



Université
Paul Sabatier
TOULOUSE III

Rapport de stage

Version 6

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

Ophélie FRAISIER

L3 SID

Du 8 avril au 28 juin 2013



Maître de stage :
Guillaume Cabanac

Tuteur de stage :
Sébastien Gerchinovitz



Rapport de stage

Version 6

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

Ophélie FRAISIER

L3 SID

Du 8 avril au 28 juin 2013



Maître de stage :
Guillaume Cabanac

Tuteur de stage :
Sébastien Gerchinovitz

Résumé du stage

Étant encore incertaine quant à ma future orientation professionnelle, j'ai décidé de profiter du stage destiné à valider ma troisième année de licence afin de découvrir le monde de la recherche. Pour cela j'ai contacté Guillaume Cabanac qui m'a proposé un stage en scientométrie¹ dans son laboratoire de recherche, l'IRIT².

La première partie de ce stage consistait à valoriser Inforsid, un congrès francophone réunissant chaque année depuis 1983 les chercheurs en Sciences de l'Information (SI). Pour cela je disposais d'une base de données contenant, pour chaque édition du congrès, son comité de programme³, la ville l'ayant accueilli et les articles présentés. Ce travail était destiné à me familiariser avec le vocabulaire de la recherche scientifique et à me permettre de me documenter au sujet des méthodes scientométriques. J'ai commencé par intégrer les données des éditions 2012 et 2013. Afin de mettre en évidence les pôles majeurs du congrès, j'ai ensuite réalisé deux cartes, une de France et une d'Europe. Pour visualiser les principaux thèmes du congrès et leur évolution j'ai par la suite créé trois nuages de mots (un nuage par décennie) représentant les concepts⁴ les plus fréquents dans les titres des articles présentés. J'ai ensuite réalisé une carte représentant 77 revues scientifiques du domaine SI sur laquelle étaient mises en évidence les revues comprenant des membres d'Inforsid dans leur comité de rédaction⁵. Enfin j'ai réalisé une carte représentant les différentes villes ayant successivement accueilli le congrès.

La seconde mission portait sur une étude de genre⁶ des membres de comités de rédaction des 77 revues scientifiques évoquées précédemment. Les résultats de cette étude étaient destinés à alimenter la rédaction d'un article scientifique. Je disposais des bibliographies de 2 850 membres : 422 femmes, 2 402 hommes et 26 membres dont le sexe n'avait pas pu être déterminé. J'ai calculé plusieurs indicateurs scientométriques, tels que la production (mesurant la quantité de documents publiés par le membre), l'homophilie (mesurant la propension du membre à collaborer avec des femmes) ou le φ -index (mesurant la capacité du membre à maintenir des collaborations).

Malheureusement nous avons rapidement réalisé que notre calcul de l'homophilie était incorrect à cause d'un manque d'information concernant le genre des coauteurs – et incalculable à cause de l'énorme quantité de données que nous aurions dû annoter manuellement. J'ai donc décidé de me documenter en consultant les articles scientifiques existants présentant des études de genre. Ceux-ci m'ont décidé à orienter mon étude vers une comparaison de générations : j'ai divisé l'échantillon en deux – les membres ayant publié leur premier article avant 2000 et les autres – et observé l'évolution des différences de productivité et de φ -index entre hommes et femmes. Dans les deux cas on remarque que les femmes de l'ancienne génération produisent moins que les hommes et ont plus de difficultés à collaborer ou à maintenir des collaborations avec leurs pairs. Cependant ces différences tendent à disparaître pour la nouvelle génération – pour laquelle, sur ces métriques, aucune différence significative n'est mesurable entre hommes et femmes.

Au moment où je rédige ce résumé, cette étude de genre n'est pas encore finalisée mais je pense orienter mon article vers cette différence entre générations.

Ce stage aura été une expérience profondément enrichissante pour moi. J'ai eu la chance de réaliser un travail de recherche avec un maître de stage ouvert à mes suggestions et toujours présent pour me conseiller. De plus j'ai appris à utiliser de nouveaux outils, concepts et méthodes qui pourront m'être utiles dans le futur.

1. Étude quantitative de la science par une démarche scientifique.

2. Institut de Recherche en Informatique de Toulouse.

3. Liste de chercheurs chargés de sélectionner les articles présentés au congrès.

4. Mots ou expressions de plusieurs mots – telles que « base de données » ou « recherche d'information ».

5. Liste de chercheurs chargés de sélectionner les articles à publier dans le journal.

6. Analyse des différences et des similitudes entre hommes et femmes.

Internship summary

Being uncertain about my future career, I decided to experience the world of research during the internship scheduled at the end of my undergraduate degree. I liaised with Guillaume Cabanac who offered me to do an internship in scientometrics¹ at his research laboratory, the IRIT.²

My work was comprised of two tasks. The first task was to promote Inforsid, a congress gathering researchers in Information Systems (IS) since 1983, using modern computing facilities. I used a database containing, for each edition of the congress, the program committee,³ the city having hosted and the papers presented. This work was intended to familiarise myself with the language of scientific research and allow me to learn about scientometric methods.

First, I integrated data from the 2012 and 2013 editions. Second, to highlight the congress's main contributors I produced two maps, one for France and one for Europe. Third, to stress the main themes and their evolution, I devised three word clouds (one cloud per decade) representing the most recurrent concepts⁴ in the titles of presented papers. Fourth, I produced a chart representing 77 scientific journals from the IS field. In this chart I emphasized the journals comprising members of Inforsid in their editorial board.⁵ Fifth, I made a map showing the cities where the 31 editions were held.

The second task concerned a gender study⁶ using as input the members of the editorial boards of the 77 aforementioned scientific journals. The main findings of this work were intended to appear in an academic article. There were 2,850 records : 422 women, 2,402 men and 26 members whose gender could not be determined. I calculated several indicators, such as researchers' production (measuring the amount of material produced by the member), homophily (measuring the propensity of members to collaborate with women) or φ -index (measuring the ability of the member to maintain collaborations).

Alas, we soon realized that our calculation of homophily was wrong because of a lack of data about the gender of coauthors – and incalculable because of the huge amount of data that we would have to manually annotate. I then decided to review the literature about gender studies. I read about forty papers spanning a large period (1992–2013). This reading suggested to me the idea to direct my study to a comparison of generations: I divided the aforementioned sample to define two sub-groups – members who published their first article before 2000 and the others – and studied the evolution of differences in productivity and φ -index between men and women. In both cases we find that women from the older generation produce less than men and have more difficulty working or maintaining collaborations with their peers. However these differences tend to disappear for the contemporary generation – on which, on these metrics, no significant differences were noted between men and women. Thus it seems that well-reported gender gap in academia tends to disappear.

At the time I write this summary, this gender study is not yet finalized, however I think I will focus my article on these differences between generations.

This placement has proved to be a deeply rewarding experience for me. I was given the opportunity to conduct a research study with an advisor open to my suggestions and always available to discuss my progress. Furthermore I learnt to use new tools, concepts and methods that will certainly prove to be useful to me in the future.

1. Quantitative study of science by a scientific approach.

2. Institut de Recherche en Informatique de Toulouse.

3. Panel of researchers responsible for selecting articles to be presented at the congress.

4. Words or term of several words – such as “information retrieval.”

5. Panel of researchers responsible for selecting articles for publication in the journal.

6. An analysis of the differences and similarities between women and men.

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 permis de réaliser ce stage.

En premier lieu je remercie Michel Daydé, directeur de l'IRIT, de m'avoir accepté comme stagiaire au sein de son établissement et Josiane Mothe, responsable de l'équipe SIG, de m'avoir permis 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 apportés. Je tiens également à remercier Sébastien Gerchinovitz, mon tuteur de stage, pour son aide.

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 apportés. Il a toujours été présent et attentif à mes remarques ou suggestions, et ce fut un véritable plaisir de travailler à ses côtés.

Enfin je remercie toutes les personnes m'ayant entouré durant ce stage et aidé à travailler et à rédiger ce rapport, à savoir Clément, Antoine, Mathieu, Mathias, Clément B., Pierre-Étienne, Rémy et Marion.

Table des matières

1	Objet et but du document	3
1.1	Présentation du projet	3
1.2	Présentation du document	3
2	Documents de référence	5
2.1	Panorama du domaine des systèmes d'information (Cabanac, 2012)	5
2.2	Cours de concepts fondamentaux de bases de données	5
2.3	Cours d'optimisation de requête	5
2.4	Cours de langage de requêtes	5
2.5	Cours de statistiques exploratoires et inférentielles	5
2.6	DUT Informatique	6
3	Terminologie	7
3.1	Recherche scientifique	7
3.2	Base de données	8
3.3	Statistiques	8
3.4	Web	9
3.5	Divers	9
4	Organisation	11
4.1	Entreprise	11
4.2	Équipe du projet	12
4.3	Planification	12

5	Valorisation d'Inforsid	15
5.1	Présentation du contexte	15
5.2	Intégration des données des éditions 2012 et 2013	16
5.3	Valorisation du congrès	18
5.4	Modification du site web	21
6	Étude de genre	27
6.1	Présentation du contexte	27
6.2	Récupération et mise en forme des données	27
6.3	Présentation des données et des tests utilisés	28
6.4	Répartition géographique	31
6.5	Comparaison de générations	31
6.6	Conclusions de l'étude	38
7	Méthodes et outils utilisés	41
7.1	Base de données	41
7.2	Analyse et mise en forme des données	42
7.3	Gestion de configuration	43
8	Assurance et contrôle qualité	45
8.1	Compte-rendus hebdomadaires	45
8.2	Revues	46
9	Bilan	49
9.1	Bilan du projet	49
9.2	Bilan personnel	49

1 — Objet et but du document

1.1 Présentation du projet

Afin de valider ma troisième année de licence j'ai dû réaliser un stage afin de mettre en pratique les connaissances acquises. J'ai choisi de me tourner vers la recherche afin d'observer l'autre facette du métier d'enseignant-chercheur, et ai donc effectué mon stage à l'IRIT¹, 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 en ce moment.

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 φ -index (Schubert, 2012; Cabanac, 2013) 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 présente le travail réalisé durant mon stage et notamment les thématiques abordées.

Je 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'introduirai également mon lieu de travail, mon collaborateur et mon planning.

J'exposerai ensuite le travail réalisé durant mes deux missions – la valorisation d'Inforsid puis l'étude de genre – ainsi que les méthodes et outils utilisés. J'exposerai également le suivi

1. Institut de Recherche en Informatique de Toulouse

organisé avec mon maître de stage et mon tuteur universitaire.

Enfin je conclurai en exposant tout ce que ce stage a pu m'apporter – tant sur le plan professionnel que sur le plan personnel – mais également tout ce qu'il a pu apporter à mon maître de stage et à l'IRIT.

2 — Documents de référence

2.1 Panorama du domaine des systèmes d'information (Cabanac, 2012)

L'article scientifique (Cabanac, 2012) – intitulé *Shaping the landscape of research in information systems from the perspective of editorial boards: A scientometric study of 77 leading journals* – 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 systèmes d'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.

2.2 Cours de concepts fondamentaux de bases de données

Le cours « Concepts fondamentaux de bases de données » de M. Morvan, M. 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 « Optimisation de requête » de M. Hameurlain, M. 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 « 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 relationnelles 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 « Statistique exploratoire » et « 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 Recherche scientifique

<i>SI / IS</i>	Systèmes d'Information ou <i>Information Systems</i> , domaine de recherche traitant de la collecte et du traitement d'informations.
<i>IA / AI</i>	Intelligence Artificielle ou <i>Artificial Intelligence</i> , 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 quantitative 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>	<i>5-year Journal Impact Factor</i> , indicateur du nombre moyen de citations de chaque article publié par le journal sur cinq ans, 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>	<i>Digital Bibliography & Library Project</i> , site web publiant des notices bibliographiques en informatique hébergé par l'université de Trèves en Allemagne existant depuis les années 1993 (Ley, 2002), consultable à l'adresse http://dblp.uni-trier.de/ .
<i>Congrès</i>	Rassemblement de chercheur-se-s travaillant sur les mêmes thèmes permettant à ceux-ci de présenter leur travail à leur pairs.
<i>Comité de programme</i>	Ensemble de chercheurs sélectionnant les thèmes des différentes sessions d'un congrès et les 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>Notice bibliographique</i>	Recueil de nombreuses données concernant une édition de congrès, telles que les articles présentés ou la composition du comité de programme.

Inforsid INformatique des ORganisations et Systèmes d'Information et de Décision, Congrès réunissant des chercheurs en SI depuis 1983.

3.2 Base de données

BD Base de Données, ensemble structuré et organisé de données permettant le stockage de grandes quantités d'informations afin d'en faciliter l'exploitation (ajout, mise à jour, recherche de données).

SGBD 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.

Table Structure stockant des données sous forme de tuples selon un schéma prédéfini.

Tuple Ensemble d'attributs caractérisant une ligne de la table – exemple : pour une table Employé un tuple contiendra (numéro d'employé, nom, prénom, service).

Clé primaire Attribut unique dans la table permettant d'identifier sans ambiguïté possible un tuple – dans l'exemple précédent la clé du tuple serait numéro d'employé.

Vue Requête enregistrée et nommée, utilisables dans les requêtes SQL comme une table et permettant de filtrer les données visibles par l'utilisateur mais aussi de clarifier l'affichage de certaines données – notamment lorsque l'on veut afficher des données provenant de plusieurs tables.

SQL *Structured Query Language*, langage permettant de créer, modifier et interroger les tables d'une base de données, mais également de gérer les droits des utilisateurs de la BD.

PL/SQL Langage de programmation créé par Oracle et permettant de créer des procédures et des fonctions au sein même d'une BD Oracle.

Procédure Portion de code effectuant un traitement sur les données – éventuellement passée en paramètres en entrée – sans renvoyer de résultat – attention : le fait qu'elle ne renvoie pas de résultat ne signifie pas forcément que l'utilisateur n'a aucun retour, une procédure peut très bien afficher des informations.

