b>
<b>7.1</b>	<b>Base de données</b>	<b>45</b>
7.1.1	Oracle Database . . . . .	45
7.1.2	SqlDeveloper . . . . .	45
7.1.3	Sql*Loader . . . . .	45
<b>7.2</b>	<b>Analyse et mise en forme des données</b>	<b>46</b>
7.2.1	Sofa Statistics . . . . .	46
7.2.2	R et RStudio . . . . .	46
7.2.3	Gnuplot . . . . .	46
7.2.4	L <sup>A</sup> T <sub>E</sub> X . . . . .	46
<b>7.3</b>	<b>Gestion de configuration</b>	<b>47</b>
7.3.1	Apache Subversion . . . . .	47
<b>8</b>	<b>Assurance et contrôle qualité . . . . .</b>	<b>49</b>
<b>8.1</b>	<b>Compte-rendus hebdomadaires</b>	<b>49</b>
<b>8.2</b>	<b>Réunions avec mon maître de stage</b>	<b>49</b>

<b>9</b>	<b>Bilan . . . . .</b>	<b>51</b>
9.1	<b>Bilan du projet</b>	<b>51</b>
9.2	<b>Bilan personnel</b>	<b>51</b>
9.3	<b>Conclusion</b>	<b>51</b>



# 1 — Objet et but du document

## 1.1 Présentation du projet

Afin de valider ma troisième année de licence j'ai du réaliser un stage afin de mettre en pratique les connaissances acquises. J'ai choisi de me tourner vers la recherche afin de voir l'autre facette du métier d'enseignant-chercheur, et ai donc effectué mon stage à l'IRIT<sup>1</sup>, dans le cadre des recherches de Guillaume Cabanac en scientométrie.

Le projet visait à approfondir l'étude des membres des comités de rédaction des revues scientifiques – ou gatekeepers – initiée dans l'article scientifique *Shaping the landscape of research in information systems from the perspective of editorial boards: A scientometric study of 77 leading journals* (Cabanac, 2012). La question de la représentation des femmes dans ces comités était au centre du projet. Cette question revêt un intérêt tout particulier pour la communauté scientifique en informatique.

Afin de me familiariser avec les termes employés dans la communauté scientifique et les techniques scientométriques ma première mission était de valoriser le congrès Inforsid, en déterminant notamment ses principaux thèmes et la contribution des villes impliquées.

Ma seconde mission était l'étude de genre à proprement parler, tout d'abord par l'étude de la distribution de plusieurs variables selon le genre des membres – par exemple, la capacité à entretenir des collaborations scientifiques mesurée par le  $\phi$ -index (Cabanac, 2012) ou le nombre d'articles publiés – puis par la valorisation des résultats obtenus par la rédaction d'un document synthétique présentant les résultats obtenus et leur discussion vis-à-vis de la littérature en matière d'étude de genre portant sur des scientifiques.

## 1.2 Présentation du document

Ce document va vous présenter le travail réalisé durant mon stage et notamment les thématiques abordées.

Je vous présenterai tout d'abord les documents m'ayant servi de base durant ce stage, ainsi que les termes nécessaires à la bonne compréhension du rapport. J'exposerai ensuite l'organisation de mon travail.

Par la suite je décrirai le travail réalisé durant ce stage ainsi que les méthodes et outils utilisés. J'exposerai également le suivi organisé avec mon maître de stage.

---

<sup>1</sup>Institut de Recherche en Informatique de Toulouse

Enfin je conclurai en faisant le bilan de ce stage.

## 2 — Documents de référence

### 2.1 *Shaping the landscape of research in information systems from the perspective of editorial boards: A scientometric study of 77 leading journals*

L'article scientifique (Cabanac, 2012) pose les bases utilisées lors de cette étude de genre.

Il se concentre sur l'étude des comités de rédaction de 77 journaux scientifiques du domaine *Science de l'Information* et discute divers indicateurs scientométriques à l'aide de statistiques descriptives. Les résultats de cet article, présentant la diversité des membres de comités de rédaction, m'a incité à proposer à Guillaume Cabanac l'étude de genre présentée ici.

Vous pourrez trouver l'article dans les annexes.

### 2.2 Cours de concepts fondamentaux de Bases de Données

Le cours de «Concepts Fondamentaux de Bases de Données» de Ms Morvan et Mokadem et Mme Yin m'a été utile afin de comprendre la structure des bases de données que j'ai eu à manipuler durant ce stage.

### 2.3 Cours d'optimisation de requête

Le cours d'«Optimisation de requête» de Ms Hameurlain et Morvan et Mme Yin m'a permis de comprendre les mécanismes d'optimisation mis en place sur certaines des bases de données que j'ai eu à utiliser.

### 2.4 Cours de langage de requêtes

Le cours de «Langage de requêtes» de Mme Pinel-Sauvagnat m'a été indispensable durant ce stage. En effet toutes mes données étaient stockées dans des bases de données et il a fallu que je les extraie mais également que je les traite à l'aide de procédures PL/SQL.

### 2.5 Cours de Statistiques Exploratoires et Inférentielles

L'extraction des données était la première étape de mon stage mais ma tâche principale était l'analyse de celles-ci. Pour cela les cours de «Statistique exploratoire» et de «Statistique Inférentielle» de M. Gendre ont été salutaires pour moi.

## 2.6 DUT Informatique

La formation que j'ai reçu durant mon DUT Informatique m'a été utile, tout spécialement les cours portant sur les bases de données de Mme Bensadoun. En effet j'ai eu à utiliser l'Oracle Web Toolkit avec lequel j'avais déjà travaillé dans le module «Bases de Données avancées».

## 3 — Terminologie

### 3.1 Projet

<i>IS</i>	Information Systems – domaine de recherche traitant de la collecte et du traitement d’informations.
<i>AI</i>	Artificial Intelligence – domaine de recherche visant à trouver des moyens susceptibles de doter les systèmes informatiques de capacités intellectuelles comparables à celles des êtres humains.
<i>Scientométrie</i>	Étude de la science par une démarche scientifique.
<i>Comité de rédaction</i>	Ensemble de chercheurs responsable des choix de publication d’un journal scientifique.
<i>5YJIF</i>	5-year Journal Impact Factor – mesure indiquant le nombre moyen de citations de chaque article publié par le journal, servant à mesurer la visibilité des revues scientifiques.
<i>Gatekeeper</i>	Nom donné en anglais aux membres de comité de rédaction des journaux scientifiques.
<i>DBLP</i>	Digital Bibliography & Library Project – site web publiant des notices bibliographiques en informatique hébergé par l’Université de Trèves en Allemagne existant depuis les années 1980.
<i>Inforsid</i>	INformatique des ORganisations et Systèmes d’Information et de Décision – Congrès réunissant des chercheurs en <i>Information Science</i> – Science de l’Information – depuis 1983.
<i>Comité de programme</i>	Ensemble de chercheurs décidant des thèmes des différentes sessions d’un congrès et des articles présentés durant celui-ci. Ce comité est constitué d’un ou plusieurs président(es) ayant pour premières tâches de choisir le reste des membres et occasionnellement des adjoint(e)s.
<i>Mot vide</i>	Mot non porteur de sens qu’il est inutile d’indexer ou d’utiliser dans une recherche, dépendant de la langue du texte.

### 3.2 Base de données

<i>BD</i>	Base de Données, ensemble structuré et organisé permettant le stockage de grandes quantités d'informations afin d'en faciliter l'exploitation (ajout, mise à jour, recherche de données).
<i>SGBD</i>	Système de Gestion de Base de Données, logiciel système destiné à gérer la définition, manipulation, cohérence, confidentialité, intégrité, sauvegarde et restauration des données et la gestion des accès concurrents, tout en cachant la complexité des opérations.
<i>SQL</i>	Structured Query Language – langage permettant de créer, modifier et interroger les tables d'une base de données.
<i>Table</i>	Structure stockant des données, sous forme de tuples, un tuple étant un enregistrement.
<i>Déclencheur</i>	Procédure provoquant un traitement particulier en fonction d'événements prédéfinis, permettant ainsi d'automatiser certains traitements assurant la cohérence et l'intégrité de la base de données.

### 3.3 Statistiques

<i>Test d'hypothèse</i>	Démarche consistant à rejeter ou à ne pas rejeter une hypothèse statistique, appelée hypothèse nulle, en fonction d'un jeu de données (échantillon).
<i>Hypothèse nulle (<math>H_0</math>)</i>	Point de vue par défaut concernant un phénomène donné. Il est nécessaire de savoir comment se comporte l'échantillon sous l'hypothèse nulle afin de pouvoir réaliser un test.
<i>p-valeur</i>	Probabilité d'obtenir la même valeur (ou une valeur encore plus extrême) du test si l'hypothèse nulle était vraie. Si cette valeur est inférieure à la valeur du seuil préalablement défini (traditionnellement 5% ou 1%), on rejette l'hypothèse nulle. En d'autres termes, la p-valeur est la probabilité de rejeter à tort l'hypothèse nulle et donc d'obtenir un faux positif.

## 4 — Organisation

### 4.1 Entreprise

J'ai effectué mon stage au sein de l'IRIT<sup>1</sup>, une unité mixte de recherche comprenant 19 équipes de recherche réparties selon sept thèmes :

1. Analyse et synthèse de l'information
2. Indexation et recherche d'informations
3. Interaction, autonomie, dialogue et coopération
4. Raisonnement et décision
5. Modélisation, algorithmes et calcul haute performance
6. Architecture, systèmes et réseaux
7. Sûreté de développement du logiciel

J'étais pour ma part intégrée à l'équipe SIG<sup>2</sup> dépendant du thème 2 «Indexation et recherche d'informations». Cette équipe est elle-même organisée autour de quatre axes de recherches :

- Conception de systèmes d'informations décisionnels
- Documents, Données Semi-Structurées et usages
- Exploration et Visualisation d'Information
- Modèles adaptatifs pour la recherche d'information

### 4.2 Équipe du projet

J'ai travaillé durant ce stage en collaboration avec mon maître de stage, Guillaume Cabanac, maître de conférence. Nous nous sommes basés sur ses précédents travaux en scientométrie.