Fonction Portion de code effectuant un traitement sur des données – éventuellement passée en paramètres en entrée – et renvoyant un résultat – exemple : une fonction renvoyant le nombre de tuples d'une table.

Déclencheur Procédure provoquant un traitement particulier en fonction d'événements prédefinis, permettant ainsi d'automatiser certains traitements pour assurer la cohérence et l'intégrité de la base de données.

3.3 Statistiques

Test d'hypothèse 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</i> (H_0)	Point de vue par défaut concernant un phénomène donné. Il est nécessaire de connaître la loi de l'échantillon sous l'hypothèse nulle afin de pouvoir réaliser un test.
α	Taux d'erreur accepté pour le test (traditionnellement 5 % ou 1 %).
<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 d' α , on rejette l'hypothèse nulle. En d'autres termes, la <i>p-valeur</i> est la probabilité de rejeter à tort l'hypothèse nulle et donc d'obtenir un faux positif.

3.4 Web

<i>HTML</i>	<i>Hypertext Markup Language</i> , langage de balisage permettant de structurer sémantiquement et de mettre en forme le contenu des pages web, d'inclure des ressources multimédias dont des images, des formulaires de saisie, et des programmes informatiques.
<i>CSS</i>	<i>Cascading Style Sheets</i> , langage informatique qui sert à décrire la présentation des documents HTML et XML.
<i>W3C</i>	<i>World Wide Web Consortium</i> , un organisme de normalisation à but non-lucratif chargé de promouvoir la compatibilité des technologies du World Wide Web.

3.5 Divers

<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.
<i>TeX</i>	Système logiciel de composition de documents, largement utilisé par les scientifiques.

4 — Organisation

4.1 Entreprise

J'ai effectué mon stage au sein de l'IRIT – Institut de Recherche en Informatique de Toulouse, une unité mixte de recherche fondée en 1990 en partenariat entre l'université Paul Sabatier de Toulouse, le Centre national de la recherche scientifique, l'ENSEEIHT, l'Institut national polytechnique de Toulouse et l'université des Sciences Sociales de Toulouse. L'IRIT comprend 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'ai pour ma part intégré l'équipe SIG¹ dépendant du thème 2 « Indexation et recherche d'informations ».

La figure 4.1 illustre cette organisation.

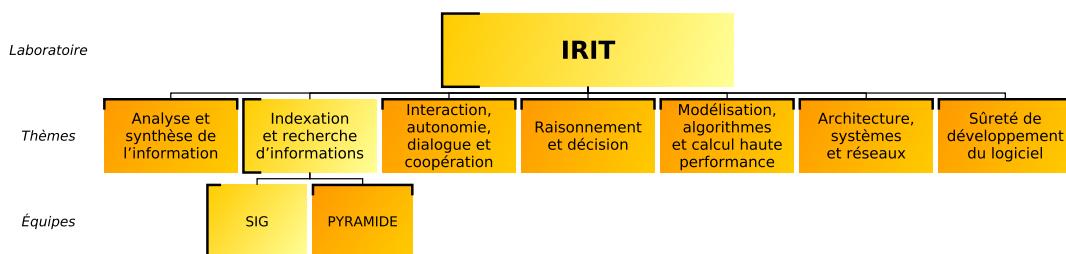
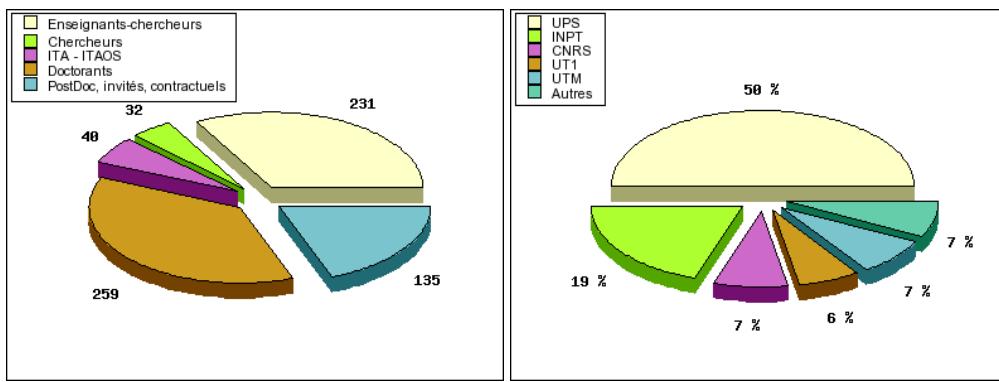


FIGURE 4.1 – Organisation des équipes de recherches de l'IRIT – il s'agit ici d'une représentation partielle se concentrant sur l'équipe SIG

L'IRIT comprend près de 700 personnes travaillant dans le monde de la recherche, comme illustré sur la figure 4.2.

1. Systèmes d'Informations Généralisés



(a) Par catégorie

(b) Par tutelle

FIGURE 4.2 – Personnel de l'IRIT (*Site officiel de l'IRIT, 2013*)

4.2 Équipe du projet

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

Guillaume Cabanac est docteur en informatique et maître de conférences dans l'équipe SIG. Il est enseignant à l'IUT Informatique de Rangueil et à l'université Paul Sabatier. Il est également membre du comité de rédaction des revues scientifiques suivantes qui sont en lien direct avec mon sujet de stage :

- Scientometrics (depuis 2013),
- Roars Transactions (depuis 2013),
- Ingénierie des Systèmes d'Information (depuis 2012).

La liste complète de ses publications est disponible sur publicationslist.org.

4.3 Planification

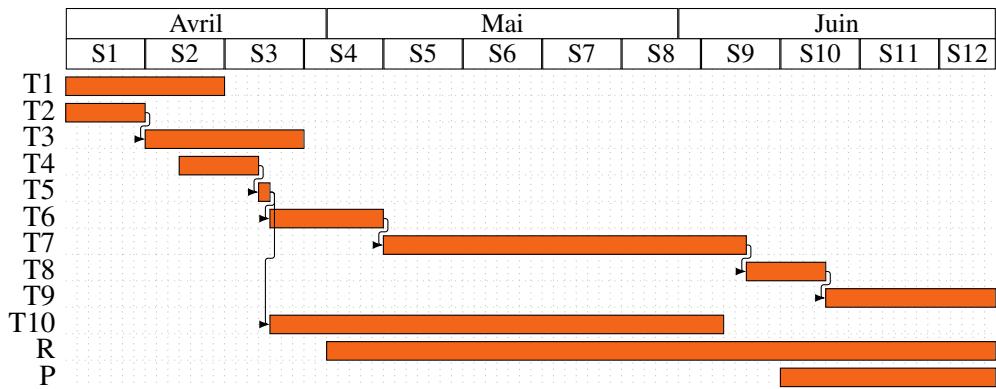
4.3.1 Planning prévisionnel

Le diagramme de Gantt prévisionnel de mon stage est représenté sur la figure 4.3.

4.3.2 Planning effectif

Le diagramme de Gantt effectif de mon stage est visible sur la figure 4.4. On peut constater que ma première mission – traitant du congrès Inforsid – a pris une semaine de plus que le temps estimé. Cependant cela n'a pas eu d'incidence fâcheuse car j'avais surestimé la durée d'autres tâches telles que l'intégration de nouvelles données dans les bases.

Je n'ai pas réalisé l'insertion des comités de rédaction du domaine IA dans la base de travail car je me suis concentrée sur mon étude et n'ai pas trouvé le temps de me concerter avec mon maître de stage sur comment mener à bien cette tâche. Bien que mon planning ait subi quelques ajustements au cours du stage j'ai donc réussi à tenir mes délais et à produire le travail demandé.



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 logiciel en 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 comités de rédaction 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 dans la base de travail.
R	Rédaction du rapport de stage.
P	Préparation de la soutenance.

FIGURE 4.3 – Diagramme de Gantt prévisionnel de mon stage

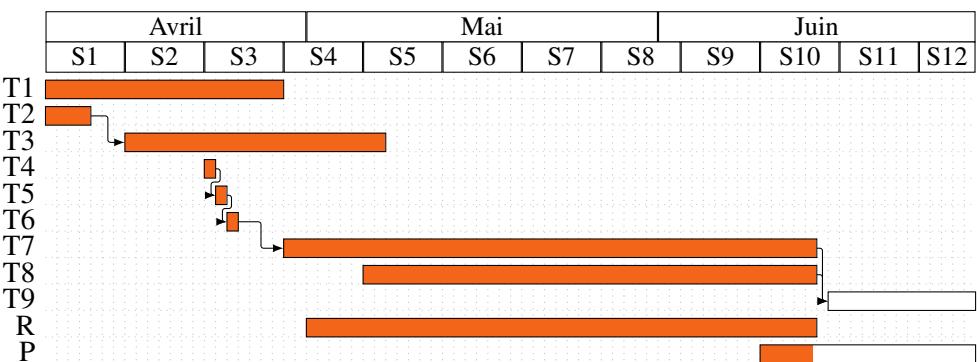


FIGURE 4.4 – Diagramme de Gantt effectif de mon stage – la description des tâches est disponible dans la figure 4.3

5 — Valorisation d'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. 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 en 2010 une base de données Oracle et une application web¹ centralisant ces informations. Celles-ci permettent, pour chaque édition du congrès, de visualiser son comité de programme et la liste des articles présentés avec leurs auteurs respectifs. L'application web comporte 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 aux 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, jusqu'alors les présidents n'avaient aucune règle ou aide pour sélectionner d'éventuels membres. Ils devaient donc se fier à leur connaissance de la communauté. Cela pouvait entraîner l'oubli de certains membres de la communauté, ou la favorisation de certains. La suggestion de membres via l'application se base sur un algorithme et permet une constitution plus éclairée, tout en s'assurant de n'oublier aucun chercheur. L'algorithme liste les chercheurs :

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

1. Consultable à l'adresse <http://www.irit.fr/~Guillaume.Cabanac/inforsid>.

- 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 2008.

Mon travail consistait à 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 et d’insérer les données des éditions 2012 et 2013 sans souci.

5.2.1 Principe de la transformation des données

Les données générales des congrès sont présentes dans un fichier texte contenant, pour chaque édition, une ligne de la forme « année, ville ». Les informations détaillées se trouvent dans 2 fichiers texte placés dans un répertoire nommé selon l’année concernée :

- un fichier contenant les membres du comité de programme,
- un fichier contenant les articles présentés durant le congrès avec leurs auteurs.

Une fois les données dans ces fichiers, l’application C crée les fichiers destinés à être insérés dans la base de données construite selon le MCD présenté en figure 5.1. La structure des fichiers est présentée dans le tableau 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	idArticle;titre;année;
congres.txt	année;idVille;
personne.txt	idPersonne;prénom;nom;
ville.txt	idVille;nomVille;paysVille;
ecrire.txt	idArticle;idPersonne;idVille;rang ²
membre.txt	idPersonne;année;rôle ³ ;idVille

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,
- 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.

5.2.2 Améliorations apportées au processus d’intégration des données 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-

2. Position à laquelle la personne apparaît lors de la déclaration des coauteurs dans l’article.

3. Président(e) = ‘P’, membre = ‘M’ ou adjoint(e) = ‘A’

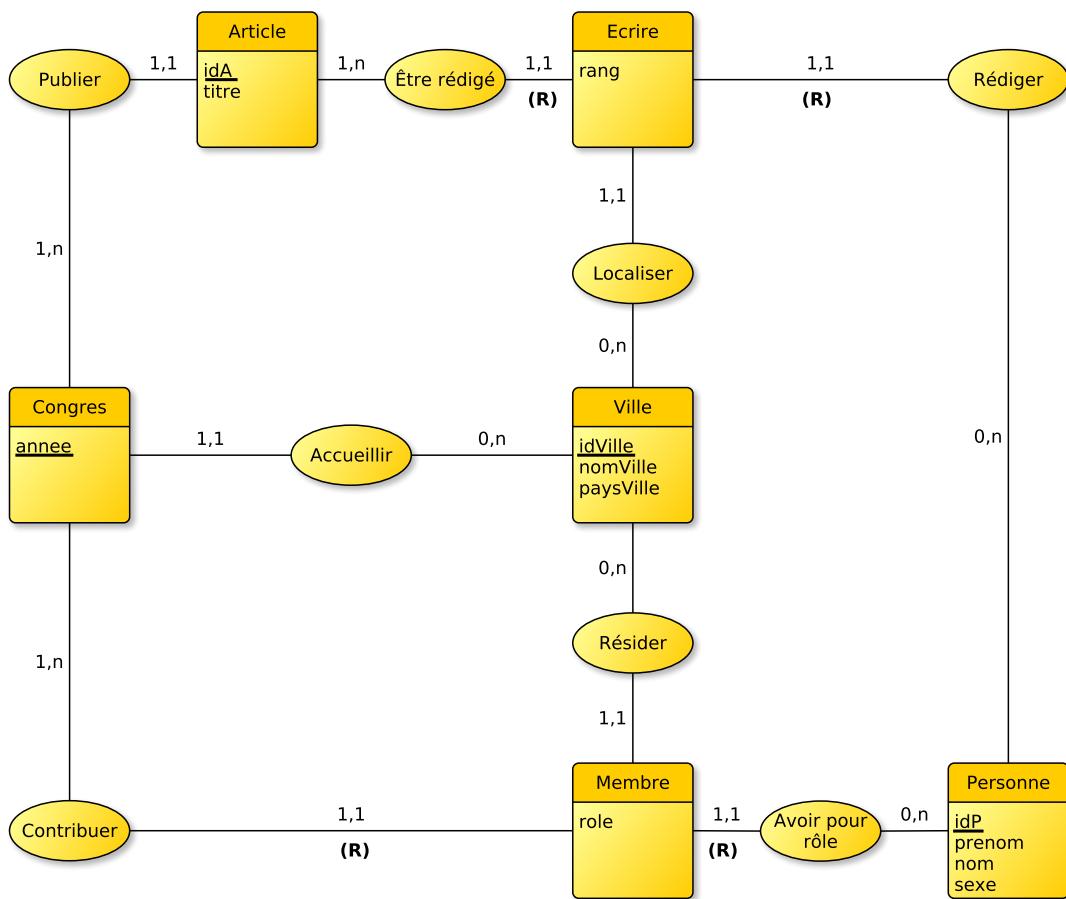


FIGURE 5.1 – Modèle Conceptuel de Données de la base utilisée par l’application (Ternisien, 2011)

ci étant dispersées dans de nombreux fichiers, il aurait été fastidieux de 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 mis en place une méthode de détection des synonymes simple et pratique pour l’utilisateur. Celui-ci devait simplement supprimer d’une liste les couples de synonymes détectés par erreur puis lancer une procédure qui prenait en charge la « fusion » des deux entités.

Cette méthode est présentée plus en détail dans les annexes.

Gestion des pays

Les pays correspondant aux villes présentes dans la base devaient avant être précisés manuellement dans un fichier texte, et la méthode de récupération des pays insérés précédemment présente dans le programme en C déclenchait souvent des erreurs.

J’ai donc décidé d’automatiser au maximum la procédure. J’ai ajouté à la base de données une table contenant de nombreux couples « Ville - Pays » et mis en place une procédure mettant à jour les villes d’Inforsid à partir de ceux-ci. Si, après une insertion de données, l’une des villes n’était pas présente dans la table de référence elle apparaissait dans une vue où l’utilisateur pouvait préciser manuellement son pays. Il n’avait ensuite qu’à relancer la procédure.

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.

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 repose sur les articles présentés lors des éditions du congrès et est défini comme suit :

- à chaque article est attribué un poids de 1,
- ce poids est attribué à la ville de rattachement que l'auteur indique dans son article,
- si l'article est rédigé par plusieurs auteurs, alors ce poids est réparti équitablement entre tous les auteurs (et donc entre leurs villes de rattachement).

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

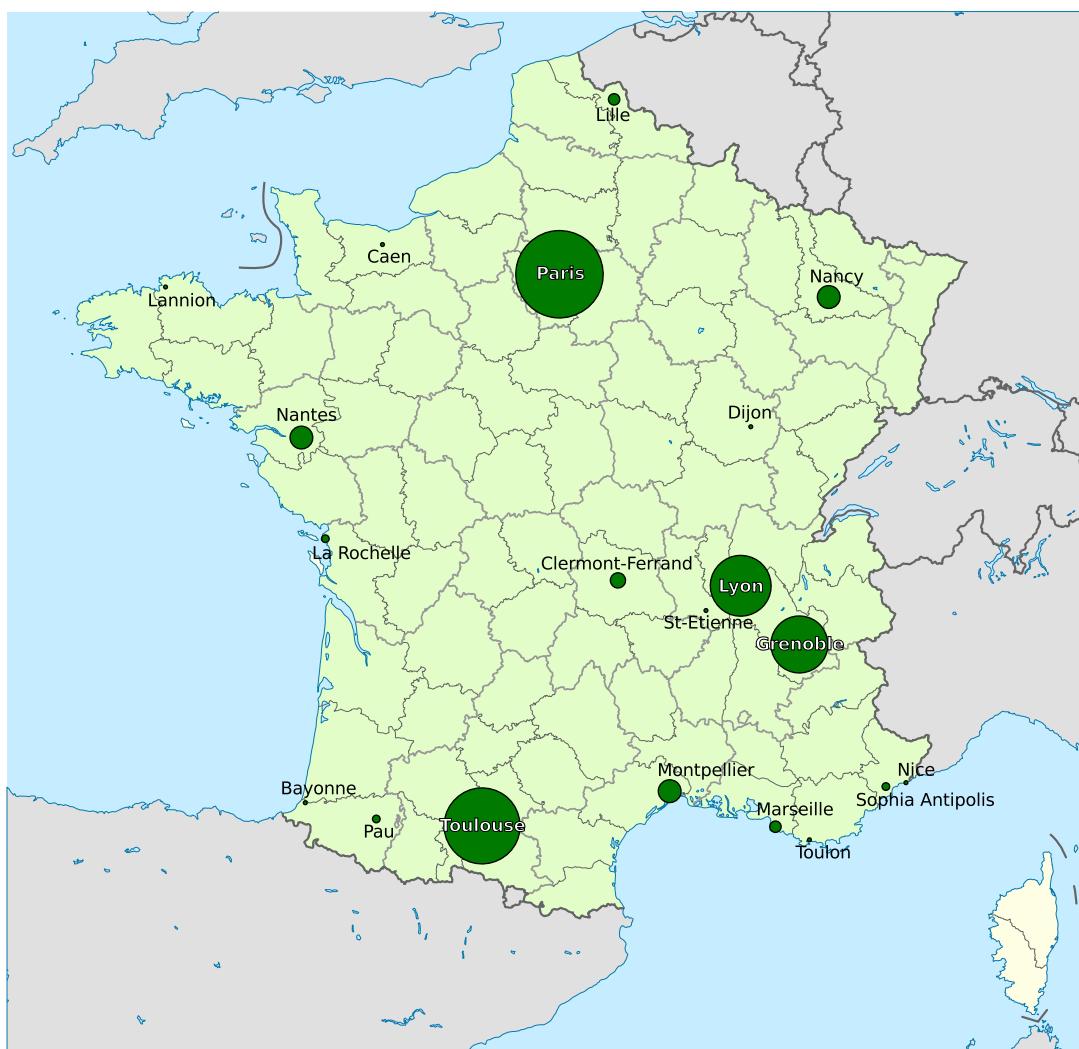


FIGURE 5.2 – Les 20 villes françaises les plus présentes dans les affiliations des auteurs d'Inforsid

Inforsid étant un congrès francophone réunissant des chercheurs de diverses nationalités différentes, j'ai ensuite souhaité représenter son rayonnement international. Pour cela j'ai utilisé les poids des pays – calculés en additionnant les poids de leurs villes – qui m'ont permis de

définir une « échelle de teinte » : plus le pays était vert plus il était impliqué dans Inforsid.

J'avais commencé à travailler sur une mappemonde, mais la majorité des pays participants étant en Europe, la lisibilité était bien trop mauvaise. Guillaume Cabanac avait 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 coté (en utilisant la même méthode que celle utilisée pour représenter les villes). Le résultat est présenté en figure 5.3.

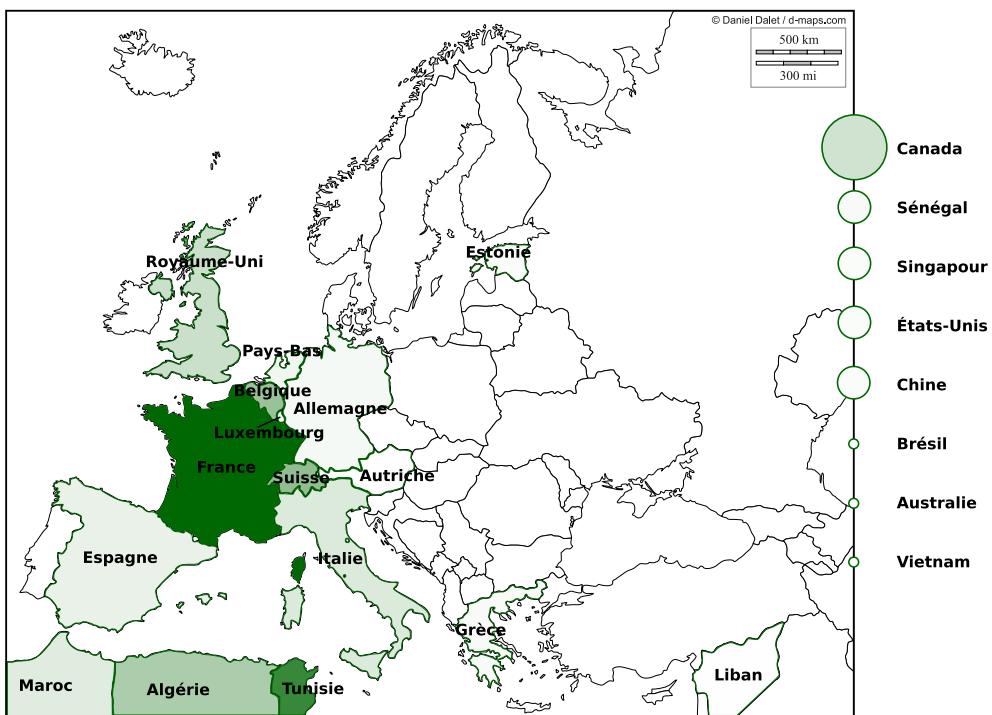


FIGURE 5.3 – Les pays des auteurs publiant à Inforsid

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 Wordle⁴.

Pour évaluer les thèmes abordés j'ai pris comme source d'information les termes employés dans les titres des articles présentés. Ces termes étaient « normalisés » afin de réaliser des calculs pertinents : suppression des accents, des caractères spéciaux et des « s » finaux. J'ai donc créé trois vues associant à chaque terme son nombre d'occurrences selon une période donnée :

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

Je me suis limitée aux 50 termes les plus fréquents afin de ne pas surcharger les nuages de mots et ainsi faciliter leur lecture – les mots vides n'étaient bien entendu pas compris parmi ceux-ci. 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. Il a ensuite fallu que je fasse un filtrage manuel dans la table

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



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

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 diminués afin de ne pas fausser les résultats – par exemple « donnee » étant présent dans « base de donnee » il fallait ôter de son poids le poids de l'expression. J'ai ainsi obtenu les wordles présents dans les figures 5.7, 5.8 et 5.9.

5.3.3 Influence des membres

Je devais identifier les chercheurs participant à Inforsid qui sont 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_dblp2013`, contenant notamment les comités de rédaction des 77 journaux traités dans (Cabanac, 2012) (pour plus d'informations, voir le chapitre 6). J'ai donc calculé l'intersection entre les membres d'Inforsid et les membres de comités de rédaction présents dans DBLP . Les chercheurs obtenus sont présentés dans le tableau 5.2.

J'ai ensuite 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, soulignant 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 national. La carte réalisée est visible dans la figure 5.11.

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_Traitemet` 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.

Afin d'améliorer la qualité du code, 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 :

- `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 désormais du plus récent au plus ancien). J'ai également modifié le code afin de n'avoir à modifier qu'un seul élément lors d'une migration de l'application sur une nouvelle base de données, et non toutes les procédures.

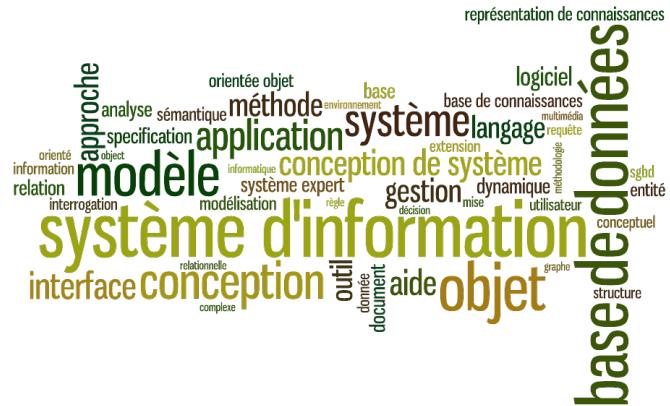


FIGURE 5.7 – Termes représentant Inforsid de 1983 à 1993 en prenant en compte les expressions



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

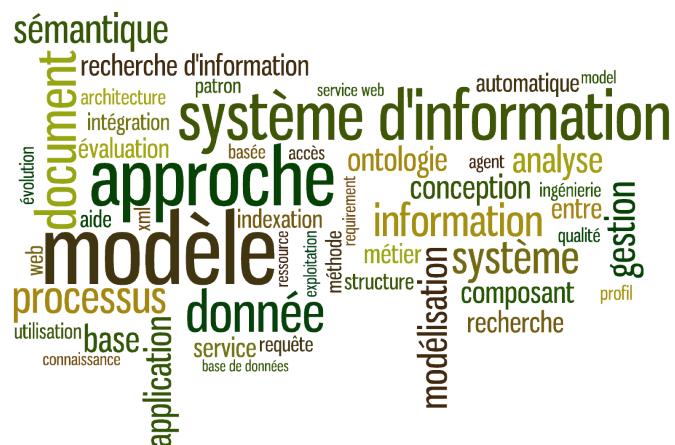


FIGURE 5.9 – Termes représentant Inforsid de 2004 à 2013 en prenant en compte les expressions

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 PIROTTÉ
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

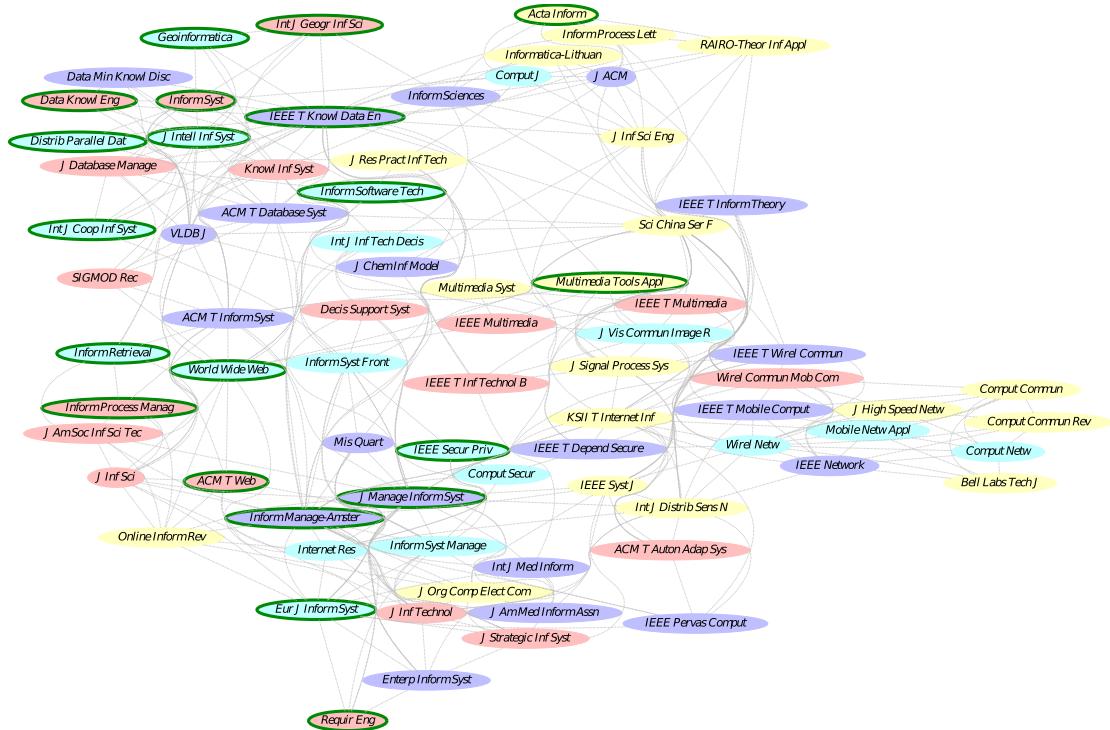


FIGURE 5.10 – Cartes des 77 principaux journaux du domaine IS (Cabanac, 2012). Les journaux ayant des comités de rédaction auxquels participent des membres d'Inforsid sont repérés par un liséré vert

5.4.2 Création d'une nouvelle charte 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, afin d'être conforme aux recommandations du W3C. Cette mise à niveau m'a permis d'utiliser Bootstrap⁵ 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 Inforsid⁶. Il aurait été préférable de faire un lien vers chaque article mais malheureusement nous n'avons pas trouvé de moyen pour réaliser ceci de manière automatique.