### 4.3 Planification

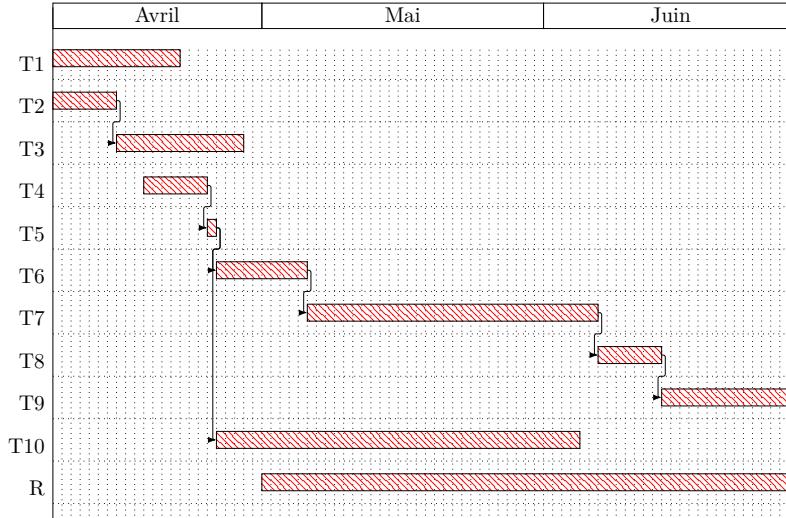
#### 4.3.1 Planning prévisionnel

Le diagramme de Gantt prévisionnel de mon stage est visible sur la figure 4.1.

#### 4.3.2 Planning effectif

<sup>1</sup>Institut de Recherche en Informatique de Toulouse

<sup>2</sup>Systèmes d'Informations Généralisés



*Tâche Description*

T1	Reprise de l'application webOracle présentant Inforsid existante afin de l'améliorer.
T2	Insertion des données d'Inforsid 2012 et 2013.
T3	Valorisation d'Inforsid (par l'utilisation de graphiques représentant les villes les plus impliquées ou de nuages de mots des thèmes abordés notamment).
T4	Compréhension du parser Java traitant la base XML de DBLP.
T5	Insertion des données à jour de DBLP dans la base de travail.
T6	Insertion des comités de rédaction du domaine IS dans la base de travail.
T7	Étude de genre des <i>boards</i> IS insérés précédemment (répartition homme-femme, partnership coefficient, ...).
T8	Valorisation des résultats obtenus.
T9	Rédaction d'un article scientifique présentant les résultats obtenus et nos conclusions.
T10	Insertion des comités de rédaction du domaine IA <sup>1</sup> dans la base de travail.
R	Rédaction du rapport de stage.

1. Intelligence Artificielle

FIGURE 4.1: Diagramme de Gantt prévisionnel de mon stage

## 5 — Inforsid

### 5.1 Présentation du contexte

Depuis sa création en 1983 le congrès Inforsid réunit chaque année les chercheurs travaillant dans le domaine de l'informatique des organisations. Il se tient chaque année dans une ville différente, et chaque édition dispose de son propre comité de programme, décidant des thèmes des différentes sessions et des articles présentés par les chercheurs. Ce comité est constitué d'un ou plusieurs président(es) ayant pour premières tâches de choisir le reste des membres et occasionnellement des adjoint(e)s.

Tout ceci représente une masse d'information non négligeable, rassemblée pour chaque congrès dans un ouvrage : les *actes du congrès*. Cependant la recherche d'une information particulière dans ces ouvrages peut être laborieuse (participation d'un chercheur à une édition ou chercheurs les plus impliqués dans le comité de programme par exemple).

Pour pallier ce problème Guillaume Cabanac et Marc Ternisien ont mis en place une application Oracle centralisant ces informations. Celle-ci permet, pour chaque édition du congrès, de visualiser son comité de programme (président(es) et membre(s)) et la liste des articles présentés avec leurs auteurs respectifs. Elle présente également une «fiche de présentation» pour chaque chercheur ayant participé à Inforsid au cours des années. Cette fiche permet de voir :

- la liste des articles présentés par le chercheur en question,
- ses participations au comité de programme,
- sa localisation au cours des années (déterminée grâce à ses différentes participations au comité de programme et les articles qu'il a présentés),
- les chercheurs avec qui il a écrit des articles présentés au congrès (avec la(les) année(s) de collaboration).

Une autre fonctionnalité est également présente dans l'application : la suggestion de membres pour la constitution du comité de programme. En effet, jusque là les présidents n'avaient aucune règle ou aide pour sélectionner d'éventuels membres, et devaient donc se fier à leurs connaissances de la communauté. Cela pouvait entraîner l'oubli de certains membres de la communauté, ou la favorisation de certains. La suggestion de membres de l'application se base elle sur un algorithme et permet une constitution plus éclairée, tout en étant certains de n'oublier aucun chercheur. L'algorithme liste les chercheurs :

- ayant présenté au moins un article lors d'un congrès depuis 2005,
- ayant écrit au moins 2 articles,

- n'ayant jamais fait partie d'un comité de programme (ceux-ci étant favorisés) ou en ayant fait partie avant 1995.

Mon travail était d'intégrer dans cette application les données des congrès Inforsid 2012 et 2013 et de valoriser le congrès.

## 5.2 Intégration des données des éditions 2012 et 2013

La transformation des données pour leur insertion dans la base de données se faisait par l'intermédiaire d'un programme en C, dont je devais donc comprendre le fonctionnement. Heureusement Marc Ternisien, le stagiaire ayant initialement développé l'application, avait fourni une documentation claire qui m'a permis de rapidement prendre en main l'application.

### 5.2.1 Fonctionnement de la transformation des données

Les données générales des congrès étaient présentes dans le fichier `Datas/data-Congres.txt` contenant pour chaque édition une ligne de la forme ‘année, ville’. Les informations détaillées étaient rentrées dans 2 fichiers texte placés dans le répertoire `Datas/[année_traitée]` :

- un fichier `membre.txt` contenant les membres du comité de programme, sous la forme ‘prénom, nom, ville’. Ce fichier contenait une ligne par membre, et les groupes des présidents, des simple membres et des adjoints étaient séparés par une ligne vide.
- un fichier `article.txt` contenant les articles présentés durant le congrès avec leurs auteurs, sous la forme ‘titre de l'article’ sur la première ligne puis pour chaque coauteur une ligne contenant ‘prénom, nom, ville’. Les différents articles étaient séparés par une ligne vide.

Les noms de personnes et de villes rentrés dans ces fichiers ne devaient pas contenir de caractères spéciaux (accents ou ç) afin de simplifier la gestion.

Une fois les données insérées dans ces fichiers, l'application C transformait ces fichiers en fichiers texte - dont la structure est présentée dans le tableau 5.1 – destinés à être insérés dans la base de données construite selon le MCD présenté en figure 5.1.

TABLEAU 5.1: Structure des fichiers destinés à être insérés dans la base de données

<i>Fichier</i>	<i>Structure</i>
article.txt	<code>idArticle;titre;année;</code>
congres.txt	<code>année;idVille;</code>
personne.txt	<code>idPersonne;prénom;nom;</code>
ville.txt	<code>idVille;nomVille;paysVille;</code>
ecrire.txt	<code>idArticle;idPersonne;idVille;rang<sup>1</sup></code>
membre.txt	<code>idPersonne;année;rôle<sup>2</sup>;idVille</code>

Toujours afin de simplifier la gestion, les noms de villes, pays et personnes étaient convertis en majuscules à cette étape. Par défaut le pays était placé à ‘FRANCE’ mais il existait 2 méthodes pour corriger ceci avant l'insertion des données dans la base :

- l'utilisateur pouvait modifier la valeur directement dans le fichier `ville.txt` généré précédemment,

<sup>1</sup>Position à laquelle la personne apparaît lors de la déclaration des coauteurs dans l'article.

<sup>2</sup>Président(e) = ‘P’, membre = ‘M’ ou adjoint(e) = ‘A’

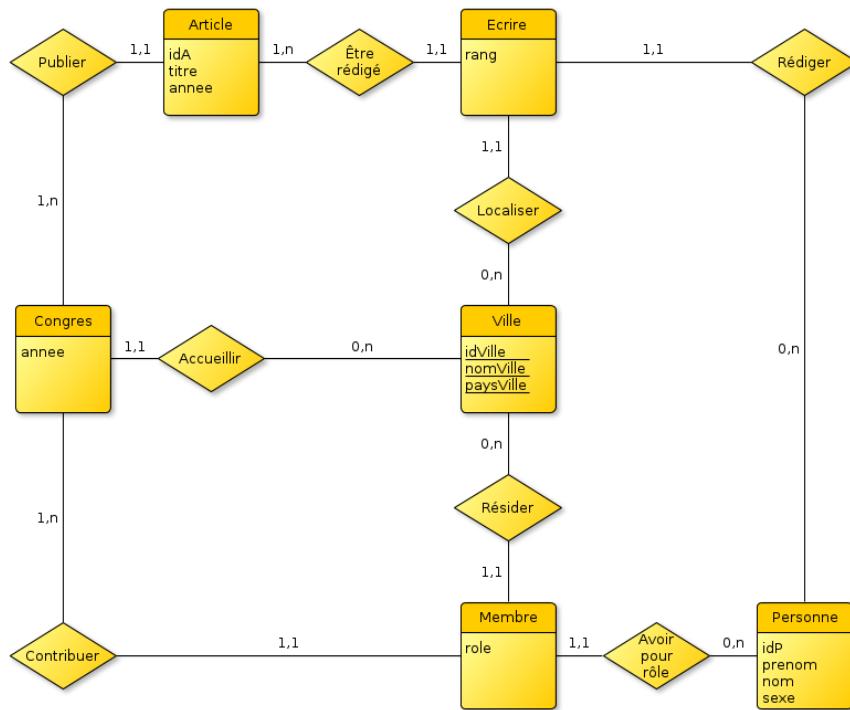


FIGURE 5.1: Modèle Conceptuel de Données de la base utilisée par l’application (*Source : dossier de conception de Marc Ternisien*)

- s'il existait déjà un fichier ville.txt lorsque le programme de transformation de données était exécuté, celui-ci récupérait les pays attribués aux villes présentes dans ce fichier.

Une fois cette étape passée il suffisait de lancer le script SqlLoader/loadData.sh permettant d'insérer ces données dans la base par l'intermédiaire de l'utilitaire Oracle SqlLoader.

### 5.2.2 Modifications apportées au processus de traitement des données

Les principales modifications que j'ai apportées se trouvent au niveau du programme en C, plusieurs données posant problèmes car ayant été codées en dur (le nombre de congrès par exemple). J'ai donc modifié ces données en les remplaçant par une variable.

J'ai également modifié la mise en forme des fichiers destinés à être insérés. En effet, n'ayant pas SqlLoader sur mon ordinateur, j'ai dû passer pour l'insertion des données par l'utilitaire d'importation des données de SqlDeveloper, utilitaire prenant en entrée des fichiers CSV. J'ai donc, simplement supprimé les «;» présents en fin de ligne afin de correspondre au format et j'ai changé l'extension des fichiers générés. Afin de permettre que ce nouveau format de fichier soit toujours importable par SqlLoader, j'ai modifié les fichiers de configurations de celui-ci.