5. Bibliothèque CSS et Javascript éditée par Twitter.

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

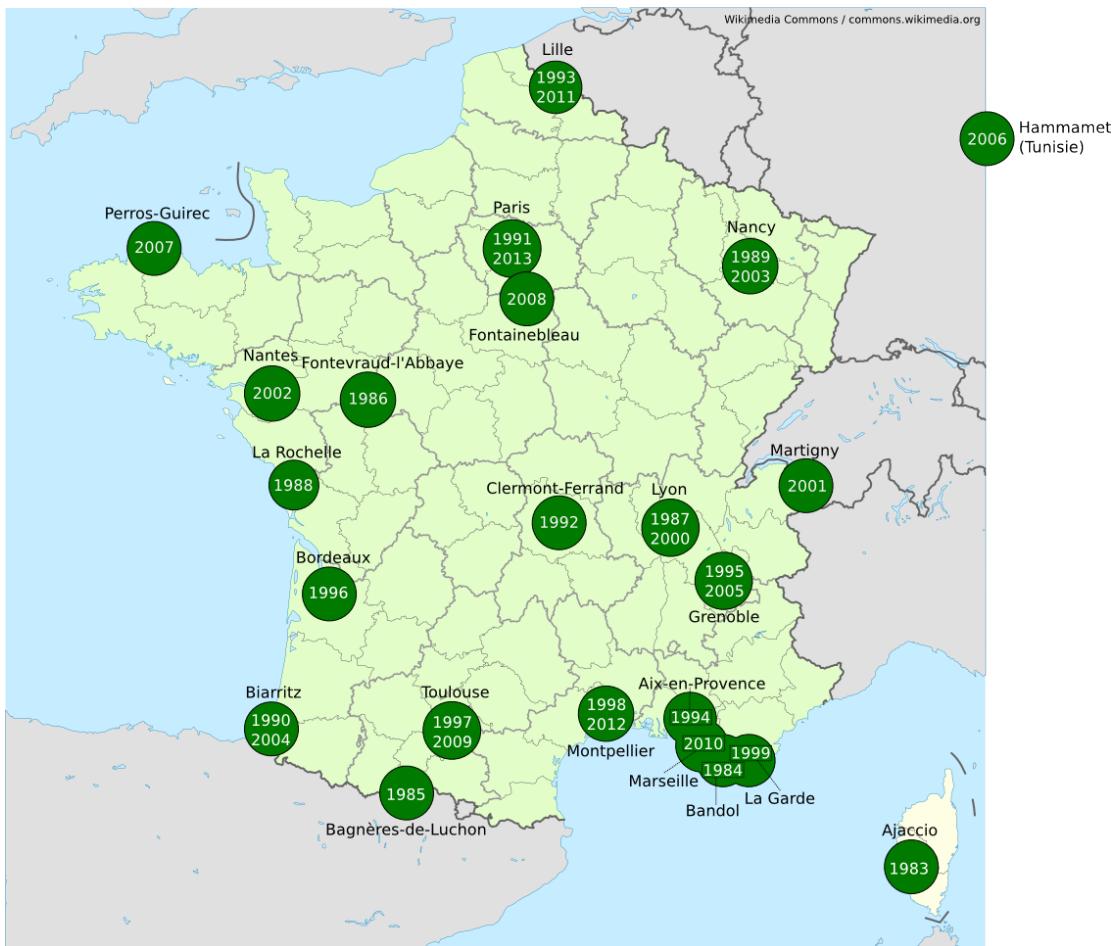


FIGURE 5.11 – Villes ayant accueilli le congrès Inforsid

Congrès Inforsid

Les éditions du congrès

1983	1984	1985	1986	1987	1988	1989			
1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
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

Suggestion de chercheurs pour le CP
 Expérimental

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



Wikimedia Commons / commons.wikimedia.org

FIGURE 5.13 – Nouvel aspect du site web

6 — Étude de genre

6.1 Présentation du contexte

Les femmes sont depuis toujours sous-représentées dans le domaine informatique (de Palma, 2001; Stross, 2008), cependant au fur et à mesure de l'évolution de la société, ces disparités s'amenuisent (Arensbergen et al., 2012).

Cette étude de genre avait pour but de questionner 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 SI.

Les résultats du stage alimenteront la discussion en cours sur la représentation des femmes en 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 le même schéma 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 Lefèuvre (stagiaire de Guillaume Cabanac en 2010). 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 réactivé les indexées et les contraintes (désactivés par SQL*Loader pour des raisons de performance lors de l'insertion des données). 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 grâce aux annotations manuelles réalisées par Guillaume Cabanac.

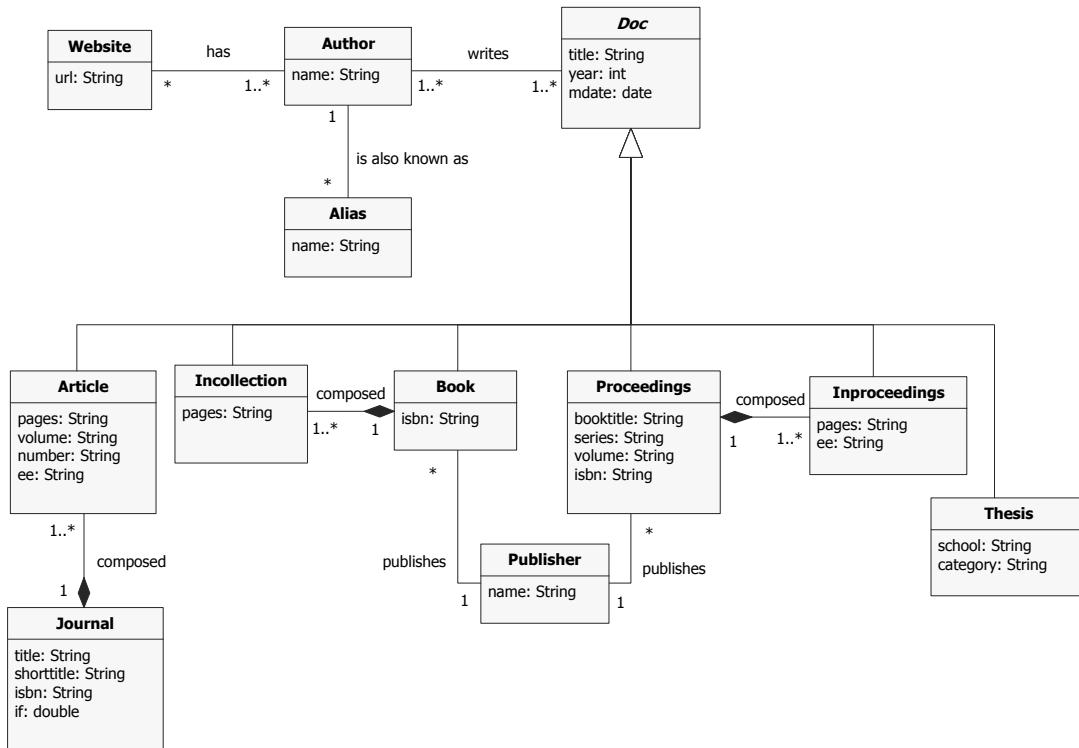


FIGURE 6.1 – Structure de la base de données contenant les données de DBLP (Lefevre, 2010; Cabanac, 2011)

6.3 Présentation des données et des tests utilisés

6.3.1 Présentation générale des données

DBLP regroupe une quantité importante de données regroupées dans un fichier XML de 1,2 Go (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.

Bien évidemment le domaine SI n'est pas représenté uniquement par ces journaux, nous ne disposons ici que d'une partie des membres de comités de rédaction du domaine. Étant donné le nombre de chercheurs présents dans notre base (voir tableau 6.1), même en prenant en compte le fait que de nombreux chercheurs ne dépendent pas du domaine SI, nous pouvions considérer qu'il s'agissait d'un échantillon d'individus indépendants et identiquement distribués. Cette propriété justifie le choix fait de réaliser des tests statistiques afin d'inférer le comportement de la population totale des membres de comités de rédaction du domaine SI, et éventuellement suggérer, de façon très prudente, quelques tendances dans la population globale des chercheurs en SI – les membres de comités de rédaction étant des chercheurs choisis par leurs pairs pour leur influence dans le milieu, et non sélectionnés au hasard, il est en effet difficile de savoir si leur comportement peut être généralisé au reste de la communauté.

TABLEAU 6.1 – Nombre d’éléments référencés dans notre base de données.

Nom de l’élément		Nombre d’éléments
Chercheur-se-s		1 265 195
Journaux scientifiques		1 346
Documents		2 265 005
<i>dont</i>	Articles	957 452
	Thèses	6 927
	Livres	9 797
Extraits de livres		21 932
Conférences		20 080
Participations à des conférences		1 247 430

6.3.2 Présentation des tests statistiques

Shapiro-Wilk (Shapiro & Wilk, 1965)

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

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

où

- $x_{(i)}$ désigne la ième statistique d’ordre , c.-à-d., le i^e 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 (James, 1971)

Le test non paramétrique 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 (Wilcoxon, 1945)

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

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

TABLEAU 6.2 – Journaux du domaine IS pour lesquels nous disposons du comité de rédaction (Cabanac, 2012).

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 Representation
GeoInformatica	JAMIA
I. J. Medical Informatics	JASIST
IEEE MultiMedia	JIT
IEEE Network	Journal of Chemical Information and Modeling
IEEE Pervasive Computing	Journal of Research and Practice in Information Technology
IEEE Security & Privacy	Knowl. Inf. Syst.
IEEE Systems Journal	MIS Quarterly
IEEE Trans. Dependable Sec. Comput.	MONET
IEEE Trans. Knowl. Data Eng.	Multimedia Syst.
IEEE Trans. Mob. Comput.	Multimedia Tools Appl.
IEEE Transactions on Information Technology in Biomedicine	Online Information Review
IEEE Transactions on Information Theory	Requir. Eng.
IEEE Transactions on Multimedia	Science in China Series F : Information Sciences
IEEE Transactions on Wireless Communications	SIGMOD Record
IJDSN	Signal Processing Systems
Inf. Process. Lett.	TAAS
Inf. Process. Manage.	TIIS
Inf. Retr.	TWEB
Inf. Sci.	VLDB J.
Inf. Syst.	Wireless Communications and Mobile Computing
Informatica, Lith. Acad. Sci.	Wireless Networks
Information & Management	World Wide Web
Information & Software Technology	
Information Systems Frontiers	
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.

Genre	Effectif
Femmes	422
Hommes	2402
Genre non déterminé	26
Total	2850

où R_i est le rang de x_i dans l'ensemble des (x_1, \dots, x_p) concaténés à l'ensemble des (y_1, \dots, y_q) , le tout classé par ordre croissant.

6.4 Répartition géographique

Une des premières questions auxquelles j'ai souhaité répondre a été : quelle est la répartition géographique des membres féminins de comités de rédaction ? Les membres de comités de rédaction de notre BD sont affiliés à 55 pays, dont 32 comptent des membres féminins.

Pour permettre une visualisation pertinente des principaux pays, j'ai décidé de réaliser un diagramme en bâton avec Gnuplot, visible en figure 6.2 – pour des raisons de place seuls les 20 pays ayant au moins trois membres féminins de comités de rédaction sont visibles.

Pour mieux visualiser la différence entre le nombre de membres masculins et féminins j'ai décidé de réaliser un autre diagramme en bâton avec cette fois-ci la proportion de membres féminins pour chacun des pays figurant dans le diagramme précédent.

On constate que le pays le plus paritaire¹ est la Turquie, bien que les femmes ne représentent qu'environ 27 % des membres de comités de rédaction – il faut cependant noter que la Turquie ne possède que onze membres de comités de rédaction, et donc seulement trois membres féminins.

Ces premières analyses confirment nos intuitions initiales sur la sous-représentation des femmes dans le domaine SI. En effet dans notre cas seul 58 % des pays comprend des membres féminins de comités de rédaction et le nombre de ces membres n'atteint jamais le tiers de membres totaux de leur pays. On peut également noter que même les pays réputés pour avoir une politique de recherche très favorable, tels que les États-Unis, ne sont pas épargnés par cette sous-représentation.

6.5 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 SI.

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.

1. Se dit d'une assemblée formée de représentants en nombre égal des parties en présence – dans notre cas une assemblée comptant autant d'hommes que de femmes (*Larousse.fr*, 2013).