### 5.2.3 Gestion des «synonymes»

Un des problèmes de la méthode d'insertion des données de l'application était qu'il était très difficile de repérer une faute de frappe avant l'insertion des données dans la base. En effet, celles-ci étant dispersées dans de nombreux fichiers, il aurait été fastidieux de devoir tous les contrôler à la recherche d'éventuelles erreurs. Or il existait de nombreux cas de noms de villes ou de personnes «synonymes», tels que «Sophia Antipolis» et «Sophia-Antipolis» ou «Cauvet Corine» et «Cauvet Corinne».

Pour résoudre ce problème, j'ai décidé de tout d'abord créer deux tables – une pour les noms

de ville et une pour les personnes – contenant les couples potentiels de synonymes, trouvés grâce à la fonction SQL soundex. L'utilisateur pouvait ainsi consulter ces couples, et supprimer les faux-positifs.

Après cela j'ai créé une procédure PL/SQL ayant pour but de «fusionner» les deux entités du couple, en modifiant les clés étrangères dans les tables les référençant et en supprimant un des deux noms. J'ai choisi de conserver le premier nom du couple, ce qui obligeait l'utilisateur à éventuellement modifier l'orthographe de celui-ci si ce n'était pas celle qui lui convenait.

La base de données était ainsi plus cohérente et la procédure facilement réutilisable lors de la prochaine insertion de données – à noter cependant, les fichiers csv initiaux n'étaient pas modifiés par cette procédure.

#### 5.2.4 Gestion des pays

Les pays correspondant aux villes présentes dans la base devaient avant être précisés manuellement dans le fichier Result/ville.csv, et la méthode de récupération des pays insérés précédemment présente dans le programme en C déclenchait souvent des bogues.

Afin d'automatiser au maximum la procédure j'ai tout d'abord créé la table inforsid\_base\_pays destinée à contenir l'identifiant d'une ville, son nom et son pays. J'y ai inséré toutes les villes françaises – récupérées sur SQL.sh<sup>3</sup>.

J'ai ensuite créé une vue inforsid\_villesSansPays recensant les villes présentes dans la base mais non présentes dans la table évoquée précédemment. Dans cette vue, l'utilisateur avait deux possibilités :

- si des villes françaises étaient dans cette vue, c'était que leur orthographe n'était pas correcte. Il pouvait donc les chercher dans la table inforsid\_base\_pays afin de trouver l'orthographe correcte et les modifier ensuite dans inforsid\_ville,
- pour les villes étrangères, il modifiait leurs pays et un déclencheur insérait alors ces nouvelles villes dans inforsid\_base\_pays.

Une fois inforsid\_villesSansPays vide, il lançait la procédure gererPays qui mettait à jour les tuples de la table inforsid\_ville.

L'avantage de cette méthode est que la base de villes présentes – et donc automatiquement gérées – augmente d'année en année au fur et à mesure des insertions de données.

#### Insertion des données des congrès Inforsid 2012 et 2013

Après avoir inséré les données correspondant au congrès de 2012 dans Dataas/dataCongres.txt j'ai créé les fichiers membre.txt et article.txt et les ai complétés à l'aide du site officiel de l'édition de DBLP.

Pour certains articles j'ai été confrontée au problème d'absence de déclaration de ville de certains auteurs. Pour ces cas, qui étaient minoritaires, j'ai cherché sur Internet la localisation des chercheurs afin d'avoir la base la plus exacte possible (il vaut mieux une ville de rattachement du chercheur, même temporairement incorrecte, que pas de ville de rattachement du tout).

### 5.3 Valorisation du congrès

#### 5.3.1 Pôles principaux

Afin de montrer les pôles principaux d'Inforsid, j'ai calculé le poids de chaque ville impliquée dans le congrès. Ce calcul reposait sur les articles présentés lors des éditions du congrès et est calculé comme suit :

- chaque article avait un poids de 1,
- ce poids était attribué à la ville de rattachement que l'auteur avait indiqué dans son article,

---

<sup>3</sup><http://sql.sh/736-base-donnees-villes-francaises>

- si l'article avait été rédigé par plusieurs auteurs, alors ce poids était réparti équitablement entre tous les auteurs (et donc entre leurs villes de rattachement).

Ces poids étaient stockés dans la table `poids_ville` contenant pour chaque ville son nom, son pays et son poids. J'ai également créé à partir des données de cette table la vue `poids_pays` contenant pour chaque pays son nom et son poids.

J'ai tout d'abord souhaité proposer une visualisation nationale d'Inforsid. Pour cela j'ai pris les 20 villes françaises ayant les poids les plus importants et je les ai représentées sur une carte en utilisant des ronds dont la taille était proportionnelle au poids. Le résultat est visible sur la figure 5.2.

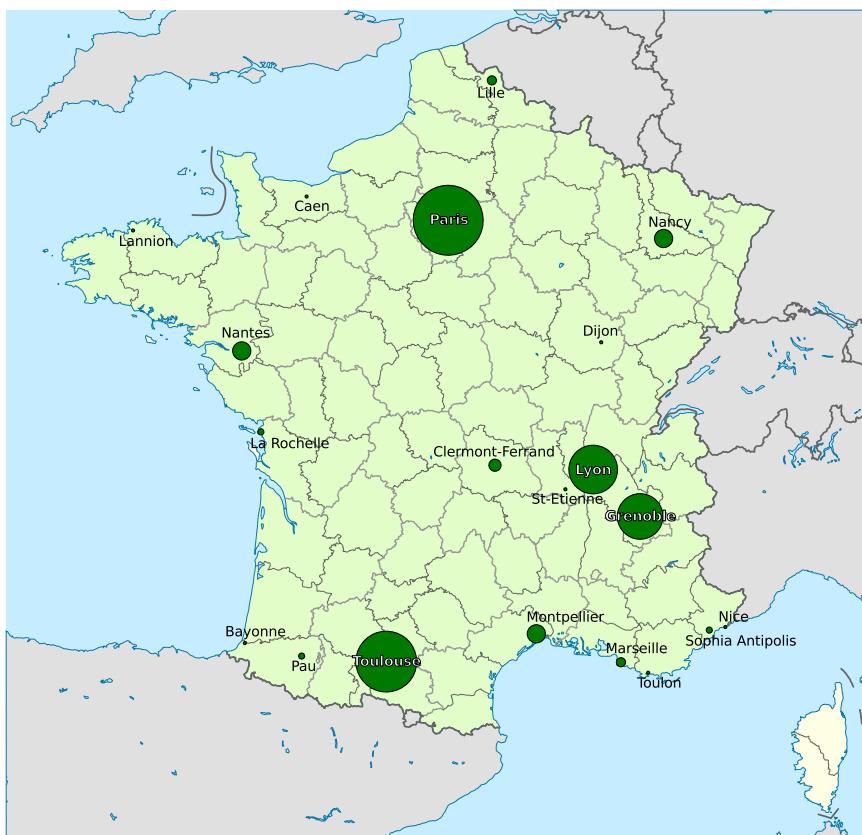


FIGURE 5.2: Les 20 villes françaises les plus présentes dans Inforsid

Inforsid étant un congrès réunissant néanmoins des chercheurs de nombreuses nationalités différentes, j'ai ensuite souhaité le représenter à l'échelle internationale. Pour cela j'ai utilisé les poids des pays qui m'ont permis de définir une «échelle de teinte» : plus le pays était vert plus il était impliqué dans Inforsid. J'ai commencé à travailler sur une mappemonde, mais la majorité des pays participants étant en Europe la lisibilité était bien trop mauvaise. Guillaume Cabanac a alors suggéré que j'utilise à la place une carte de l'Europe et que je fasse figurer les pays n'y apparaissant pas sous forme de ronds de tailles différentes à côté (en utilisant la même méthode que celle utilisée pour représenter les villes). Le résultat est présenté en figure 5.3.

### 5.3.2 Thèmes du congrès

Afin de valoriser d'Inforsid il a été décidé de représenter l'évolution des thèmes traités par le congrès. Pour illustrer cela je devais réaliser des nuages de mots à l'aide de Wordle<sup>4</sup>.

<sup>4</sup><http://www.wordle.net/>

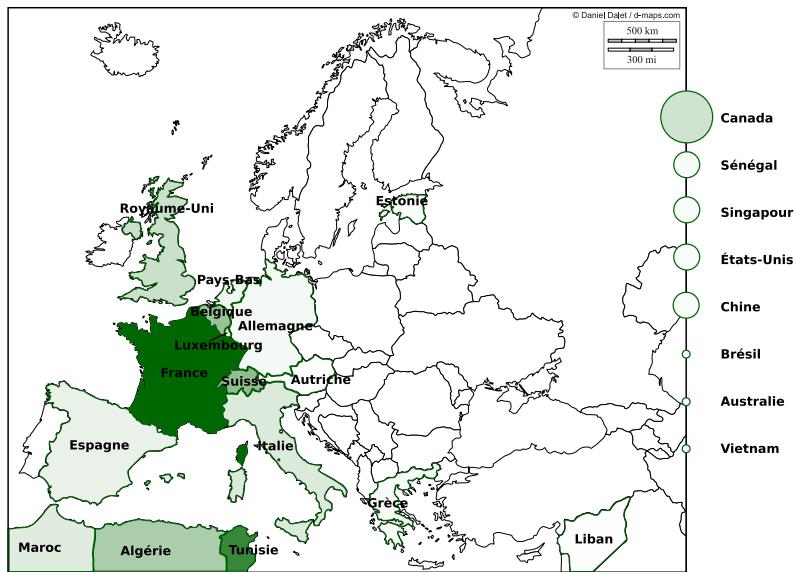


FIGURE 5.3: Les pays participant à Inforsid

Pour évaluer les thèmes abordés durant les congrès Inforsid j'ai pris comme source d'information les termes employés dans les titres d'articles présentés. J'ai donc créé une procédure PL/SQL enlevant les caractères spéciaux des titres (suppression de la ponctuation et des accents), les «s» finaux des mots – afin que certains n'apparaissaient pas en double, au singulier et au pluriel – et insérant chaque terme dans la table Mots avec l'année durant laquelle l'article a été présenté.

J'ai ensuite créé trois vues associant à chaque terme son nombre d'occurrence selon une période donnée :

- la première vue comptait les occurrences des mots des éditions du congrès jusqu'en 1993,
- la seconde vue comptait les occurrences des mots de 1994 à 2003,
- la troisième vue comptait les occurrences des mots de 2004 à 2013.

Ces vues étaient limités aux 50 termes les plus fréquents afin de ne pas surcharger les nuages de mots et ainsi faciliter leur lecture. Pour réaliser cette sélection j'ai supprimé de la table Mots les mots vides – français ou anglais, Inforsid publiant parfois des articles dans cette langue – car ils représentaient les termes les plus fréquents. Ce traitement manuel était nécessaire car Wordle peut enlever automatiquement les mots vides lors de la création d'un nuage de mot, mais d'une langue seulement.

Il m'a ensuite suffit d'exporter ces données sous la forme de trois fichiers texte ayant la structure suivante : terme : nombreOccurrences. La suppression des «s» finaux a imposé une relecture obligatoire de ces fichiers texte afin de vérifier qu'il n'y avait pas de terme pour lesquels le «s» final était nécessaire (le seul mot dans notre cas a été «processus»). J'ai ainsi obtenu les nuages de mots présentés en figures 5.4, 5.5 et 5.6.

Cependant ces nuages de mots avaient une faiblesse : ils ne prenaient pas en compte les expressions de deux ou trois mots. Pour pallier ce problème j'ai créé une procédure PL/SQL détectant toutes ces expressions et les stockant dans une table avec leur année et leur nombre d'occurrences. Afin de faciliter la suppression des expressions non pertinentes cette procédure supprime celles commençant ou terminant avec des mots vides.

Il a ensuite fallu que je fasse un filtrage manuel dans la table afin de ne garder que les expressions pertinentes pour nous, puis que j'uniformise les poids des mots en fonction des expressions conservées. En effet les poids des termes présents dans ces expressions devaient être

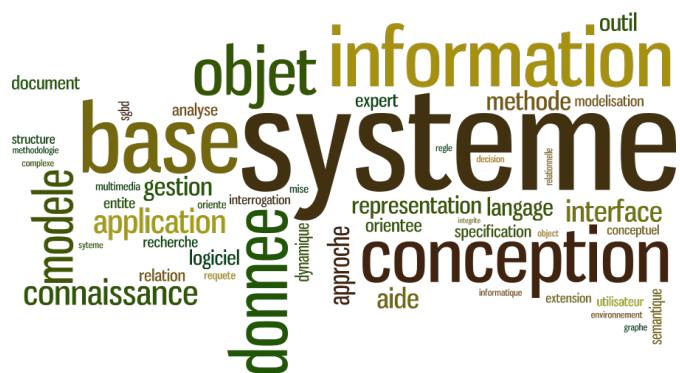


FIGURE 5.4: Termes représentant Inforsid de 1983 à 1993

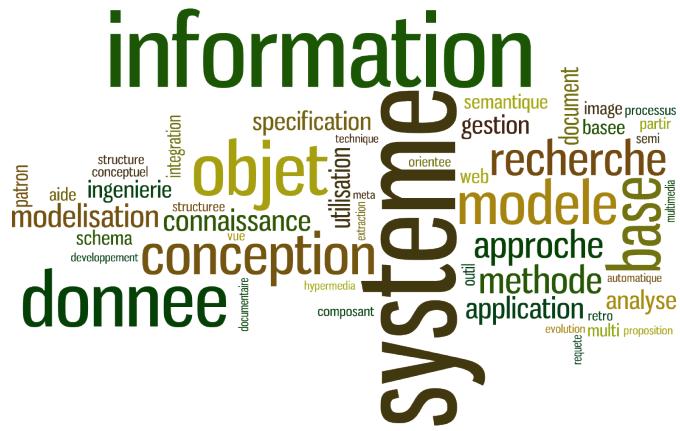


FIGURE 5.5: Termes représentant Inforsid de 1994 à 2003



FIGURE 5.6: Termes représentant Inforsid de 2004 à 2013

diminués afin de ne pas fausser les résultats – par exemple «donnee» étant présent dans «base de donnee» il faut ôter de son poids le poids de l'expression.

Ceci étant fait j'ai modifié les vues présentant les termes les plus présents pour les périodes 1983-1993, 1994-2003 et 2004-2013 afin de prendre en compte les expressions trouvées précédemment. Enfin j'ai rajouté manuellement les «s» finaux lorsqu'il y avait besoin, les accents et les apostrophes, et ainsi obtenu les wordles présents dans les figures 5.7, 5.8 et 5.9.

### 5.3.3 Influence des membres

Je devais chercher si des chercheurs participant à Inforsid étaient membres de comités de rédaction de journaux scientifiques du domaine IS. Au moment où j'ai eu à réaliser cette tâche je disposais de la base cabanac\_db1p2013, contenant notamment les comités de rédaction des 77 journaux traités dans l'article (Cabanac, 2012) (pour plus d'informations, voir la section 5.2.1). J'ai donc fait l'intersection entre les membres d'Inforsid et les chercheurs présents dans DBLP membres de comités de rédaction. Les chercheurs obtenus sont présenté dans le tableau 5.2.

J'ai ensuite fait la correspondance entre les noms des 20 journaux présents et ceux figurant sur la carte des 77 principaux journaux scientifiques figurant dans (Cabanac, 2012). Les correspondances sont présentées dans le tableau 5.4.

J'ai enfin mis en valeur sur la carte les journaux trouvés. La carte finale est visible en figure 5.10. On constate que la majorité des journaux mis en valeur sont dans la partie supérieure gauche de la carte, impliquant une similarité des thèmes traités.

### 5.3.4 Villes ayant accueilli le congrès

Les retours des utilisateurs sur notre site web présentant Inforsid ont montré qu'il pourrait être pertinent de réaliser une carte montrant où les congrès avaient eu lieu, afin de montrer le fait qu'il s'agissait bien d'un congrès francophone et pas simplement d'un congrès toujours hébergé par la ou les même villes.

Afin de réaliser cette carte j'ai repris le fond de carte utilisé pour représenter les villes les plus présentes dans le congrès, et plutôt que d'indiquer le nom de la ville dans le cercle j'y ai indiqué la (les) année(s) pour laquelle la ville avait accueilli le congrès. J'obtiens ainsi la carte présentée dans la figure 5.11, que j'ai incluse sur la page d'accueil de l'application web.

## 5.4 Modification du site web

### 5.4.1 Optimisation des procédures PL/SQL

L'application web était constituée de 6 procédures :

- Inforsid\_Accueil affichait la page d'accueil permettant d'accéder aux résumés des différentes éditions du congrès, de rechercher un chercheur ou d'accéder aux suggestions pour le comité de programme.
- Inforsid\_Fiche\_Annee affichait le résumé d'une édition.
- Inforsid\_Statut\_Personne affichait le résumé du chercheur.
- Inforsid\_Traitement recherchait un chercheur dans la base et affichait la liste des personnes trouvées ou directement la fiche du chercheur s'il n'y avait qu'un seul résultat.
- Inforsid\_Suggestions\_CP affichait la liste des membres proposés par le système pour la constitution du comité de programme
- Inforsid\_Footer affichait le pied-de-page X/HTML.

J'ai pour ma part réorganisé le code de chaque procédure afin d'avoir le plus possible la structure suivante : récupération puis affichage des données. De plus, afin de diminuer les répétitions de code, j'ai créé 2 nouvelles procédures :

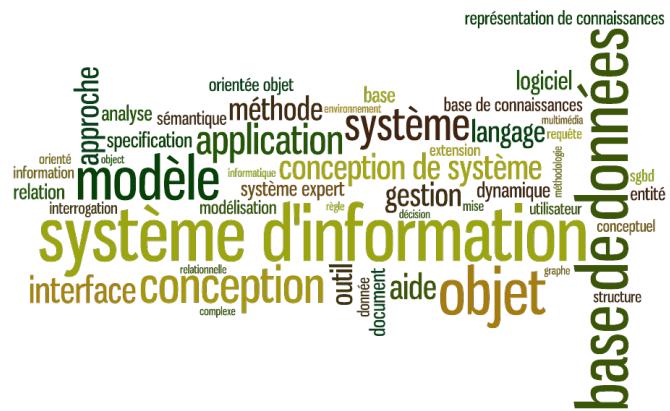


FIGURE 5.7: Termes représentant Inforsid de 1983 à 1993 en prenant en compte les expressions



FIGURE 5.8: Termes représentant Inforsid de 1994 à 2003 en prenant en compte les expressions



FIGURE 5.9: Termes représentant Inforsid de 2004 à 2013 en prenant en compte les expressions

TABLEAU 5.2: Membres d'Inforsid membres de comités de rédaction de journaux scientifiques du domaine IS

Nom du journal	Auteur
Acta Inf.	Elisa BERTINO
Data Knowl. Eng.	Georges GARDARIN
Data Knowl. Eng.	Jacky AKOKA
Data Knowl. Eng.	Colette ROLLAND
Data Knowl. Eng.	Elisa BERTINO
Data Knowl. Eng.	Stefano SPACCAPIETRA
Distributed and Parallel Databases	Elisa BERTINO
Distributed and Parallel Databases	Patrick VALDURIEZ
EJIS	Frantz ROWE
GeoInformatica	Robert LAURINI
GeoInformatica	Michel SCHOLL
IEEE Security & Privacy	Elisa BERTINO
IEEE Trans. Knowl. Data Eng.	Elisa BERTINO
Inf. Process. Manage.	Iadh OUNIS
Inf. Retr.	Josiane MOTHE
Inf. Retr.	Jacques SAVOY
Inf. Syst.	Alain PIROTTE
Information & Management	Imed BOUGHZALA
Information & Management	Moez LIMAYEM
Information & Software Technology	Colette ROLLAND
Int. J. Cooperative Inf. Syst.	Boualem BENATALLAH
Int. J. Cooperative Inf. Syst.	Elisa ERTINO
Int. J. Cooperative Inf. Syst.	Barbara PERNICI
International Journal of Geographical Information Science	Christophe CLARAMUNT
J. Intell. Inf. Syst.	Olivier PIVERT
J. Intell. Inf. Syst.	Elisa BERTINO
J. of Management Information Systems	Jacky AKOKA
Multimedia Tools Appl.	Harald KOSCH
Multimedia Tools Appl.	Chabane DJERABA
Requir. Eng.	Klaus POHL
Requir. Eng.	Oscar PASTOR
Requir. Eng.	Eric DUBOIS
Requir. Eng.	John MYLOPOULOS
Requir. Eng.	Emmanuel LETIER
Requir. Eng.	Neil A. M. MAIDEN
Requir. Eng.	Colette ROLLAND
TWEB	Elisa BERTINO
World Wide Web	Patrick VALDURIEZ

TABLEAU 5.4: Correspondance des noms de journaux stockés dans DBLP et figurant dans l'article de Guillaume Cabanac

Nom du journal dans DBLP	Nom du journal dans l'article de G. Cabanac
Acta Inf.	Acta Inform
Data Knowl. Eng.	Data Knowl Eng
Distributed and Parallel Databases	Distrib Parallel Dat
EJIS	Eur J Inform Syst
GeoInformatica	GeoInformatica
IEEE Security & Privacy	IEEE Secur Priv
IEEE Trans. Knowl. Data Eng.	IEEE T Knowl Data En
Inf. Process. Manage.	Inform Process Manag
Inf. Retr.	InformRetrieval
Inf. Syst.	InformSyst
Information & Management	InformManage-Amster
Information & Software Technology	Inform Software Tech
Int. J. Cooperative Inf. Syst.	Int J Coop Inf Syst
International Journal of Geographical Information Science	Int J Geogr Inf Sci
J. Intell. Inf. Syst.	J Intell Inf Syst
J. of Management Information Systems	J Manage Inform Syst
Multimedia Tools Appl.	Multimedia Tools Appl
Requir. Eng.	Requir Eng
TWEB	ACM T Web
World Wide Web	World Wide Web