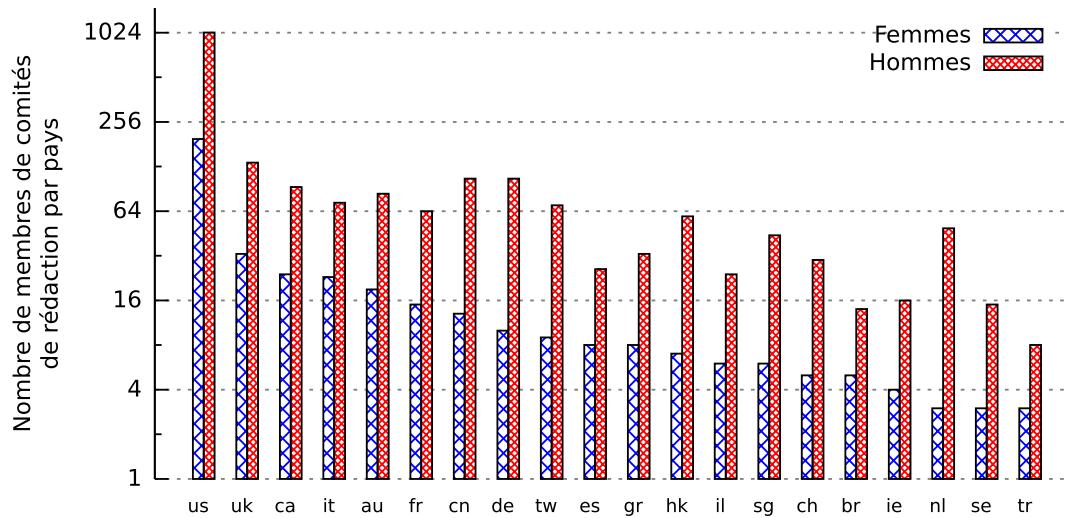


FIGURE 6.2 – Pays des membres féminins de comités de rédaction – à titre de comparaison le nombre de membres masculins de chaque pays est également indiqué – pour des raisons de place seuls les 20 pays ayant au moins trois membres féminins apparaissent ici.

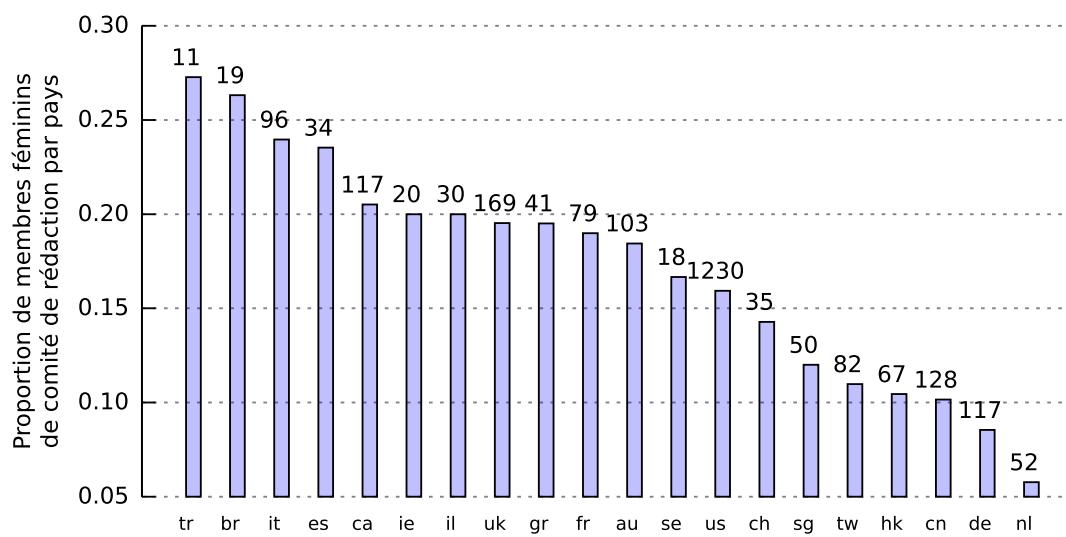


FIGURE 6.3 – Proportion des membres féminins de comités de rédaction des 20 pays figurant dans la figure 6.2 – le nombre total de membres de comités de rédaction du pays est indiqué au dessus du bâton correspondant.

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.

L'année 2000 a été choisie comme limite car les articles lus suggéraient que l'écart de production observé entre hommes et femmes se creusait principalement durant les dix premières années de carrières, il fallait donc que ma nouvelle génération commence au plus tard en 2003. Afin d'avoir une quantité suffisamment importante de données à analyser j'ai décidé de la faire commencer un peu plus tôt, et donc de sélectionner une date facilement repérable, comme 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 26 chercheurs pour lesquels le genre n'a pas pu être déterminé ne sont pas pris en compte ici.

	Femmes	Hommes	Total
Ancienne génération [1958 – 1999]	289	1882	2171
Nouvelle génération [2000 – 2013]	133	520	653
Total	422	2402	2824

Production et productivité

Le premier élément que j'ai souhaité tester a été la production 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 s'estomper avec les années.

Pour calculer la production d'un chercheur j'ai décidé d'utiliser la métrique de production de l'état de l'art (Egghe, Rousseau, & Van Hooydonk, 2000) 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 $r \in \llbracket 1; n \rrbracket$ le rang du chercheur dans la liste de ceux-ci.

Les productions selon le genre de l'ancienne et de la nouvelle génération ainsi calculées sont représentées en figures 6.4 et 6.5.

On constate que la différence clairement visible pour l'ancienne génération (où la médiane pour les femmes vaut 19,38 et celle des hommes 27,13, soit un delta de 7,75) s'atténue fortement pour la nouvelle (où cette fois-ci la médiane pour les femmes vaut 8,24 contre 8,84 pour les hommes, soit un delta de 0,60). Bien évidemment la production de l'ancienne génération est supérieure à celle de la nouvelle génération étant donné que ses chercheurs ont bien plus d'années de carrière à leur actif.

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 s'estompait avec le temps.

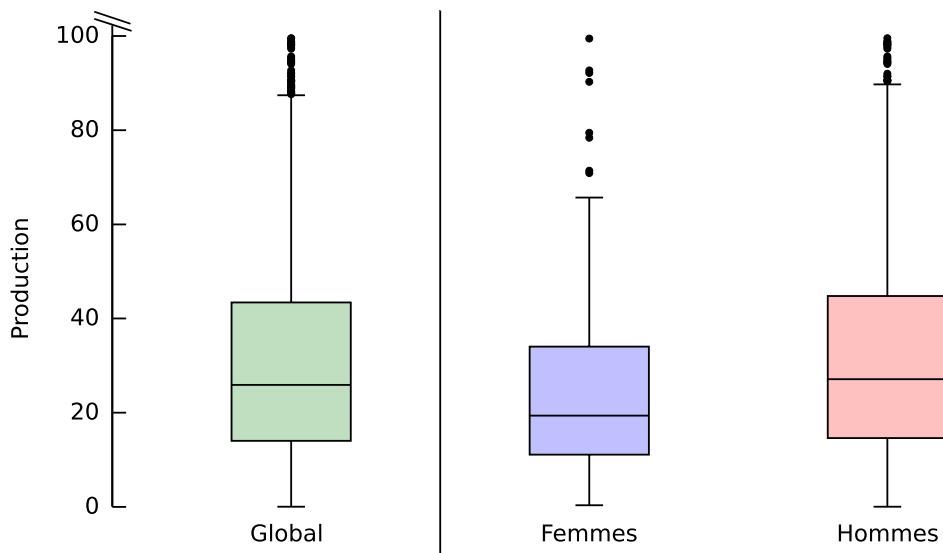


FIGURE 6.4 – Production de l'ancienne génération de chercheurs par genre – les 78 chercheurs (3,5 % de la génération) ayant une production supérieure à 100 n'apparaissent pas sur ce graphique.

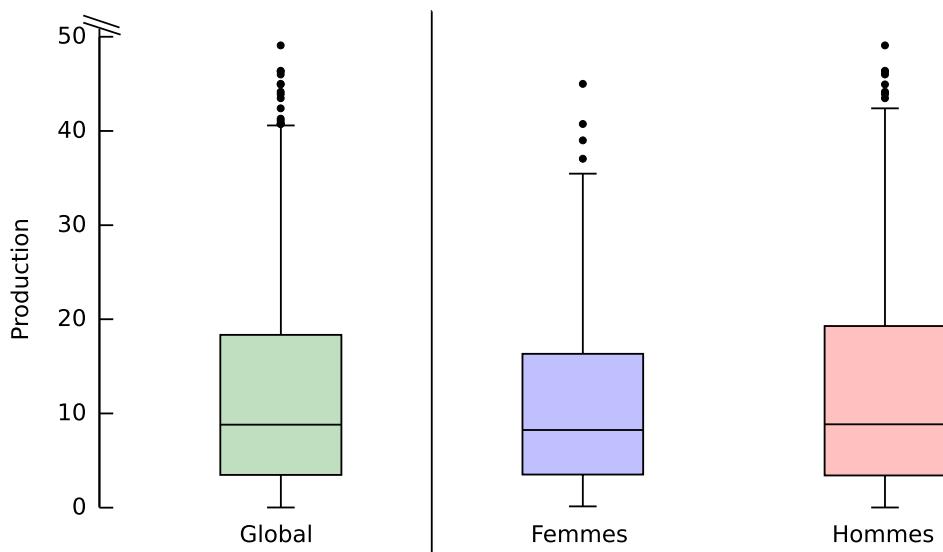


FIGURE 6.5 – Production de la nouvelle génération de chercheurs par genre – les 20 chercheurs (3,5 % de la génération) ayant une production supérieure à 50 n'apparaissent pas sur ce graphique.

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. Ils indiquent 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 « Production des chercheuses de l'ancienne génération », « Production des chercheurs de l'ancienne génération », « Production des chercheuses de la nouvelle génération » et « Production des chercheurs de la nouvelle génération », indiquant si ces échantillons ne suivent pas une loi normale.

Échantillon testé		<i>W</i>	<i>p-value</i>
Ancienne génération [1958 – 1999]	Femmes	0,7158	$< 2,200 \cdot 10^{-16}$
	Hommes	0,8045	$< 2,200 \cdot 10^{-16}$
Nouvelle génération [2000 – 2013]	Femmes	0,7080	$6,235 \cdot 10^{-15}$
	Hommes	0,7818	$< 2,200 \cdot 10^{-16}$

Étant donné 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 présentés dans le tableau 6.6. Les deux tests concordent et suggèrent que :

- la différence de production entre membres masculins et féminins des comités de rédaction de l'ancienne génération est significative d'un point de vue statistique,
- 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	<i>p-value</i>
Ancienne génération [1958 – 1999]	KS	0,1599	$5,4240 \cdot 10^{-6}$
	W	221549,0000	$3,7830 \cdot 10^{-7}$
Nouvelle génération [2000 – 2013]	KS	0,0958	0,2857
	W	32960,5000	0,4043

Cette démarche présentait néanmoins un biais non négligeable : au sein d'une même génération, deux chercheurs peuvent avoir une durée de carrière très dissemblable. Pour pallier ce problème j'ai décidé, sur une suggestion de Guillaume, de diviser la production de chaque chercheur par la durée de sa carrière (calculée en fonction du plus ancien et du plus récent de ses documents présents dans la BD) afin d'obtenir sa productivité annuelle. J'ai également écarté les membres ayant participé à la rédaction de moins de 5 documents, ce qui nous fournit de nouveaux effectifs, présentés dans le tableau 6.7. Nous obtenons donc la productivité annuelle par chercheur visible dans les figures 6.6 et 6.7.

À nouveau, on observe une différence plus importante entre hommes et femmes pour l'ancienne génération (1,00 pour les femmes contre 1,19 pour les hommes) que pour la nouvelle (où la médiane des femmes vaut 0,98 contre 1,06 pour les hommes).

J'ai suivi la même démarche que pour les résultats précédents, les résultats du test de Shapiro-Wilk est visible dans le tableau 6.8. On peut voir qu'ici encore aucun de mes échantillons n'est

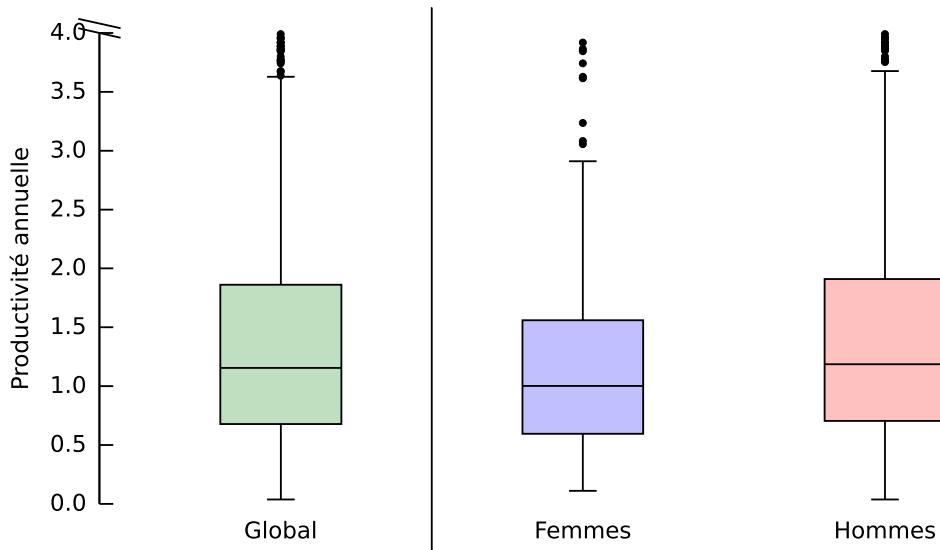


FIGURE 6.6 – Productivité annuelle de l’ancienne génération de chercheurs par genre – les 74 chercheurs (3,5 % de la génération) ayant une productivité annuelle supérieure à 4 n’apparaissent pas sur ce graphique.

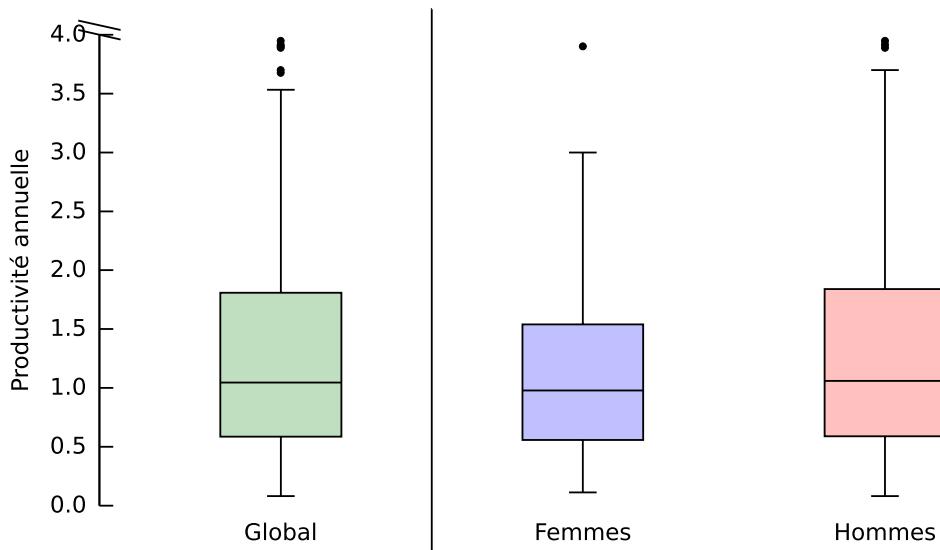


FIGURE 6.7 – Productivité annuelle de l’ancienne génération de chercheurs par genre – les 74 chercheurs (3,5 % de la génération) ayant une productivité annuelle supérieure à 4 n’apparaissent pas sur ce graphique.

TABLEAU 6.7 – Répartition des membres de comités de rédaction – les chercheurs pour lesquels le genre n'a pas pu être déterminé et ceux ayant participé à la rédaction de moins de 5 documents ne sont pas pris en compte ici.

	Femmes	Hommes	Total
Ancienne génération [1958 – 1999]	284	1849	2171
Nouvelle génération [2000 – 2013]	116	446	562
Total	400	2295	2695

distribué selon la loi normale – les *p-values* étant inférieures à 5 %. J'ai donc réutilisé les même tests que précédemment, les résultats sont visibles dans le tableau 6.9.

TABLEAU 6.8 – Résultats du test de Shapiro-Wilk sur les échantillons « Productivité annuelle des chercheuses de l'ancienne génération », « Productivité annuelle des chercheurs de l'ancienne génération », « Productivité annuelle des chercheuses de la nouvelle génération » et « Productivité annuelle des chercheurs de la nouvelle génération », indiquant si ces échantillons ne suivent pas une loi normale.

Échantillon testé		W	<i>p-value</i>
Ancienne génération [1958 – 1999]	Femmes	0,8023	$< 2,200 \cdot 10^{-16}$
	Hommes	0,8143	$< 2,200 \cdot 10^{-16}$
Nouvelle génération [2000 – 2013]	Femmes	0,7435	$5,983 \cdot 10^{-13}$
	Hommes	0,8251	$< 2,200 \cdot 10^{-16}$

TABLEAU 6.9 – Résultats des tests de Kolmogorov-Smirnov (*KS*) et Wilcoxon (*W*) indiquant si la différence de productivité annuelle entre hommes et femmes est significative.

Échantillon testé		Statistique de test	<i>p-value</i>
Ancienne génération [1958 – 1999]	<i>KS</i>	000000,1191	$1,8570 \cdot 10^{-7}$
	<i>W</i>	227603,5000	$2,9790 \cdot 10^{-8}$
Nouvelle génération [2000 – 2013]	<i>KS</i>	000000,0949	0,3778
	<i>W</i>	24209,0000	0,2871

Ces nouvelles analyses corroborent donc le résultat des précédentes : l'écart de production entre hommes et femmes tend à diminuer voire à disparaître avec le temps.

φ -index

J'ai ensuite pensé qu'il serait intéressant de comparer également l'évolution du φ -index entre générations (Schubert, 2012). Le φ -index mesure la *partnership ability* – l'aptitude à collaborer avec d'autres chercheurs et à conserver des collaborations. Schubert (2012) défini le φ -index comme ceci :

Un auteur a un φ -index de φ si, avec φ de ses n coauteurs, il a écrit au moins φ articles, et s'il n'a pas plus de φ articles en commun avec ses $n - \varphi$ autres coauteurs.

Prenons l'exemple d'Albert Einstein, auquel est attribué 272 articles scientifiques. Parmi ces articles, seuls 44 ont été co-écrits avec des pairs (24 co-auteurs au total). Einstein possède un

φ -index de 3 car il a écrit au moins trois articles avec trois de ses coauteurs (8 avec W. Mayer, 4 avec W.J. de Haas et 4 avec N. Rosen) et pas plus de 3 avec les 21 restants.

J'ai calculé cette mesure à l'aide de l'article (Cabanac, 2013). Les boîtes à moustaches correspondantes sont visibles en figures 6.8 et 6.9.

Même si l'écart entre les médianes des membres masculins et féminins est le même pour les deux générations (ancienne génération : médiane de 5 pour les femmes et 6 pour les hommes, nouvelle génération : médiane de 3 pour les femmes et 4 pour les hommes) les échantillons hommes et femmes semblent plus proches au niveau de la distribution dans la nouvelle 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 6.10).

TABLEAU 6.10 – Résultats du test de Shapiro-Wilk sur les échantillons « φ -index des chercheuses de l'ancienne génération », « φ -index des chercheurs de l'ancienne génération », « φ -index des chercheuses de la nouvelle génération » et « φ -index des chercheurs de la nouvelle génération », indiquant si ces échantillons suivent une loi normale.

Échantillon testé		<i>W</i>	p-value
Ancienne génération [1958 – 1999]	Femmes	0,9454	$7,026 \cdot 10^9$
	Hommes	0,9630	$< 2,200 \cdot 10^{-16}$
Nouvelle génération [2000 – 2013]	Femmes	0,9263	$2,032 \cdot 10^{-6}$
	Hommes	0,9192	$4,576 \cdot 10^{-16}$

TABLEAU 6.11 – Résultats des tests de Kolmogorov-Smirnov (*KS*) et Wilcoxon (*W*) indiquant si la différence de φ -index entre hommes et femmes est significative.

Échantillon testé		Statistique de test	p-value
Ancienne génération [1958 – 1999]	<i>KS</i>	0,1547	$1,2310 \cdot 10^{-05}$
	<i>W</i>	225853,5000	$3,7830 \cdot 10^{-07}$
Nouvelle génération [2000 – 2013]	<i>KS</i>	0,0770	0,5566
	<i>W</i>	32007,0000	0,1808

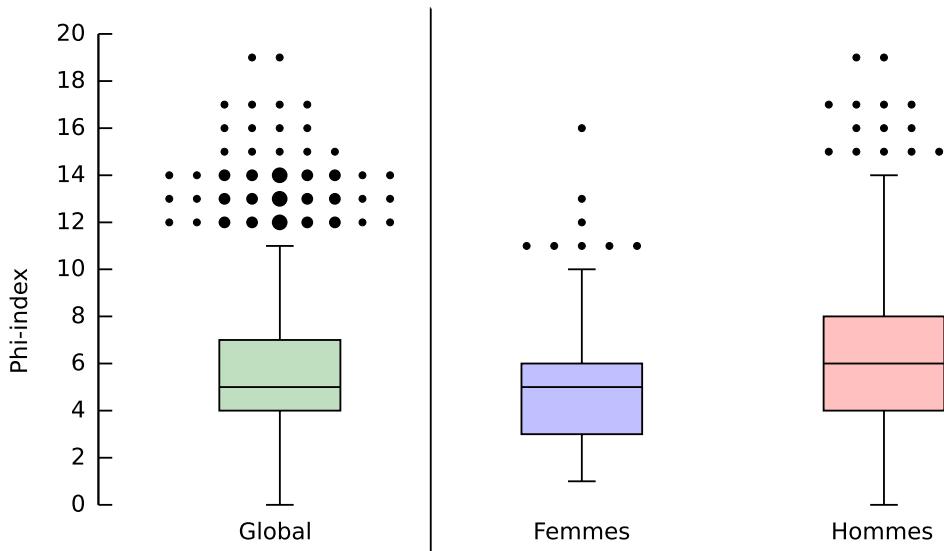
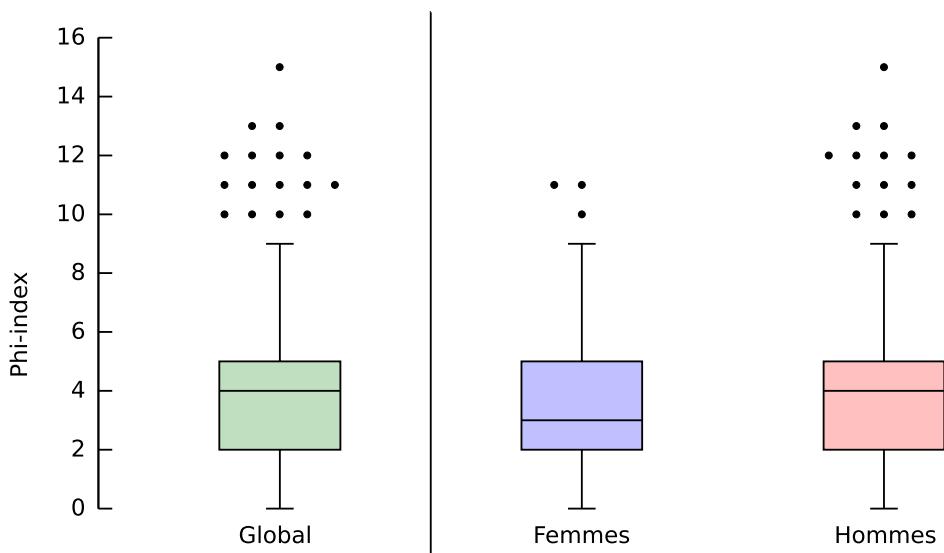
On constate que, comme pour la productivité, les deux tests concluent aux mêmes résultats (voir tableau 6.11) :

- la différence de φ -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.

6.6 Conclusions de l'étude

Toutes les analyses menées sur les membres de comités de rédaction de 77 principales revues du domaine SI nous permettent de tirer plusieurs conclusions que nous n'avions pas décelées dans les articles existants.

Tout d'abord la sous-représentation majeure des femmes dans le domaine SI est un phénomène global : des 55 pays auxquels sont affiliées les membres féminins de comités de rédaction, seuls 32 comprennent des membres féminins, soit seulement 58 %. Parmi ces 32 pays seuls deux ont

FIGURE 6.8 – φ -index de l'ancienne génération de chercheurs par genreFIGURE 6.9 – φ -index de la nouvelle génération de chercheurs par genre

plus de 30 % de membres féminins : la Croatie qui a deux membres féminins et la Slovénie qui a 2 membres féminins et 4 membres masculins.

Nous pouvons ensuite noter que l'écart de productivité et de φ -index entre hommes et femmes observé dans de nombreux articles scientifiques – et pouvant constituer une « justification » de la quasi absence des femmes aux postes de pouvoir, celles-ci étant supposément moins performante que les hommes (Hildrun et al., 2012; Nakhaie, 2002) – tend à diminuer voire à disparaître dans les nouvelles générations de chercheurs.

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 Database 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 des assistants de création et de configuration de bases de données.

7.1.2 SqlDeveloper

SQL Developer est un environnement de développement intégré 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 existantes mais également les fonctions, procédures, déclencheurs, séquences et autres objets présents dans la base.

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 de données 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 primaires,
- optimiser le mode de chargement «direct», c.-à-d. avec désactivation des éventuelles contraintes et indexes pour améliorer la vitesse de chargement des données,
- 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 produit 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 afficher les graphiques à l'écran ou les stocker dans des fichiers dans de nombreux formats.

Le programme peut être utilisé interactivement, et est accompagné d'une aide en ligne. L'utilisateur saisit 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^AT_EX

L^AT_EX est un langage et un système de composition de documents créé par Leslie Lamport en 1983.

Du fait de sa relative simplicité, il est devenu l'outil privilégié d'écriture de documents scientifiques employant TeX. 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).

L^AT_EX 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 ou alinéas par exemple) est dévolue au logiciel lors d'une compilation ultérieure.

7.2.5 BibTeX

BibTeX est un logiciel de gestion de références bibliographiques et un format de fichier conçu par Oren Patashnik et Leslie Lamport en 1985 pour L^AT_EX. Il sert à gérer et traiter des bases bibliographiques.

7.3 Gestion de configuration

7.3.1 Apache Subversion

Subversion – souvent abrégé SVN – est un logiciel de gestion de versions,. 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,
- des postes clients sur lesquels se trouvent :
 - les fichiers recopiés depuis le serveur, éventuellement modifiés localement depuis la dernière opération de synchronisation,
 - 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 de façon concorrente il le détecte et laisse l'utilisateur décider des portions du fichiers à modifier sur le serveur, sauf si les zones modifiées du fichier ne se « chevauchent » pas, dans ce cas il se charge d'intégrer les modifications sans intervention de l'utilisateur.

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 et Guillaume Cabanac me le rendait ensuite accompagné de ses annotations et observations. Il s’agissait d’un moyen de garder une trace de mon avancée pour mon maître de stage tout en m’imposant de rédiger et présenter progressivement l’avancée de mon travail.

8.1.1 Semaine 1

Guillaume a pu valider grâce à ce compte-rendu ma compréhension du mode de fonctionnement de l’insertion des données dans la base alimentant l’application web présentant Inforsid et prendre connaissance des modifications apportées à certaines procédures.

J’avais exprimé dans ce compte-rendu mon intention de déterminer les villes les plus importantes du congrès en calculant leur poids et pour ce faire il m’a conseillé d’utiliser une pondération fractionnelle pour les articles – c’est-à-dire diviser le poids de l’article par le nombre d’auteurs (Egghe et al., 2000).

Il m’a également demandé de changer quelques éléments de mise en forme de mon compte-rendu afin d’être plus proche d’une mise en page d’article scientifique.