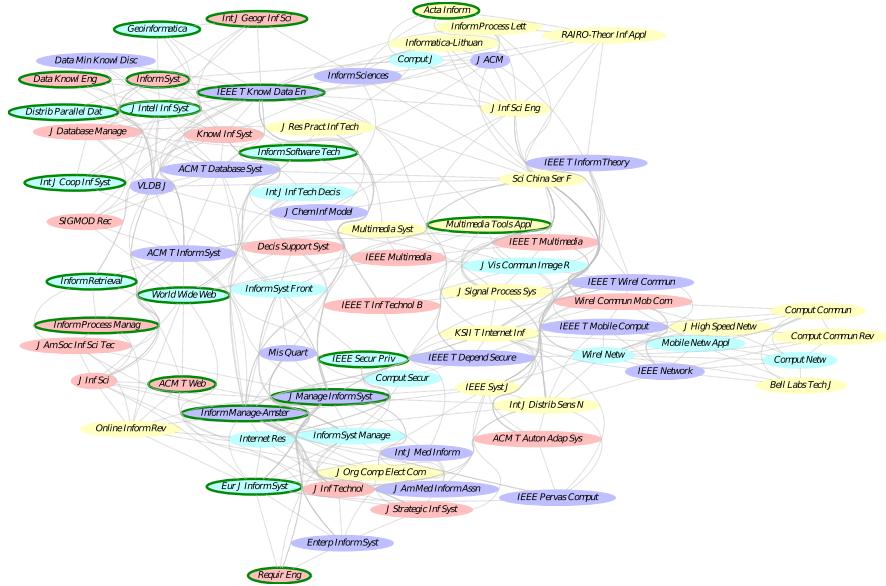


FIGURE 5.10: Cartes des 77 principaux journaux du domaine IS. Les journaux ayant des comités de rédaction auxquels participent des membres d'Inforsid sont repérés par un liséré vert

- Inforsid\_Header affichant l'en-tête X/HTML des pages avec les titres passés en paramètre (un paramètre pour le titre de la page affichée par le navigateur et un pour le titre principal <h1> affiché dans la page).
- Inforsid\_Retour\_Accueil affichant un lien permettant de retourner à l'accueil.

À la demande de Guillaume Cabanac j'ai également supprimé la mention de la personne ayant présenté le plus d'articles présente sur la page d'accueil et inversé l'ordre de présentation des participations au comité de programme et de sa localisation (présenté précédemment dans l'ordre chronologique et maintenant du plus récent au plus ancien). J'ai également créé un paquetage nommé inforsid\_conf contenant la procédure getBd qui renvoie le nom de la base de données, et modifié les pages HTML générées en conséquence. Lors d'une migration, le seul élément à modifier à présent est donc ce paquetage.

#### 5.4.2 Création d'un nouveau style graphique

Afin d'améliorer le site et de pouvoir utiliser les fonctionnalités HTML les plus récentes j'ai choisi de passer le site web en HTML5. Il a fallu pour cela que je modifie de nombreux éléments de style qui n'étaient plus supportés par la dernière version de HTML, tels que les balises align ou font. Cette mise à niveau m'a permise d'utiliser Bootstrap<sup>5</sup> afin de modifier le graphisme du site. Vous pouvez voir l'ancien graphisme sur la figure 5.12 et le nouveau sur la figure 5.13.

J'ai inséré sur la page d'accueil les graphiques réalisés (cartes et nuages de mots présentés tout au long de cette section) afin de permettre au visiteur de mieux connaître le congrès.

J'ai également ajouté des liens vers les actes Inforsid<sup>6</sup>. Il aurait été préférable de faire un lien vers chaque article mais malheureusement nous n'avons pas trouvé de moyen pour réalisé ceci de manière automatique.

<sup>5</sup>Bibliothèque CSS et Javascript éditée par Twitter.

<sup>6</sup>Présents à l'adresse <https://liris.cnrs.fr/inforsid/?q=Actes\%20Inforsid>

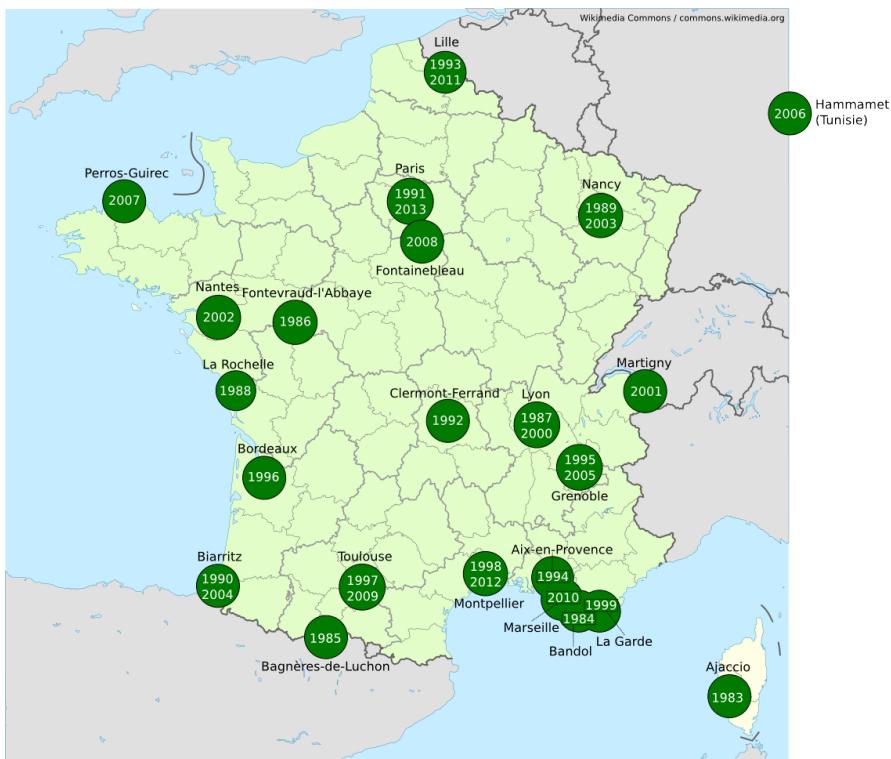


FIGURE 5.11: Villes ayant accueilli le congrès Inforsid

### Congrès Inforsid

**Les éditions du congrès**

1983	1984	1985	1986	1987	1988	1989
1990	1991	1992	1993	1994	1995	1996
2000	2001	2002	2003	2004	2005	2006
2010	2011					

**Les chercheurs**

Rechercher

**Information**

Vous trouverez ci-dessous la liste de tous les congrès Inforsid depuis 1983.  
Le comité de programme et la liste des articles sont présents pour chaque année.  
Ce congrès a rassemblé au cours de ces années **1198 membres** !  
De plus, ce n'est pas moins de **715 articles** qui ont été publiés.  
L'auteur qui a écrit le plus d'articles est **Dominique RIEU** avec ses **22 articles** !

Le site officiel de l'association est <http://inforsid.irit.fr>  
(attention : expérimental) [Suggestions de chercheurs](#) pour constituer le CP.

FIGURE 5.12: Ancien aspect du site web

**Anthologie des congrès Inforsid**

INFormatique des ORganisations et Systèmes d'Information et de Décision

Ce congrès a rassemblé sur 31 ans 1375 chercheurs qui ont présenté 805 articles.

**Éditions du congrès**

1983 1984 1985 1986 1987 1988 1989

1990 1991 1992 1993 1994 1995 1996 1997 1998 1999

2000 2001 2002 2003 2004 2005 2006 2007 2008 2009

2010 2011 2012 2013

**Chercheurs**

exemple : Flory

**Suggestions de chercheurs**

**Où le congrès Inforsid a-t-il eu lieu ?**

The map shows the European continent with a focus on Northern France. A green circle highlights the region around Lille, which is labeled with the years 1993 and 2011. The source of the map is cited as "Wikimedia Commons / commons.wikimedia.org".

FIGURE 5.13: Nouvelle apparence du site web

## 6 — Étude de genre

### 6.1 Présentation du contexte

Les femmes sont depuis toujours sous-représentées dans le domaine informatique, cependant au fur et à mesure de l'évolution de la société, ces disparités s'amenuisent.

Cette étude avait pour but d'étudier le statut actuel des chercheuses dans le milieu informatique par des méthodes scientométriques en se concentrant sur les membres de comités de rédaction de 77 revues scientifiques du domaine IS.

Les résultats du stage alimenteront la discussion en cours sur la représentation des femmes dans les sciences.

### 6.2 Récupération et mise en forme des données

Les données sur lesquelles nous devions travailler étaient celles de DBLP. Or la base contenant ces données datait de 2010 – ayant été créée dans le cadre des précédentes recherches de Guillaume Cabanac – et n'avait pas été mise à jour depuis. Il a donc été décidé que le plus simple était que je crée une nouvelle base contenant les données à jour, et ayant la même architecture que l'ancienne (voir figure 6.1).

Pour cela il a tout d'abord fallu que je récupère le fichier XML de ces données et que je le traite avec l'analyseur Java conçu par Anaïs Lefevre. J'ai constaté que j'avais un certain nombre d'exceptions lancées lors de son traitement mais le message étant «Exception muselée par Anaïs», j'en ai déduit que celles-ci ne gênaient pas l'insertion des données.

Toute la procédure d'insertion était claire et documentée, ce qui m'a permis de savoir exactement comment procéder. Ainsi après avoir créé les tables nécessaires et inséré les données à l'aide de Sql\*Loader j'ai tenté de réactiver les indexées et les contraintes (désactivées à cause de l'insertion des données en mode direct). Un problème s'est posé au niveau des contraintes car certaines clés étrangères n'étaient pas présentes dans la table mère. Étant donné qu'elles étaient assez peu nombreuses j'ai supprimé les tuples les contenant afin de pouvoir réactiver toutes les contraintes. Je suppose que ce problème est la conséquence des exceptions lancées précédemment.