8.1.2 Semaine 2

Ce compte-rendu présentait ma gestion des synonymes et des pays dans la base de données, ainsi que la modification de la page d’accueil qu’il m’avait demandé. Il a également pu valider les cartes des villes et pays contribuant le plus au congrès.

8.1.3 Semaine 3

Ce compte-rendu présentait les nuages de mots réalisés pour représenter les thèmes du congrès et la carte des journaux scientifiques s’appuyant sur des membres d’Inforsid dans leur comité de rédaction. Guillaume a validé les premiers mais en revanche la carte ne lui semblait pas correcte car mettant en valeur trop peu de journaux par rapport à ses estimations, et je devais donc la reprendre la semaine suivante. Il a également pu valider le nouvel aspect de l’application et m’a demandé d’ajouter à ma liste de tâche de la semaine suivante l’ajout de liens vers les actes complets du congrès.

8.1.4 Semaine 4

Ce compte-rendu faisant état de ma dernière semaine passée à travailler sur la valorisation d'Inforsid. Guillaume a ainsi pu valider les nuages de mots finaux – contenant des expressions de plusieurs mots – et la carte des journaux scientifiques s'appuyant sur des membres d'Inforsid afin que je les intègre sur le site web.

J'ai également présenté les premières données de l'étude de genre.

8.1.5 Semaine 5

Grâce à ce compte-rendu Guillaume a pu donner son aval pour l'ajout de la carte des villes ayant accueilli le congrès Inforsid sur le site web.

De plus il a pu prendre connaissance des calculs de φ -index et d'homophilie réalisés pour l'étude de genre et des premiers graphiques réalisés avec Gnuplot.

Il m'a demandé d'utiliser BibTeX pour les documents suivants afin d'avoir des références correctement présentées.

8.1.6 Semaine 6

Ce compte-rendu présentait plusieurs figures représentant les données extraites précédemment. Il introduisait également les résultats des premiers tests statistiques de l'étude, pour lesquels Guillaume m'a demandé de revoir la présentation.

8.1.7 Semaine 7

Le début de la septième semaine a été principalement consacré à la conception d'un nouveau calcul de l'homophilie. J'exposais donc dans ce compte-rendu les résultats trouvés sur un plus petit échantillon – 24 *gatekeepers* de *JASIST* – et expliquais l'impossibilité d'étendre cette mesure à un échantillon plus important.

Je présentais également les conclusions tirées de mes lectures d'articles scientifiques traitant d'études de genre et les pistes que je souhaitais suivre la semaine suivante. Guillaume a validé ces pistes et m'a également proposé de me pencher sur l'évolution du nombre de nouveaux chercheurs dans un journal scientifique.

8.1.8 Semaine 8

J'avais principalement travaillé sur la comparaison de générations cette semaine là. Sur ce point Guillaume a validé mon travail.

J'avais également commencé à analyser l'évolution de nouveaux auteurs par an dans *JASIST* mais Guillaume m'a conseillé de normaliser les valeurs obtenues afin de pouvoir les comparer.

8.1.9 Semaine 9

Cette semaine avait été consacrée à la normalisation des données obtenues la semaine précédente. J'avais également intégré dans la base de données alimentant l'application web présentant Inforsid les données des articles présentés lors de l'édition 1983, récupérées pas Guillaume Cabanac lors de la 31^e édition du congrès à Paris.

8.2 Revues

En plus des compte-rendus hebdomadaires et de la réunion associée au commentaire des annotations nous avions généralement deux réunions par mois afin de pouvoir discuter de vive voix des prochaines tâches à réaliser ou des éventuelles difficultés que je rencontrais.

8.2.1 8 avril 2013

Cette réunion a eu lieu le 8 avril 2013 à 8h00 dans la salle de réunion du quatrième étage de l'IRIT. Étaient présents Guillaume Cabanac (mon maître de stage) et moi-même.

Cette réunion avait pour but principal de réaliser toutes les procédures administratives afin que je puisse disposer d'un poste de travail – création d'un compte informatique et attribution d'un badge.

Guillaume Cabanac m'a ensuite présenté plus en détail ma première mission – la valorisation d'Inforsid – et octroyé les accès et autorisations nécessaires pour mon travail.

8.2.2 19 avril 2013

Cette réunion a eu lieu le 19 avril 2013 à 13h30 dans la salle de réunion du troisième étage de l'IRIT. Étaient présents Guillaume Cabanac (mon maître de stage), et moi-même.

Elle avait principalement pour but de vérifier que je m'intégrais bien à l'IRIT et que mon travail apportait une véritable valorisation au congrès Inforsid. J'ai pu ainsi montrer les premières cartes réalisées – villes et pays les plus impliqués dans le congrès – et m'assurer qu'elles convenaient au besoin exprimé par les membres.

8.2.3 30 avril 2013

Cette réunion a eu lieu le 30 avril 2013 à 9h00 dans la salle de réunion du quatrième étage de l'IRIT. Étaient présents Guillaume Cabanac (mon maître de stage), et moi-même.

J'ai pu présenter les derniers éléments de valorisation d'Inforsid réalisés. Guillaume les a trouvés pertinents tout en me faisant remarquer que les nuages de mots pourraient prendre en compte les expressions de plusieurs mots afin d'être plus pertinents. Je devais donc trouver un moyen de faire cela la semaine suivante.

8.2.4 7 mai 2013

Cette réunion a eu lieu le 7 mai 2013 à 9h30 dans le bureau de Guillaume Cabanac à l'IRIT. Étaient présents Guillaume Cabanac (mon maître de stage), Sébastien Gerchinovitz (mon tuteur de stage), et moi-même.

J'ai présenté le travail réalisé et les tâches restantes à mon tuteur.

8.2.5 21 mai 2013

Cette réunion a eu lieu le 21 mai 2013 à 14h00 au quatrième étage de l'IRIT. Étaient présents Guillaume Cabanac (mon maître de stage), et moi-même.

Guillaume m'a fait part d'un problème qu'il avait découvert dans mon calcul de l'homophilie, ce qui rendait tout les résultats calculés précédemment pour cette métrique erronés. Je devais donc reprendre ce calcul si je voulais pouvoir l'utiliser pour mon étude de genre.

8.2.6 27 mai 2013

Cette réunion a eu lieu le 27 mai 2013 à 9h30 au quatrième étage de l'IRIT. Étaient présents Guillaume Cabanac (mon maître de stage), Sébastien Gerchinovitz (mon tuteur de stage), et moi-même.

Sébastien m'a fait remarquer qu'il serait bon que j'améliore la présentation des données analysées dans mes comptes-rendus en précisant notamment le nombres d'observations. Il m'a également donné des conseils sur l'utilisation de tests statistiques et la présentation de leurs résultats.

8.2.7 4 juin 2013

Cette réunion a eu lieu le 4 juin 2013 à 16h30 dans la salle de réunion du quatrième étage de l'IRIT. Étaient présents Guillaume Cabanac (mon maître de stage), et moi-même.

Guillaume m'a fait part des retours positifs sur la valorisation Inforsid, qu'il avait présenté lors du 31^e congrès.

Il m'a également fait quelques remarques sur mon rapport de stage, notamment sur ma représentation des schémas des bases de données utilisées.

Enfin il m'a donné ses retours sur mon compte-rendu hebdomadaire de la semaine 8 (voir section 8.1.8).

8.2.8 7 juin 2013

Cette réunion a eu lieu le 6 juin 2013 à 15h00 dans la salle de réunion du bâtiment 1R1 de l'université Paul Sabatier. Étaient présents Guillaume Cabanac (mon maître de stage), Sébastien Gerchinovitz (mon tuteur de stage), et moi-même.

Nous avons discuté des éventuels biais présents dans notre méthodologie de recherche et de l'utilisation des tests statistiques.

Cette réunion nous a permis de reconsidérer l'approche que nous avions vis-à-vis de notre échantillon.

9 — Bilan

9.1 Bilan du projet

Ce stage aura permis aux membres d'Inforsid d'obtenir plusieurs éléments de valorisation pour leur congrès, ce qui est un élément non négligeable pour présenter son activité, non seulement au grand public, mais aussi à des organismes officiels. Les éléments obtenus permettent de présenter les thèmes abordés par Inforsid mais également la diversité et l'influence des membres constituant sa communauté, tant sur le plan national qu'international. Tous ces éléments de valorisation ont été présentés à la communauté par Guillaume Cabanac lors de la 31^e édition du congrès et ont été considérés pertinents par celle-ci. Le site web d'Inforsid a même choisi de réutiliser certains de ces éléments de valorisation (visibles à l'adresse <http://inforsid.fr>).

Quant à l'étude de genre, les résultats obtenus permettent de cerner plus distinctement la position des femmes au sein de la communauté SI. Ils pourront servir de base à un article scientifique afin d'informer les autres chercheurs. Si je n'ai pas le temps de rédiger cet article durant les deux semaines de stage qu'il me reste, mon maître de stage pourra toujours se servir de ces données pour ses propres recherches.

9.2 Bilan personnel

Il m'a été donné une occasion de découvrir le monde de la recherche en tant que participante, et cette expérience a été très enrichissante pour moi. J'ai pu appréhender les problématiques auxquelles sont confrontés les enseignants-chercheurs et mieux comprendre ce milieu professionnel qui jusque là était à mes yeux assez abscon.

De plus Guillaume Cabanac m'a laissé une grande autonomie dans mon travail, ce qui m'a motivée à me documenter par moi-même et à trouver des solutions appropriées aux problèmes rencontrés. Il était néanmoins toujours présent pour me conseiller et m'apporter ses remarques sur mes analyses. Cette méthode de travail était très valorisante car j'ai pu apporter mes idées sans avoir l'impression d'uniquement enrichir les recherches de quelqu'un d'autre.

Ce stage m'aura également permis de découvrir de nouveaux outils – tels que Gnuplot ou Sofa Statistics – et de m'améliorer dans l'usage d'autres – notamment L^AT_EX pour n'en citer qu'un. J'ai également découvert de nombreuses possibilités des langages SQL et PL/SQL, tout en retravaillant des compétences acquises durant mon DUT Informatique et restées inutilisées tout au long de l'année, telles que l'utilisation de déclencheurs. Tous ces éléments me seront

sans aucun doute utiles lors de la suite de mes études et lors de ma vie professionnelle.

Même si je suis encore indécise quand mon futur, ce stage m'aura permis de vraiment apprêhender le travail de chercheur et de pouvoir effectuer un jugement éclairé lorsque je devrai choisir mon projet professionnel.

Références

- Abramo, G., D'Angelo, C. A., & Caprasecca, A. (2009, juin). Gender differences in research productivity: A bibliometric analysis of the Italian academic system. *Scientometrics*, 79(3), 517–539. doi: 10.1007/s11192-007-2046-8
- Arensbergen, P., van der Weijden, I., & Besselaar, P. (2012). Gender differences in scientific productivity: a persisting phenomenon? *Scientometrics*, 93(3), 857–868. doi: 10.1007/s11192-012-0712-y
- Cabanac, G. (2011, mai). Accuracy of inter-researcher similarity measures based on topical and social clues. *Scientometrics*, 87(3), 597–620. doi: 10.1007/s11192-011-0358-1
- Cabanac, G. (2012). Shaping the landscape of research in Information Systems from the perspective of editorial boards: A scientometric study of 77 leading journals. *JASIST*, 63(5), 977–996. doi: 10.1002/asi.22609
- Cabanac, G. (2013). Experimenting with the partnership ability ϕ -index on a million computer scientists. *Scientometrics*. doi: 10.1007/s11192-012-0862-y
- Cunningham, S. J., & Dillon, S. M. (1997, mai). Authorship patterns in information systems. *Scientometrics*, 39(1), 19–27. doi: 10.1007/BF02457428
- de Palma, P. (2001, juin). Why women avoid computer science [viewpoint]. *Communications of the ACM*, 44(6), 27–30. doi: 10.1145/376134.376145
- 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. doi: 10.1002/(SICI)1097-4571(2000)51:2<145::AID-ASI6>3.0.CO;2-9
- Hegarty, P., & Pratto, F. (2010). Handbook of gender research in psychology. In (pp. 191–211). Springer-Verlag. doi: 10.1007/978-1-4419-1465-1_10
- Hildrun, K., Alexander, P., & Johannes, S. (2012). Research evaluation. Part II: gender effects of evaluation : are men more productive and more cited than women? *Scientometrics*, 93(1), 17–30. doi: 10.1007/s11192-012-0658-0

- James, F. (1971). *Statistical Methods in Experimental Physics*. World Scientific Publishing Co. Pte. Ltd.
- Jump, P. (2013, 7 mars). *Male domination of philosophy ‘must end’*. Times Higher Education. Consulté sur <http://www.timeshighereducation.co.uk/news/male-domination-of-philosophy-must-end/2002324.article>
- Laender, A. H. F., & Oliveira, A. L. (Eds.). (2002). *String Processing and Information Retrieval, 9th International Symposium, SPIRE 2002, Lisbon, Portugal, September 11-13, 2002, Proceedings* (Vol. 2476). Springer.
- Larousse.fr. (2013, juin). Consulté sur <http://www.larousse.fr>
- Ledin, A., Bornmann, L., Gannon, F., & Wallon, G. (2007, novembre). A persistent problem. Traditional gender roles hold back female scientists. *EMBO reports*, 8(11), 982–987. doi: 10.1038/sj.embo.7401109
- Lefevre, A. (2010). *Relations entre individus appartenant à un réseau social académique – Caractérisation des relations et comparaison à la perception humaine* (Rapport technique). Université Paul Sabatier.
- Ley, M. (2002). The DBLP Computer Science Bibliography : Evolution, Research Issues, Perspectives. In A. H. F. Laender & A. L. Oliveira (Eds.), *SPIRE* (Vol. 2476, pp. 1–10). Springer. doi: 10.1007/3-540-45735-6_1
- Long, J. S. (1992). Measures of Sex Differences in Scientific Productivity. *Social Forces*, 71(1), 159–178. Consulté sur <http://www.jstor.org/stable/2579971>
- Mauleón, E., Hillán, L., Moreno, L., Gómez, I., & Bordons, M. (2013, avril). Assessing gender balance among journal authors and editorial board members. *Scientometrics*, 95(1), 87–114. doi: 10.1007/s11192-012-0824-4
- Nakhaie, M. R. (2002, mai). Gender Differences in Publication among University Professors in Canada. *Canadian Review of Sociology/Revue canadienne de sociologie*, 39(2), 151–179. doi: 10.1111/j.1755-618X.2002.tb00615.x
- Penas, C. S., & Willett, P. (2006, juin). Gender differences in publication and citation counts in librarianship and information science research [brief communication]. *Journal of Information Science*, 32(5), 480–485. doi: 10.1177/0165551506066058
- Prpić, K. (2002). Gender and productivity differentials in science. *Scientometrics*, 55(1), 27–58. doi: 10.1023/A:1016046819457
- Rossiter, M. W. (1993, mai). The Matthew Matilda Effect in Science. *Social Studies of Science*, 23(2), 325–341. doi: 10.1177/030631293023002004
- Schubert, A. (2012, avril). A hirsch-type index of co-author partnership ability. *Scientometrics*, 91(1), 303–308. doi: 10.1007/s11192-011-0559-7
- Shapiro, S. S., & Wilk, M. B. (1965, décembre). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4), 591–611. doi: 10.1093/biomet/52.3-4.591
- Shaw, C. (2013, 7 mars). *Global action on gender inequality in HE: what’s needed?* The Guardian. Consulté sur <http://www.guardian.co.uk/higher-education-network/2013/mar/07/international-womens-day-gender-inequality>