Il a ensuite fallu que je crée plusieurs indexées, et là encore un problème s'est posé avec des termes présents en double alors qu'ils étaient censés être uniques. En regardant les termes incriminés j'ai constaté qu'il s'agissait certainement d'un problème d'encodage lors de la lecture

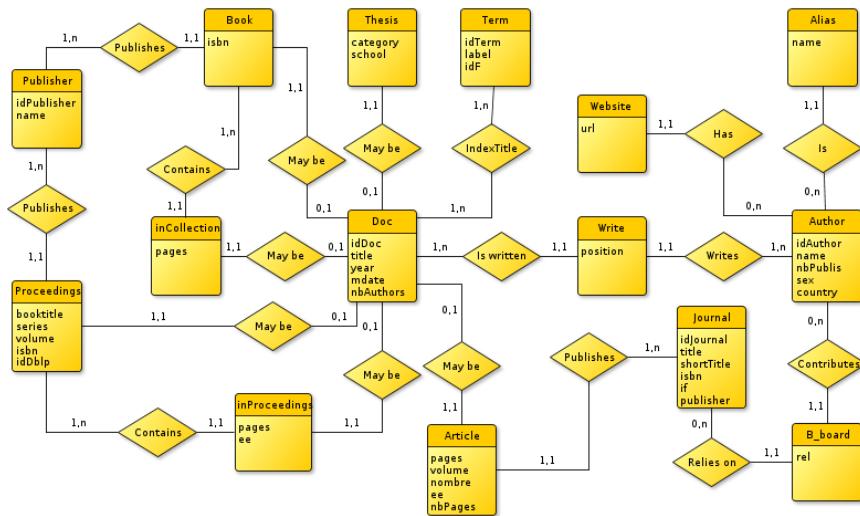


FIGURE 6.1: MCD de la base contenant les données de DBLP

du fichier XML car ils étaient tous constitués presque exclusivement du caractère «?». Étant donné qu'ils s'agissait de termes complètement inexploitables je les ai supprimés, en supprimant d'abord les tuples les concernant dans `indexTitle`, et ai créé l'index sur les termes restants.

J'ai ensuite continué la procédure en gérant les alias et mettant à jour les colonnes dérivées. J'ai finalement inséré les comités de rédaction des différents journaux et mis à jour le sexe et pays de chaque chercheur appartenant à un comité de rédaction.

## 6.3 Présentation des données et des tests utilisés

### 6.3.1 Présentation générale des données

DBLP regroupe une quantité importante de données (voir tableau 6.1), mais notre étude s'est concentrée sur les membres de comités de rédaction de 77 journaux scientifiques du domaine IS.

Les journaux dont nous connaissons les comités de rédaction sont présentés dans le tableau 6.2 et les effectifs sur lesquels nous avons travaillé dans le tableau 6.3.

TABLEAU 6.1: Nombre d'éléments référencés dans notre base de données.

Chercheur-se-s	1 265 195
Journaux scientifiques	1 346
Documents	2 265 005
dont	
Articles	957 452
Thèses	6 927
Livres	9 797
Extraits de livres	21 932
Conférences	20 080
Participations à des conférences	1 247 430

TABLEAU 6.2: Journaux du domaine IS pour lesquels nous disposons du comité de rédaction.

ACM Trans. Database Syst.	International Journal of Information
ACM Trans. Inf. Syst.	Technology and Decision Making
Acta Inf.	Internet Research
Bell Labs Technical Journal	IS Management
Comput. J.	ITA
Computer Communication Review	J. ACM
Computer Communications	J. Database Manag.
Computer Networks	J. High Speed Networks
Computers & Security	J. Inf. Sci. Eng.
Data Knowl. Eng.	J. Information Science
Data Min. Knowl. Discov.	J. Intell. Inf. Syst.
Decision Support Systems	J. of Management Information Systems
Distributed and Parallel Databases	J. Org. Computing and E. Commerce
EJIS	J. Strategic Inf. Sys.
Enterprise IS	J. Visual Communication and Image
GeoInformatica	Representation
I. J. Medical Informatics	JAMIA
IEEE MultiMedia	JASIST
IEEE Network	JIT
IEEE Pervasive Computing	Journal of Chemical Information and
IEEE Security & Privacy	Modeling
IEEE Systems Journal	Journal of Research and Practice in In-
IEEE Trans. Dependable Sec. Comput.	formation Technology
IEEE Trans. Knowl. Data Eng.	Knowl. Inf. Syst.
IEEE Trans. Mob. Comput.	MIS Quarterly
IEEE Transactions on Information	MONET
Technology in Biomedicine	Multimedia Syst.
IEEE Transactions on Information	Multimedia Tools Appl.
Theory	Online Information Review
IEEE Transactions on Multimedia	Requir. Eng.
IEEE Transactions on Wireless Com-	Science in China Series F: Information
munications	Sciences
IJDSN	SIGMOD Record
Inf. Process. Lett.	Signal Processing Systems
Inf. Process. Manage.	TAAS
Inf. Retr.	TIIS
Inf. Sci.	TWEB
Inf. Syst.	VLDB J.
Informatica, Lith. Acad. Sci.	Wireless Communications and Mobile
Information & Management	Computing
Information & Software Technology	Wireless Networks
Information Systems Frontiers	World Wide Web
Int. J. Cooperative Inf. Syst.	
International Journal of Geographical	
Information Science	

TABLEAU 6.3: Nombre de membres de comités de rédaction référencés dans notre base de données.

Femmes	422
Hommes	2402
Sexe non déterminé	26
Total	2850

### 6.3.2 Présentation des tests statistiques

#### Shapiro-Wilk

Ce test a pour hypothèse nulle que l'échantillon testé est issu d'une population normalement distribuée et la statistique de test utilisée est

$$W = \frac{(\sum_{i=1}^n a_i x_{(i)})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

où

- $x_{(i)}$  désigne la ième statistique d'ordre , i.e., le ième plus petit nombre dans l'échantillon,
- $\bar{x}$  est la moyenne de l'échantillon,
- la constante  $a_i$  est donnée par

$$(a_1, \dots, a_n) = \frac{m^T V^{-1}}{\sqrt{m^T V^{-1} V^{-1} m}}$$

où  $m = (m_1, \dots, m_n)^T$  et  $m_1, \dots, m_n$  sont les espérances des statistiques d'ordre d'un échantillon de variables indépendantes et identiquement distribuée suivant une loi normale, et  $V$  est la matrice de variance-covariance de ces statistiques d'ordre.

#### Kolmogorov-Smirnov

Le test de Kolmogorov-Smirnov est utilisé pour déterminer si un échantillon suit bien une loi donnée connue ou bien si deux échantillons suivent la même loi. Sa statistique vaut

$$D = \max(|F(x) - F(y)|)$$

où

- $F(x)$  est la fonction de répartition du premier échantillon  $(x_1, \dots, x_p)$ ,
- $F(y)$  est la fonction de répartition du second échantillon  $(y_1, \dots, y_q)$ .

#### Wilcoxon / Mann-Whitney U

Le test de Wilcoxon – aussi appelé test  $U$  de Mann-Whitney – permet de tester si deux échantillons ont la même loi. Sa statistique de test est

$$W = \sum_{i=1}^p R_i$$

où  $R_i$  est le rang de  $x_i$  dans l'ensemble des  $(x_1, \dots, x_p)$  et des  $(y_1, \dots, y_q)$  classés par ordre croissant.

## 6.4 Comparaison de générations

En me documentant sur les différentes études de genre déjà réalisées j'ai pris connaissance de l'article (Arensbergen, van der Weijden, & Besselaar, 2012) comparant la différence de

productivité entre hommes et femmes entre deux générations de chercheurs. Arensbergen, van der Weijden, & Besselaar (2012) constatent dans cet article que cette différence – clairement établie dans la génération de chercheurs établis pour laquelle les hommes produisent plus que les femmes – a tendance à disparaître voire à s'inverser dans le domaine des sciences sociales.

J'ai pensé qu'il pourrait être intéressant d'analyser nos données de la même façon afin de voir s'il y avait une évolution, positive ou négative, de la place des femmes dans la communauté IS.

Afin de déterminer les limites des générations de chercheurs à comparer j'ai choisi de me baser sur la date du premier document – article publié, livre ou participation à une conférence – archivé sur DBLP.

À partir de ces dates j'ai divisé les chercheurs comme suit :

- une «ancienne» génération de chercheurs ayant réalisé leur premier document avant 2000,
- une «nouvelle» génération de chercheurs ayant réalisé leur premier document après ou en 2000.

Les 2 850 membres de comités de rédaction étaient donc répartis selon les effectifs présentés dans le tableau 6.4 – le total ne vaut pas 2 850 car je n'ai pris en compte que les chercheurs pour lesquels le sexe était clairement déterminé.

TABLEAU 6.4: Répartition des membres de comités de rédaction – les chercheurs pour lesquels le genre n'a pas pu être déterminé ne sont pas pris en compte ici.

	Femmes	Hommes	Total
Ancienne génération	289	1882	2171
Nouvelle génération	133	520	653
Total	422	2402	2824

### Productivité

Le premier élément que j'ai souhaité tester a été la productivité des chercheurs. En effet il a souvent été noté une différence significative d'articles produits entre hommes et femmes (Nakhaie, 2002; Prpić, 2002; Penas & Willett, 2006; Abramo, D'Angelo, & Caprasecca, 2009; Symonds, Gemmell, Braisher, Gorringe, & Elgar, 2006; Ledin, Bornmann, Gannon, & Wallon, 2007; Taylor, Fender, & Burke, 2006; Xie & Shauman, 1998). Je souhaitais donc voir si cette différence avait tendance à disparaître avec les années.

Pour calculer la productivité d'un chercheur j'ai décidé d'utiliser la métrique de productivité présentée dans (Egghe, Rousseau, & Van Hooydonk, 2000), et ayant pour formule :

$$g(r, n) = \frac{2^{n-r}}{2^n - 1}$$

avec  $n \in \mathbb{N}_+^*$  le nombre total d'auteurs du document et  $r \in \mathbb{N}_+^*$  le rang du chercheur dans la liste de ceux-ci.

Les productivités selon le genre de l'ancienne et de la nouvelle génération ainsi calculées sont représentées en figures 6.2 et 6.3 .

On constate que la différence clairement visible pour l'ancienne génération s'atténue fortement pour la nouvelle.

Afin d'en avoir le cœur net j'ai décidé d'effectuer des tests statistiques pour déterminer :

1. si la différence de production entre hommes et femmes était significative,
2. si cette différence diminuait avec le temps.

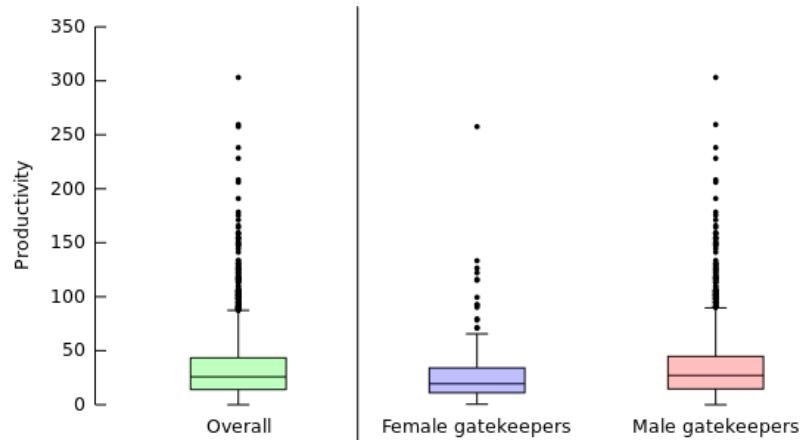


FIGURE 6.2: Productivité de l'ancienne génération de chercheurs par genre

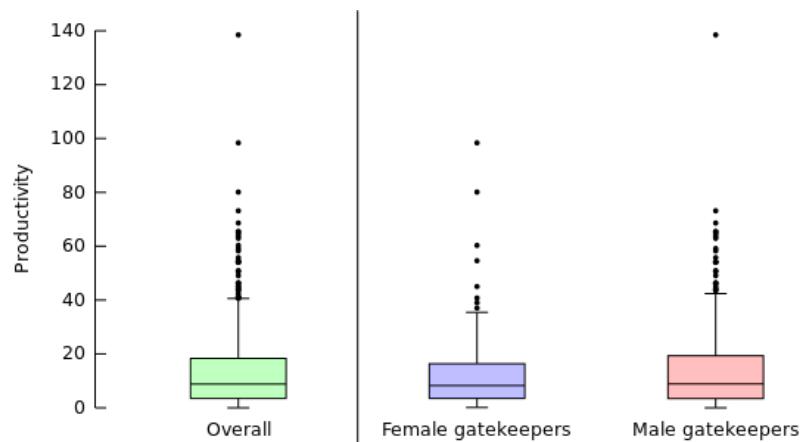


FIGURE 6.3: Productivité de la nouvelle génération de chercheurs par genre

Afin de choisir quels tests utiliser pour cela, j'ai tout d'abord choisi de vérifier la normalité de mes échantillons représentant la productivité des chercheurs et chercheuses de chaque génération avec un test de Shapiro-Wilk. Les résultats de ces tests sont présentés dans le tableau 6.5. On peut voir qu'aucun de mes échantillons n'est distribué selon la loi normale – les p-values étant largement inférieures à 5%.

TABLEAU 6.5: Résultats du test de Shapiro-Wilk sur les échantillons «Productivité des chercheuses de l'ancienne génération», «Productivité des chercheurs de l'ancienne génération», «Productivité des chercheuses de la nouvelle génération» et «Productivité des chercheurs de la nouvelle génération», indiquant si ces échantillons suivent une loi normale.

Échantillon testé		W	p-value
Ancienne génération	Femmes	0,7158	$< 2,2 \cdot 10^{-16}$
	Hommes	0,8045	$< 2,2 \cdot 10^{-16}$
Nouvelle génération	Femmes	0,708	$6,235 \cdot 10^{-15}$
	Hommes	0,7818	$< 2,2 \cdot 10^{-16}$

Étant donnée que nous ne connaissons pas la loi de ceux-ci, j'ai décidé d'utiliser des tests non paramétriques – Kolmogorov-Smirnov et Wilcoxon – pour les comparer. Les résultats de ces tests sont visibles dans le tableau 6.6. On peut voir que les deux nous permettent de tirer les mêmes conclusions :

- la différence de productivité entre membres masculins et féminins des comités de rédaction de l'ancienne génération est significative,
- cette différence s'est atténuée au point de ne plus être significative pour la nouvelle génération.

TABLEAU 6.6: Résultats des tests de Kolmogorov-Smirnov (KS) et Wilcoxon (W) indiquant si la différence de production entre hommes et femmes est significative.

Échantillon testé		Statistique de test	p-value
Ancienne génération	KS	0,1599	$5,424 \cdot 10^{-06}$
	W	221549,0	$3,783 \cdot 10^{-07}$
Nouvelle génération	KS	0,0958	0,2857
	W	32960,5	0,4043

### $\varphi$ -index

J'ai ensuite pensé qu'il serait intéressant de comparer également l'évolution du  $\varphi$ -index entre générations. Le  $\varphi$ -index mesure la *partnership ability* – l'aptitude à collaborer avec d'autres chercheurs et à conserver des collaborations. J'ai calculé cette mesure à l'aide de l'article (Cabanac, 2013). Les boîtes à moustaches correspondantes sont visibles en figures 6.4 et 6.5.

On peut voir que même si la médiane du  $\varphi$ -index est légèrement inférieure à celle des hommes dans la nouvelle génération, la différence est tout de même beaucoup moins flagrante que pour l'ancienne génération.

J'ai décidé d'appliquer la même démarche statistique que pour la production étant donné qu'ici aussi mes échantillons ne suivaient pas une loi normale (voir tableau 6.7).

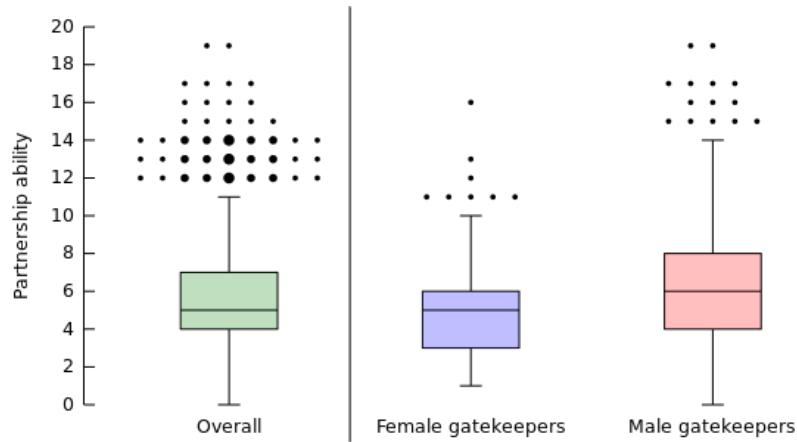


FIGURE 6.4:  $\varphi$ -index de l'ancienne génération de chercheurs par genre

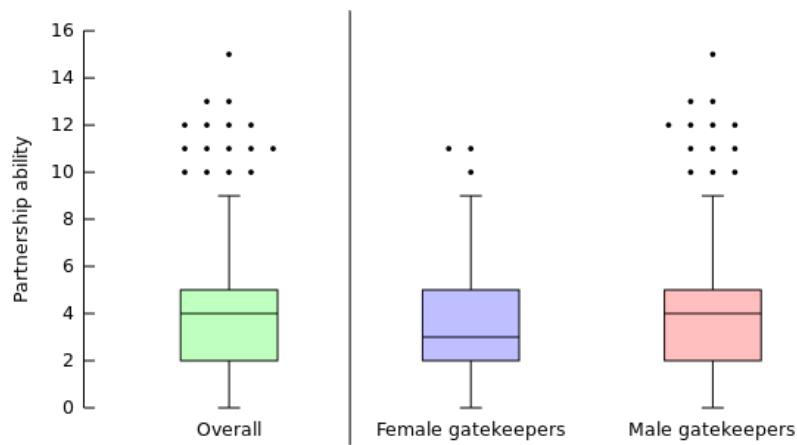


FIGURE 6.5:  $\varphi$ -index de la nouvelle génération de chercheurs par genre

TABLEAU 6.7: Résultats du test de Shapiro-Wilk sur les échantillons « $\varphi$ -index des chercheuses de l'ancienne génération», « $\varphi$ -index des chercheurs de l'ancienne génération», « $\varphi$ -index des chercheuses de la nouvelle génération» et « $\varphi$ -index des chercheurs de la nouvelle génération», indiquant si ces échantillons suivent une loi normale.

Échantillon testé		W	p-value
Ancienne génération	Femmes	0,9454	$7,026 \cdot 10^{-9}$
	Hommes	0,963	$< 2,2 \cdot 10^{-16}$
Nouvelle génération	Femmes	0,9263	$2,032 \cdot 10^{-6}$
	Hommes	0,9192	$4,576 \cdot 10^{-16}$

On constate que, comme pour la productivité, les deux tests nous donnent les même résultats (voir tableau 6.8) :

- la différence de  $\varphi$ -index entre membres masculins et féminins des comités de rédaction de l'ancienne génération est significative,
- cette différence s'est atténuée au point de ne plus être significative pour la nouvelle génération.

## 6.5 Conclusion de l'étude

TABLEAU 6.8: Résultats des tests de Kolmogorov-Smirnov (KS) et Wilcoxon (W) indiquant si la différence de  $\varphi$ -index entre hommes et femmes est significative.

Échantillon testé		Statistique de test	p-value
Ancienne génération	KS	0,1547	$1,231 \cdot 10^{-5}$
	W	225853,5	$3,783 \cdot 10^{-7}$
Nouvelle génération	KS	0,077	0,5566
	W	32007,0	0,1808



## 7 — Méthodes et outils utilisés

### 7.1 Base de données

#### 7.1.1 Oracle Database

Oracle Database est un SGBD relationnel fourni par Oracle Corporation, leader mondial des bases de données. Il s'agit d'un SGBD d'entreprise : il est puissant, capable de manipuler de grandes quantités d'informations et peut être utilisé par des milliers d'utilisateurs simultanément.

La première version d'Oracle (Oracle 4) est commercialisée en 1984 sur les machines IBM. Depuis Oracle Corporation n'a cessé de faire évoluer son produit, multipliant les plates-formes matérielles supportées (plus d'une centaine aujourd'hui) et améliorant les performances. Oracle se décline en plusieurs versions afin de mieux répondre aux besoins des entreprises.

Outre la base de données, Oracle fournit de nombreux outils formant un véritable environnement de travail, permettant notamment une administration graphique d'Oracle (les outils d'administration les plus connus sont Oracle Manager (SQL\*DBA), Network Manager, Oracle Enterprise Manager et Import/Export, un outil permettant d'échanger des données entre deux bases Oracle), de s'interfacer avec des produits divers et d'assistants de création et de configuration de bases de données.

#### 7.1.2 SqlDeveloper

SQL Developer est un outil graphique fourni gratuitement par Oracle qui simplifie le développement et l'administration des bases de données Oracle en permettant de visualiser de manière plus pratique leur contenu. Il présente les tables présentes dans la base mais également les fonctions, procédures, triggers, séquences ou autres objets présents.

SQL Developer offre une solution complète pour développer des applications PL/SQL, une feuille de travail pour lancer des requêtes et des scripts, une console d'administration pour gérer la base de données, une interface de retour, et d'autres outils que nous n'avons pas eu à utiliser lors de notre projet.