Site officiel de l'IRIT. (2013, juin). Consulté sur <http://www.irit.fr/>

Stross, R. (2008, 15 novembre). *Digital Domain – What Has Driven Women Out of Computer Science?* The New York Times. Consulté sur http://www.nytimes.com/2008/11/16/business/16digi.html?_r=0

Symonds, M. R., Gemmell, N. J., Braisher, T. L., Gorringe, K. L., & Elgar, M. A. (2006, décembre). Gender Differences in Publication Output : Towards an Unbiased Metric of Research Performance. *PLoS ONE*, 1(1), e127. doi: 10.1371/journal.pone.0000127

Taylor, S. W., Fender, B. F., & Burke, K. G. (2006, avril). Unraveling the Academic Productivity of Economists: The Opportunity Costs of Teaching and Service. *Southern Economic Journal*, 72(4), 846–859. Consulté sur <http://ideas.repec.org/a/sej/ancoec/v724y2006p846-859.html>

Ternisien, M. (2011). *Document de conception* (Rapport technique). Université Paul Sabatier.

The Accidental Mathematician. (2013, 9 février). *Gender Bias 101 For Mathematicians*. Blog « ilaba ». Consulté sur <http://ilaba.wordpress.com/2013/02/09/gender-bias-101-for-mathematicians/>

The Librarian Kate. (2013, 3 mars). *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 novembre). *More Gender Bias Uncovered : Conference Symposia Organizers*. Blog « womeninwetlands ». Consulté sur <http://blog.katekosturski.info/who-rule-the-world-girls-a-look-at-the-scholarly-literature-on-gender-and-librarianship-part-2/>

Wilcoxon, F. (1945, décembre). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6), 80–83. Consulté sur <http://www.jstor.org/stable/3001968>

Xie, Y., & Shauman, K. A. (1998). Sex Differences in Research Productivity: New Evidence about an Old Puzzle. *American Sociological Review*, 63(6), 847–870. Consulté sur <http://www.jstor.org/stable/2657505>

Annexes

Ces annexes vont vous présenter certains problèmes rencontrés durant mon stage et les solutions mises en œuvre pour y pallier d'un point de vue technique.

Gestion des « synonymes » dans la base de données Inforsid

L'insertion des données des congrès Inforsid dans la base de données alimentant l'application web était basée sur des fichiers texte : nous disposions de deux fichiers par édition, un contenant les informations sur les membres du comité de programme et un contenant celles sur les articles présentés (titre et auteur(s)), plus un fichier contenant les villes ayant accueilli chaque édition. À partir de ces fichiers, un programme en langage C avait pour tâche de déterminer la liste des chercheurs et des villes.

Le principal problème de cette méthode était que les fautes de frappe sont très difficilement repérables avant l'insertion des données dans la base. En effet, les noms de chercheurs et de villes étant dispersées dans de nombreux fichiers, il aurait été fastidieux de devoir tous les contrôler à la recherche d'éventuelles erreurs. De fait la base de données était « polluée » 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 soundex¹. Cette méthode retourne la représentation phonétique du paramètre passé, et permet donc de comparer des noms et prénoms ne s'écrivant pas pareil mais se prononçant de la même façon. Il faut noter que cette méthode a cependant une limite non négligeable : elle se base sur la prononciation américaine des consonnes. Chaque terme est représenté sous la forme d'une chaîne de caractères composée d'une lettre (l'initiale du terme) et de 3 chiffres représentant les consonnes suivantes, deux consonnes ayant la même prononciation étant codées par le même chiffre².

J'ai également dû prendre en compte le fait que nous ne disposions pas du prénom complet pour certains chercheurs mais seulement de l'initiale, et que nous devions donc, lorsque c'est le cas, ne comparer que la phonétique du nom de famille. Enfin j'ai traité le cas où une personne

1. Documentation officielle : http://docs.oracle.com/cd/B19306_01/server.102/b14200/functions148.htm.

2. Plus d'informations à l'adresse <http://www.archives.gov/research/census/soundex.html>.

change de nom au cours de sa carrière (nous avions par exemple dans la base « Karen Sauvagnat » et « Karen Pinel-Sauvagnat »). Le code SQL ayant permis la création de ces tables est présenté dans l'extrait de code 9.1.

```

1 -- villes
2 create table inforsid_syVilles as
3   select v1.idV as id1, v1.nomville as nom1, v2.idV as id2, v2.nomville
4     as nom2
5   from inforsid_ville v1, inforsid_ville v2
6  where soundex(v1.nomville) = soundex(v2.nomville)
7  and v1.nomville > v2.nomville
8  order by 2, 4 ;
9
10 -- personnes
11 create table inforsid_syPersonnes as
12   select p1.idP as id1, p1.nom as nom1, p1.prenom as prenom1, p2.idP as
13     id2, p2.nom as nom2, p2.prenom as prenom2
14   from inforsid_personne p1, inforsid_personne p2
15  where ((( regexp_substr(p2.nom, '\w+', 1, 2) is not null
16    and (p2.nom like p1.nom||'-' or p2.nom like '%-'||p1.nom))
17    or (regexp_substr(p1.nom, '\w+', 1, 2) is not null
18      and (p1.nom like '%-'||p2.nom or p1.nom like p2.nom||'-%'))
19    and p1.prenom = p2.prenom
20    and p1.nom > p2.nom)
21  or (soundex(p1.nom) = soundex(p2.nom)
22    and (soundex(p1.prenom) = soundex(p2.prenom)
23      or (length(p1.prenom) <= 4 and substr(p1.prenom,
24        length(p1.prenom), 1) = '.')
25      or (length(p2.prenom) <= 4 and substr(p2.prenom,
26        length(p2.prenom), 1) = '.'))
27    and (p1.nom > p2.nom
28      or (p1.nom = p2.nom
29        and p1.prenom > p2.prenom)))
30  order by 2, 3, 5, 6 ;

```

EXTRAIT DE CODE 9.1 – Code SQL permettant la création des tables détectant les « synonymes » dans la base de données alimentant l'application web présentant Inforsid.

Après la création de ces tables l'utilisateur pouvait ainsi consulter ces couples, et supprimer les faux-positifs, grâce à l'utilitaire de visualisation des données d'une table de SQL*Developer (voir figure 9.1).

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. Le code de cette procédure est présenté dans l'extrait de code 9.2.

```

1 create or replace procedure gererSy is
2 begin
3   --traitement des villes
4   for coupleV in (select id1, nom1, id2 from inforsid_syVilles) loop
5     update inforsid_congres
6       set idV = coupleV.id1
7       where idV = coupleV.id2;
8     update inforsid_ecrire
9       set idV = coupleV.id1
10      where idV = coupleV.id2;
11     update inforsid_membre
12       set idV = coupleV.id1
13       where idV = coupleV.id2;

```

ID1	NOM1	PRENOM1	ID2	NOM2	PRENOM2
-1	597 ANDRES	F.	812 ANDRE	Pascal	
-2	1212 ANDRES	Samuel	597 ANDRES	F.	
-3	142 ARRONATEGUI	U.	818 ARNAUD	Nicolas	
-4	282 ATTQUI	A.	837 AOUADI	Hamed	
-5	1319 AUDEH	Bissan	282 ATTQUI	A.	
-6	1214 AYAD	Sarah	282 ATTQUI	A.	
-7	33 BEGUE	J.M.	123 BEC	Michel	
-8	177 BENOUHIBA	F.Z.	1179 BENAFIA	Ali	

FIGURE 9.1 – Utilitaire de visualisation des données d'une table de SQL*Developer grâce auquel l'utilisateur pouvait supprimer les couples de « synonymes » détectés par erreur. La table présentée ici est celle recensant les synonymes parmi les noms de chercheurs et les lignes indiquées en rouge sont des couples détectés par erreur destinés à être supprimés.

```

14      delete from inforsid_ville
15          where idV = coupleV.id2;
16      update inforsid_ville
17          set nomVille = coupleV.nom1
18          where idV = coupleV.id1;
19  end loop;
-- traitement des personnes
20  for coupleP in (select id1, nom1, prenom1, id2 from
21      inforsid_syPersonnes) loop
22      update inforsid_ecrire
23          set idP = coupleP.id1
24          where idP = coupleP.id2;
25      update inforsid_membre
26          set idP = coupleP.id1
27          where idP = coupleP.id2;
28      delete from inforsid_personne
29          where idP = coupleP.id2;
30      update inforsid_personne
31          set nom = coupleP.nom1, prenom = coupleP.prenom1
32          where idP = coupleP.id1;
33  end loop;
34  commit;
35 end gererSy;
```

EXTRAIT DE CODE 9.2 – Code SQL permettant la création de la procédure ayant pour tâche de fusionner les deux entités des couples « synonymes » dans la base de données alimentant l'application web présentant Inforsid.

La base de données était ainsi plus cohérente et la procédure facilement réutilisable. Il est cependant à noter que les fichiers CSV initiaux n'étaient pas modifiés par cette procédure, et qu'il était donc nécessaire de la relancer à chaque nouvelle insertion de données.

Gestion des pays dans la base de données Inforsid

Auparavant, pour spécifier le pays correspondant aux villes destinées à être insérées dans la base de données Inforsid, il fallait parcourir le fichier contenant la liste de villes résultant du traitement des fichiers texte contenant les données de chaque édition par le programme en

C. Celui-ci contenait une fonction destinée à récupérer les pays précédemment spécifiés par l'utilisateur si une liste de ville existait déjà (résultat d'un précédent traitement des données) mais cette fonctionnalité avait souvent tendance à corrompre la liste finale.

J'ai alors choisi d'automatiser au maximum la procédure d'attribution d'un pays à une ville. J'ai tout d'abord créé une table référence destinée à contenir l'identifiant d'une ville, son nom et son pays dans laquelle j'ai inséré toutes les villes françaises – téléchargées sur SQL.sh³.

J'ai ensuite créé une vue recensant les villes présentes dans la liste des villes concernées par le congrès Inforsid 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 et il devait donc les corriger (il pouvait chercher l'orthographe correcte dans la table référence),
- pour les villes étrangères, il modifiait leur pays et un déclencheur insérait alors ces nouvelles villes dans la table référence (le code de ce déclencheur est présenté dans l'extrait de code 9.3).

```

1 create or replace trigger ajoutVilleBase
2     instead of update on inforsid_villesSansPays
3         for each row
4 declare
5     v_idNewV    inforsid_base_pays.idV%type;
6 begin
7     select max(idV)+1 into v_idNewV
8     from inforsid_base_pays;
9     insert into inforsid_base_pays(idV, ville, pays)
10        values (v_idNewV, :new.nomVille, :new.nomPays);
11 end ajoutVilleBase;
```

EXTRAIT DE CODE 9.3 – Code SQL permettant la création du déclencheur insérant une nouvelle ville et le pays correspondant dans la table référence dans la base de données alimentant l'application web présentant Inforsid.

Une fois toutes les villes corrigées ou insérées dans la table référence, la vue était donc vide. L'utilisateur devait alors lancer une procédure qui mettait à jour les villes concernées par le congrès Inforsid. Le code de cette procédure est visible dans l'extrait de code 9.4.

```

1 create or replace procedure gererPays is
2     v_pays inforsid_base_pays.pays%type;
3 begin
4     for v in (select idV, nomVille, nomPays from inforsid_ville) loop
5         select pays into v_pays
6             from inforsid_base_pays
7                 where lower(ville) = lower(v.nomVille);
8         if v.nomPays <> v_pays then
9             update inforsid_ville
10                set nomPays = v_pays
11                  where idV = v.idV;
12         end if;
13     end loop;
14     commit;
15 exception
16     when no_data_found then
17         dbms_output.put_line('/!\ Une ville n''est pas dans base_pays.');
18     when others then
19         dbms_output.put_line(sqlcode || ' : ' || sqlerrm);
20 end;
```

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

EXTRAIT DE CODE 9.4 – Code SQL de la procédure mettant à jour les pays des villes impliquées dans le congrès Inforsid.

La détermination « manuelle » de l’identifiant de la nouvelle ville (`max(idV)+1`) n’est pas robuste aux accès concurrents mais étant donné que l’insertion des données n’est effectuée que par un seul utilisateur (Guillaume Cabanac ou l’un(e) de ses stagiaires) l’utilisation d’une séquence aurait eu un intérêt minime dans notre cas.

L’avantage de cette méthode est que la base de villes présentes – et donc automatiquement générées – augmente d’année en année au fur et à mesure des insertions de données. Elle diffère cependant fortement de la méthode mise en place précédemment, au sens où l’attribution des pays se fait à présent après l’insertion des villes dans la base.