#### 7.1.3 Sql\*Loader

SQL\*Loader est un utilitaire de chargement spécifique pour les bases Oracle. Il permet d'insérer dans une ou plusieurs tables des données issues d'un fichier texte.

Il permet notamment de :

- charger des fichiers texte externes dans Oracle avec des fichiers d'entrée au format fixe ou variable (avec séparateur),
- utiliser des fonctions SQL,
- générer des clés uniques,
- optimiser le mode de chargement «direct»,
- gérer les logs et les erreurs avec possibilité de reprise.

## 7.2 Analyse et mise en forme des données

### 7.2.1 Sofa Statistics

SOFA Statistics – Statistics Open For All – est un logiciel de statistique libre mettant en avant la simplicité d'utilisation et d'apprentissage et la propreté des sorties graphiques.

Il permet de :

- faire des graphiques,
- produire des tableaux récapitulatifs,
- effectuer plusieurs tests statistiques de base.

### 7.2.2 R et RStudio

R est un langage de programmation libre et un environnement mathématique utilisés pour le traitement de données et l'analyse statistique. Il s'agit de l'un des logiciels les plus utilisés par les analystes.

RStudio est un environnement de développement multiplateforme gratuit et open source pour R, permettant de travailler avec celui-ci de manière plus confortable.

### 7.2.3 Gnuplot

Gnuplot est un logiciel libre qui sert à produire des représentations graphiques en deux ou trois dimensions de fonctions numériques ou de données. Le programme fonctionne sur de nombreux systèmes d'exploitation et peut envoyer les graphiques à l'écran ou dans des fichiers dans de nombreux formats.

Le programme peut être utilisé interactivement, et est accompagné d'une aide en ligne. L'utilisateur entre en ligne de commande des instructions qui ont pour effet de produire un tracé. Il est aussi possible d'écrire des scripts gnuplot qui, lorsqu'ils sont exécutés, génèrent les graphiques de l'utilisateur.

### 7.2.4 L<sup>A</sup>T<sub>E</sub>X

L<sup>A</sup>T<sub>E</sub>Xest un langage et un système de composition de documents créé par Leslie Lamport en 1983.

Du fait de sa relative simplicité, il est devenu la méthode privilégiée d'écriture de documents scientifiques employant TeX<sup>1</sup>. Il est particulièrement utilisé dans les domaines techniques et scientifiques pour la production de documents de taille moyenne ou importante (thèse ou livre, par exemple). Néanmoins, il peut être aussi employé pour générer des documents de types variés (par exemple, des lettres, ou des transparents).

LaTeX exige du rédacteur de se concentrer sur la structure logique de son document, son contenu, tandis que la mise en page du document (césure des mots, alinéas) est laissée au logiciel lors d'une compilation ultérieure.

---

<sup>1</sup>Système logiciel de composition de documents, largement utilisé par les scientifiques, particulièrement en mathématiques, physique, bio-informatique, astronomie et informatique.

## 7.3 Gestion de configuration

### 7.3.1 Apache Subversion

Subversion – souvent abrégé SVN – est un logiciel de gestion de versions, distribué sous licence Apache et BSD. Il fonctionne sur le mode client-serveur, avec :

- un Serveur informatique centralisé et unique où se situent :
  - les fichiers constituant la référence (le dépôt ou *repository* en anglais),
  - un logiciel serveur Subversion tournant en tâche de fond
- des postes clients sur lesquels se trouvent :
  - les fichiers recopiés depuis le serveur, éventuellement modifiés localement depuis,
  - un logiciel client permettant la synchronisation entre chaque client et le serveur de référence.

SVN facilite grandement le travail collaboratif en incluant une gestion des conflits – si deux clients ont modifiés un fichier en même temps il le détecte et laisse l'utilisateur décider des portions du fichiers à modifier sur le serveur.

Dans mon cas – étant donné que j'étais la seule à travailler sur le projet, Guillaume se contentant de consulter les documents sans les modifier – il avait principalement une fonction de sauvegarde des données.



## 8 — Assurance et contrôle qualité

### 8.1 Compte-rendus hebdomadaires

Je devais rédiger chaque semaine un compte-rendu. Guillaume Cabanac me le rendait ensuite accompagné de ses observations.

Il s'agissait à la fois d'un moyen de garder une trace de mon avancée et de permettre à mon maître de stage de savoir ce que je faisais mais également un bon exercice de rédaction afin de me familiariser avec la présentation de mon travail.

### 8.2 Réunions avec mon maître de stage

En plus des compte-rendus hebdomadaires nous avions généralement deux à trois réunions par mois afin de pouvoir discuter de vive voix des prochaines tâches à réaliser ou des éventuelles difficultés que je rencontrais.



## **9 — Bilan**

- 9.1 Bilan du projet**
- 9.2 Bilan personnel**
- 9.3 Conclusion**



## Références

- Abramo, G., D'Angelo, C. A., & Caprasecca, A. (2009, Jun). Gender differences in research productivity: A bibliometric analysis of the Italian academic system. *Scientometrics*, 79(3), 517-539. Retrieved from <http://dx.doi.org/10.1007/s11192-007-2046-8> doi: 10.1007/s11192-007-2046-8
- Arensbergen, P., van der Weijden, I., & Besselaar, P. (2012, Dec). Gender differences in scientific productivity: a persisting phenomenon? *Scientometrics*, 93(3), 857-868. Retrieved from <http://dx.doi.org/10.1007/s11192-012-0712-y> doi: 10.1007/s11192-012-0712-y
- Cabanac, G. (2012). 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. Retrieved from [http://www.irit.fr/publis/SIG/2012\\\_JASIST\\\_C.pdf](http://www.irit.fr/publis/SIG/2012\_JASIST\_C.pdf) doi: 10.1002/asi.22609
- Cabanac, G. (2013). Experimenting with the partnership ability  $\varphi$ -index on a million computer scientists. *Scientometrics*. Retrieved from [http://www.irit.fr/publis/SIG/2013\\\_Scientometrics\\\_C.pdf](http://www.irit.fr/publis/SIG/2013\_Scientometrics\_C.pdf) doi: 10.1007/s11192-012-0862-y
- Cunningham, S. J., & Dillon, S. M. (1997, May). Authorship patterns in information systems. *Scientometrics*, 39(1), 19-27. Retrieved from <http://dx.doi.org/10.1007/BF02457428> doi: 10.1007/BF02457428
- Egghe, L., Rousseau, R., & Van Hooydonk, G. (2000). Methods for accrediting publications to authors or countries: Consequences for evaluation studies. *JASIST*, 51(2), 145-157.
- Hegarty, P., & Pratto, F. (2010). *Interpreting and Communicating the Results of Gender-Related Research*. Springer-Verlag. Retrieved from [http://dx.doi.org/10.1007/978-1-4419-1465-1\\\_10](http://dx.doi.org/10.1007/978-1-4419-1465-1\_10) doi: 10.1007/978-1-4419-1465-1\\_10
- Hildrun, K., Alexander, P., & Johannes, S. (2012, Oct). Research evaluation. Part II: gender effects of evaluation: are men more productive and more cited than women? *Scientometrics*, 93(1), 17-30. Retrieved from <http://dx.doi.org/10.1007/s11192-012-0658-0> doi: 10.1007/s11192-012-0658-0
- Jump, P. (2013, 7 March). *Male domination of philosophy 'must end'*. Times Higher Education.

- Ledin, A., Bornmann, L., Gannon, F., & Wallon, G. (2007, Nov). A persistent problem. Traditional gender roles hold back female scientists. *EMBO reports*, 8(11), 982-987. Retrieved from <http://dx.doi.org/10.1038/sj.embor.7401109> doi: 10.1038/sj.embor.7401109
- Long, J. S. (1992). Measures of Sex Differences in Scientific Productivity. *Social Forces*, 71(1), pp. 159-178. Retrieved from <http://www.jstor.org/stable/2579971>
- Mauleón, E., Hillán, L., Moreno, L., Gómez, I., & Bordons, M. (2013, Apr). Assessing gender balance among journal authors and editorial board members. *Scientometrics*, 95(1), 87-114. Retrieved from <http://dx.doi.org/10.1007/s11192-012-0824-4> doi: 10.1007/s11192-012-0824-4
- Nakhaie, M. R. (2002, May). Gender Differences in Publication among University Professors in Canada\*. *Canadian Review of Sociology/Revue canadienne de sociologie*, 39(2), 151-179. Retrieved from <http://dx.doi.org/10.1111/j.1755-618X.2002.tb00615.x> doi: 10.1111/j.1755-618X.2002.tb00615.x
- Penas, C. S., & Willett, P. (2006, Jun). Brief communication: Gender differences in publication and citation counts in librarianship and information science research. *Journal of Information Science*, 32(5), 480-485. Retrieved from <http://dx.doi.org/10.1177/0165551506066058> doi: 10.1177/0165551506066058
- Prpić, K. (2002). Gender and productivity differentials in science. *Scientometrics*, 55(1), 27-58. Retrieved from <http://dx.doi.org/10.1023/A:1016046819457> doi: 10.1023/A:1016046819457
- Rossiter, M. W. (1993, May). The Matthew Matilda Effect in Science. *Social Studies of Science*, 23(2), 325-341. Retrieved from <http://dx.doi.org/10.1177/030631293023002004> doi: 10.1177/030631293023002004
- Shaw, C. (2013, 7 March). *Global action on gender inequality in HE: what's needed?* The Guardian.
- Symonds, M. R., Gemmell, N. J., Braisher, T. L., Gorringe, K. L., & Elgar, M. A. (2006, Dec). Gender Differences in Publication Output: Towards an Unbiased Metric of Research Performance. *PLoS ONE*, 1(1), e127. Retrieved from <http://dx.doi.org/10.1371/journal.pone.0000127> doi: 10.1371/journal.pone.0000127
- Taylor, S. W., Fender, B. F., & Burke, K. G. (2006, April). Unraveling the Academic Productivity of Economists: The Opportunity Costs of Teaching and Service. *Southern Economic Journal*, 72(4), 846-859. Retrieved from <http://ideas.repec.org/a/sej/ancoec/v724y2006p846-859.html>
- The Accidental Mathematician. (2013, 9 February). *Gender Bias 101 For Mathematicians*. Personal blog «ilaba».
- The Librarian Kate. (2013, 3 March). *Who Rule The World? Girls–A Look at the Scholarly Literature on Gender and Librarianship (Part 2)*. Personal blog «katekosturski».
- The Singular Scientist. (2012, 29 November). *More Gender Bias Uncovered: Conference Symposia Organizers*. Personal blog «womeninwetlands».
- Xie, Y., & Shauman, K. A. (1998). Sex Differences in Research Productivity: New Evidence about an Old Puzzle. *American Sociological Review*, 63(6), pp. 847-870. Retrieved from <http://www.jstor.org/stable/2657505>

## **Annexes**